NUMÉRO SPÉCIAL

POUR LA SCIENCE

Novembre 2013 - n° 433

www.pourlascience.fr

Édition française de Scientific American

BIG BANG numérique

Les données massives changent-elles le monde ?

Fragiles archives numériques
Comment protéger la vie privée
Tout le savoir en libre accès ?
Les mémoires de demain
L'ADN, une mémoire modèle ?

Allemagne : 10,20 € - Belgique : 7,8 € - Canada /S : 11,95 CAD - Guadeloupe/St Martin /S : 7,90 € - Guyane /S : 7,90 € - Italie : 7,90 € - Luxembourg : 7,90 € - Maroc : 69 MAD - Martinique /S : 7,9 € - Nle Calédonie Wallis/S : 1100 XPF - Polynésie Française/S : 1100 XPF - Portugal : 7,90 € - Réunion/A : 9,90 € - Suisse : 12 CHF

M 02687 - 433 - F: 6,60 € - RD

PERMETTRE À VOTRE ÉPARGNE DE FAIRE DES PETITS C'EST AUSSI ÇA LA GMF.

TBWAPARIS



Une **épargne** performante et sûre pour préparer les **projets de demain.**

Avec les contrats d'assurance vie et les supports en euros de la GMF, vous vous constituez petit à petit une épargne performante qui vous permettra de financer tous vos projets d'avenir, en toute sécurité.

ÉPARGNE
3,05%*
TAUX SERVI EN 2012

Trophées d'or •2013• des Contrats d'Assurance Vie
Le Revenu
MULTÉO
Catégorie contrats et fonds en euros

Rendez-vous sur www.gmf.fr
ou appelez le
N° Vert 0 800 88 11 62
APPEL GRATUIT DEPUIS UN POSTE FIXE

GMF
Assurément Humain

*3,05 %, taux servi en 2012 : frais de gestion déduits et avant prélèvements sociaux pour tous les contrats d'assurance vie en euros de la GMF et les supports réguliers de Multéo et de Certigo. Les performances passées ne préjugent pas des performances futures.

GMF VIE - Société anonyme au capital de 179 676 528 € entièrement versé. Entreprise régie par le Code des assurances. R.C.S. Paris B 315 814 806 - Siège social : 76, rue de Prony - 75857 Paris Cedex 17 - Adresse postale : 1, rue Raoul Dautry - 95127 Ermont Cedex.

POUR LA SCIENCE
www.pourlascience.fr
8, rue Férou - 75278 Paris cedex 06
Standard : **01 55 42 84 00**

Groupe POUR LA SCIENCE
Directrice de la rédaction : Françoise Pétry
Pour la Science
Rédacteur en chef : Maurice Mashaal
Rédacteurs : François Savatier, Marie-Neige Cordonnier, Philippe Ribeau-Gésippe, Guillaume Jacquemont, Sean Bailly
Dossiers Pour la Science
Rédacteur en chef adjoint : Loïc Mangin
Cerveau & Psycho
Rédactrice en chef : Françoise Pétry
Rédacteur : Sébastien Bohler
L'Essentiel Cerveau & Psycho
Rédactrice : Bénédicte Salthun-Lassalle
Directrice artistique : Céline Lapert
Secrétariat de rédaction/Maquette : Sylvie Sobelman, Pauline Bilbault, Raphaël Queruel, Ingrid Leroy, Caroline Vanhoove
Site Internet : Philippe Ribeau-Gésippe
Marketing : Élise Abib et Ophélie Maillet assistées d'Anaïs Grelet
Direction financière : Anne Gusdorf
Direction du personnel : Marc Laumet
Fabrication : Jérôme Jalabert assisté de Marianne Sigogne
Presse et communication : Susan Mackie
Directrice de la publication et Gérante : Sylvie Marcé
Conseillers scientifiques : Philippe Boulanger et Hervé This
Ont également participé à ce numéro :
Philippe Gérard, Jean-Claude Guillemin, Jacques Laskar, Gilles Laurent, Caroline Martel, Alessandro Papitto, Laurence Perotto, Christophe Pichon, Sean Raymond, Pascale Thiollier-Dumartin

PUBLICITÉ France
Directeur de la Publicité : Jean-François Guillotin (jf.guillotin@pourlascience.fr), assisté de Nada Mellouk-Raja
Tél. : 01 55 42 84 28 ou 01 55 42 84 97 • Fax : 01 43 25 18 29
SERVICE ABONNEMENTS
Ginette Bouffaré. Tél. : 01 55 42 84 04
Espace abonnements :
http://tinyurl.com/abonnements-pourlascience
Adresse e-mail : abonnements@pourlascience.fr
Service des abonnements - 8 rue Férou - 75278 Paris cedex 06
Commande de livres ou de magazines :
0805 655 255 (numéro vert)
DIFFUSION DE POUR LA SCIENCE
Contact kiosques : À Juste Titres ; Benjamin Boutonnet
Tel : 04 88 15 12 41
Canada : Edipresse : 945, avenue Beaumont, Montréal, Québec, H3N 1W3 Canada.
Suisse : Servidis : Chemin des châlets, 1979 Chavannes - 2 - Bogis
Belgique : La Caravelle : 303, rue du Pré-aux-oies - 1130 Bruxelles.
Autres pays : Éditions Belin : 8, rue Férou - 75278 Paris Cedex 06.

SCIENTIFIC AMERICAN Editor in chief : Mariette DiChristina. Editors : Fred Guterl, Ricky Rusting, Philip Yam, Mark Fischetti, Christine Gorman, Anna Kuchment, Michael Moyer, Gary Stix, Kate Wong. President : Steven Inchcoombe. Executive Vice President : Michael Florek.
Toutes demandes d'autorisation de reproduire, pour le public français ou francophone, les textes, les photos, les dessins ou les documents contenus dans la revue « Pour la Science », dans la revue « Scientific American », dans les livres édités par « Pour la Science » doivent être adressées par écrit à « Pour la Science S.A.R.L. », 8, rue Férou, 75278 Paris Cedex 06.
© **Pour la Science S.A.R.L.** Tous droits de reproduction, de traduction, d'adaptation et de représentation réservés pour tous les pays. La marque et le nom commercial « Scientific American » sont la propriété de Scientific American, Inc. Licence accordée à « Pour la Science S.A.R.L. ».
En application de la loi du 11 mars 1957, il est interdit de reproduire intégralement ou partiellement la présente revue sans autorisation de l'éditeur ou du Centre français de l'exploitation du droit de copie (20, rue des Grands-Augustins - 75006 Paris).
OJD PRESSE PAYANTE Diffusion Certifiée 2012

Évolution ou révolution ?

Le mot de révolution est galvaudé. Au moindre changement, à la moindre évolution, d'aucuns prédisent rapidement une révolution. Mais si l'on s'en tient à sa définition, il s'agit d'« un changement brusque et important dans l'ordre social et moral, d'une transformation complète. » Quelles ont été les « vraies » révolutions scientifiques : l'invention de l'imprimerie ? La révolution industrielle ? Le séquençage du génome humain ? L'avènement des neurosciences ?

Selon Victor Hugo, « L'invention de l'imprimerie est le plus grand événement de l'histoire. » Un événement, pas une révolution. Pourtant, permettre une diffusion et une conservation du savoir aurait sans doute mérité le vocable. La révolution industrielle porte en elle son qualificatif. Le séquençage du génome et l'avènement des neurosciences ont laissé croire à des révolutions. Mais les désillusions ont suivi l'engouement des débuts. La génomique n'a pas apporté les progrès escomptés, sans doute parce que l'on avait trop négligé le dialogue entre gènes et environnement. Les neurosciences ont beaucoup promis, mais divers blocages se font jour.

Si les livres se conservent des centaines d'années, les mémoires numériques sont éphémères.

Sans doute est-ce le lot de toutes les révolutions de passer par des phases d'effervescence, puis de déception. Est-ce le cas pour le big bang numérique ? Aujourd'hui, nous serions dans la phase d'effervescence : l'accès pour tous au savoir universel, la mise en mémoire de tout ce que produisent l'être humain, la nature et la science ; la puissance des outils numériques est considérable, pleine de promesses pour la recherche scientifique et technique. Mais… elle présente la fragilité du colosse aux pieds d'argile, car, si les livres se conservent des centaines d'années, les mémoires numériques sont éphémères.

Dès lors, faut-il sélectionner les données à conserver ? Comment les organiser, les archiver, les partager ? Quels supports utiliser pour assurer leur pérennité ? Par ce que le numérique modifie déjà, par ce qu'il va transformer, le terme de révolution est justifié. Ou plutôt celui de big bang ouvrant une ère nouvelle de communication et de partage. À condition que l'homme n'ait pas à vivre avec un *Big Brother* hypermnésique – selon l'expression de Gilles Dowek – surveillant sans cesse ses faits et gestes, voire ses pensées. À condition qu'il ait le droit à l'oubli. Et si les concepteurs de supports numériques s'inspiraient de l'ADN, gigantesque mémoire du vivant, et de la mémoire biologique, naturellement dotée de la capacité d'oublier ? ■

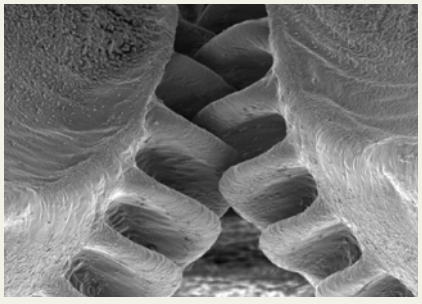
SOMMAIRE

1 ÉDITO

4 BLOC-NOTES
Didier Nordon

Actualités

6 **Les insectes ont aussi inventé les engrenages**



7 **L'Univers est-il hyperbolique ?**

11 **Des bactéries pour combattre l'obésité ?**

12 **De l'oxygène atmosphérique très ancien**

... et bien d'autres sujets.

Opinions

18 POINT DE VUE
ISET : un test génétique prénatal à encadrer
Floriane Blanc

19 DÉVELOPPEMENT DURABLE
Énergies renouvelables : le coût de l'intermittence
Renaud Crassous et Fabien Roques

23 VRAI OU FAUX
Certains autistes sont-ils surdoués ?
Isabelle Soulières

24 ENTRETIEN
La difficile transmission... d'un autoapprentissage
Entretien avec Philippe Descola et l'un de ses élèves, Emmanuel de Vienne, autour de la méthode de travail des ethnologues.

Ce numéro comporte deux encarts d'abonnement *Pour la Science* brochés sur la totalité du tirage et un encart Voyage en Turquie *(Dialogue Marketing)* posé sur la totalité de la diffusion abonnés
Illustration de couverture : © agsandrew/shutterstock.com



BIG BANG numérique

Les données massives changent-elles le monde ?

26 SCIENCE ET SOCIÉTÉ
L'explosion mémorielle change la donne
Gilles Dowek

32 INFORMATIQUE
Un déluge de données
Serge Abiteboul et Pierre Senellart

36 PHYSIQUE
Les pétaoctets du LHC
Faïrouz Malek

38 BIOLOGIE
Génomes, protéomes et transcriptomes à foison
Guy Perrière

40 ASTRONOMIE
SDSS, un relevé... astronomique !
Céline Reylé

42 SCIENCES DE L'INFORMATION
Au cœur d'un centre d'archivage
Francis Daumas et Marion Massol

48 SCIENCES DE L'INFORMATION
La numérisation des bibliothèques
Frédéric Blin

© Pour la Science - n° 433 - Novembre 2013

52 SCIENCES DE L'INFORMATION
Dictionnaires et encyclopédies à l'ère numérique
Rémi Mathis

54 SCIENCES DE L'INFORMATION
La science accessible à tous
Christine Berthaud et Agnès Magron

56 SCIENCES SOCIALES
Orienter la société grâce aux données massives
Alex Pentland

62 INFORMATIQUE
Comment préserver l'anonymat ?
Tristan Allard, Benjamin Nguyen et Philippe Pucheral

70 INFORMATIQUE
La protection des informations sensibles
Rémy Chrétien et Stéphanie Delaune

78 TECHNOLOGIE
La quête d'un support numérique durable
Franck Laloë et Erich Spitz

86 INFORMATIQUE
Organiser le partage pour préserver les données
Roberto Di Cosmo

92 PHYSIQUE
Les mémoires du futur
Luca Perniola

100 BIOTECHNOLOGIE
Stocker de l'information dans l'ADN
Christophe Dessimoz

102 GÉNÉTIQUE
L'ADN, mémoire numérique du vivant
Vincent Daubin, Simon Penel et Éric Tannier

110 NEUROSCIENCES
La mémoire biologique est-elle codée ?
Franck Chaillan

© Pour la Science - n° 433 - Novembre 2013

Regards

118 HISTOIRE DES SCIENCES
Cinquante ans de chaos
Adilson Motter et David Campbell
En 1963, le météorologiste américain Edward Lorenz montrait que la prédictibilité déterministe est dans certains cas illusoire.

124 LOGIQUE & CALCUL
La quête du pavé apériodique unique
Jean-Paul Delahaye
On espérait depuis longtemps découvrir un pavé unique avec lequel le pavage du plan serait forcément non périodique. Un tel pavé a été trouvé... enfin, presque.

132 SCIENCE & FICTION
Le sexe chez les extraterrestres
J.-Sébastien Steyer et Roland Lehoucq

134 ART & SCIENCE
Promenade dans un paysage... épigénétique
Loïc Mangin

136 IDÉES DE PHYSIQUE
Des gouttes bien calibrées
Jean-Michel Courty et Édouard Kierlik

141 SCIENCE & GASTRONOMIE
Il faut faire sauter le riz *après* l'avoir cuit à l'eau
Hervé This

142 À LIRE

POUR LA SCIENCE
Maintenant sur tablette et sur smartphone !
Téléchargez gratuitement l'application « Pour la Science »



Le 1er numéro est offert !
- lecture intuitive
- sommaire interactif
- contenus enrichis
- feuilletage hors-connexion

Disponible sur App Store
DISPONIBLE SUR Google play

➜ SOMMES-NOUS SI MALADES ?

Il y a les médecins, les infirmiers et les secouristes, bien sûr. Mais aussi les assistants sociaux, les éducateurs, les psychologues, les conseillers en économie sociale et familiale, les conseillers conjugaux, les aides aux personnes âgées, les travailleurs du *care*, les sophrologues et les nombreux autres dispensateurs de techniques de bien-être, les prédicateurs de toutes sortes de religions montrant la voie vers le bonheur, les inventeurs de solutions miracles, etc., etc. Quiconque voudrait oublier que l'humanité est une espèce malade n'a qu'à se remémorer la liste impressionnante des professions consacrées uniquement à guérir des maux – maux du corps, maux de l'âme – pour revenir à une appréciation plus véridique de la situation.



➜ L'ILLUSION DE LA PENSÉE UNIVERSELLE

Si je suis athée, je ne considérerai pas la foi comme une tendance universelle de l'être humain. Si j'ai de l'aversion pour les mathématiques, je ne verrai pas en elles un langage universel. Si je crois à la nécessité de réguler l'économie, je ne prendrai pas la compétition entre les marchés pour un idéal universel.

Une conception ne peut passer à mes yeux pour universelle que si elle est mienne. Voilà une évidence dont on oublie de se méfier. Pire, on se laisse parfois aller à une espèce de réciproque implicite. J'ai telle conception ; or je ne suis ni idiot ni aveuglé de préjugés ; donc elle est universelle. Ou devrait l'être.

Parmi les idées d'un individu, en vérité, n'ont valeur universelle que celles qui ne dépendent ni de sa personnalité ni de sa culture, donc qui n'expriment rien de lui. Voilà qui n'est pas engageant. D'abord, il n'est pas donné à un individu d'être certain qu'une idée qui l'habite est indépendante de lui. En outre, à supposer qu'elle le soit effectivement, il n'a aucune raison de s'attacher à quelque chose d'aussi désincarné. Voilà pourquoi je soupçonne quiconque professe l'universalisme de vouloir imposer à tous des idées qui, en fait, expriment sa culture, voire sa seule individualité.

➜ PLOMBS FAUTIFS

Rldasedlrad les dlcmhypbgf.

Ce n'est pas un message codé. Enfin, peut-être que si.

Ce n'est pas une série de fautes d'impression : une telle accumulation, vous n'y pensez pas. En fait, si, c'en est une.

Ce charabia illisible ne saurait être le titre d'un texte français. Et pourquoi pas ?

Ce borborygme imprononçable ne se traduit sûrement pas. Mais si. En anglais.

Cessons les mystères, expliquons-nous. *Rldasedlrad les dlcmhypbgf* n'est pas un message codé, en ce sens que je suis tombé dessus sans m'être mêlé de codage. Mais rien ne garantit que jamais personne n'a utilisé cette suite de lettres pour crypter quelque message.

En 1923, Valery Larbaud a lu dans un journal l'annonce de la proche parution de l'un de ses livres, intitulé, selon le journal, *Rldasedlrad les dlcmhypbgf*. Cette information a produit la commotion qu'on devine : le titre du livre que Larbaud avait en préparation n'était évidemment pas cet ahurissant fatras de lettres ! Les responsables de la bourde étaient les linotypistes du journal. Il arrivait parfois que les fautes d'impression chamboulent une ligne entière. Comme quoi, nous avons progressé : certes, les fautes d'impression continuent d'être abondantes dans nos journaux, mais elles laissent en général deviner, sous ce qui est, ce qui aurait dû être. Tel n'était pas le cas du temps de Larbaud. Ce dernier s'est vengé avec humour des linotypistes dans un texte intitulé, cela va de soi, « Rldasedlrad les dlcmhypbgf », paru en 1927 dans le recueil *Jaune Bleu Blanc*.

Une erreur fréquente des linotypes anglaises était de faire apparaître la suite *Etaoin shrdlu*. Erreur si classique, que *Etaoin shrdlu* fait l'objet d'un article sur Wikipédia. Je propose donc de traduire le français *Rldasedlrad les dlcmhypbgf* par l'anglais *Etaoin shrdlu*. Ma traduction n'est pas belle, mais elle est fidèle.

➜ LE COMTE POSEUR

« Le désespoir est la plus petite de nos erreurs », a écrit Isidore Ducasse, alias Lautréamont (*Poésies*, II). Quand on sait que cet auteur, qui n'eut aucun succès de son vivant, est mort à peine moins jeune qu'Évariste Galois, on ne peut qu'être bouleversé par sa sincérité. Il a pressenti le tragique de la condition humaine, dont il allait donner un exemple.

Mais, si on sait aussi que Vauvenargues avait écrit : « Le désespoir est la plus grande de nos erreurs », la phrase de Lautréamont acquiert la tonalité d'un jeu littéraire, qui la rend beaucoup moins émouvante. On peut même y soupçonner un brin de pose.

À quel point le sens d'une phrase change selon ce qu'on sait du contexte dans lequel elle a été écrite !



© Pour la Science - n° 433 - Novembre 2013

➜ PENSÉES VERLANES

Critiquant l'accumulation des horreurs dans les livres de Sade, qu'il jugeait lassante, l'écrivain Jean Paulhan disait : « Le pire est l'ennemi du mal. »

Inspirés par Paulhan, inversons terme à terme des formules connues, afin d'en obtenir de nouvelles. Une citation éculée de Coubertin (« Il n'est pas nécessaire d'espérer pour entreprendre, ni de réussir pour persévérer ») donne ainsi : « Il n'est pas nécessaire de craindre pour renoncer, ni d'échouer pour abandonner », réflexion qui, ma foi, en vaut d'autres. Inverser la boutade « La culture, c'est ce qui reste quand on a tout oublié », ne mène à rien de nouveau, puisqu'on obtient : « L'inculture, c'est ce à quoi on parvient quand on retient tout. » Or Talleyrand avait dit : « [Les émigrés] n'ont rien appris ni rien oublié. » Le dicton « Les grands esprits se rencontrent » devient « Les petits corps s'évitent » : de quoi inquiéter les physiciens qui placent trop d'espoirs dans les collisionneurs.

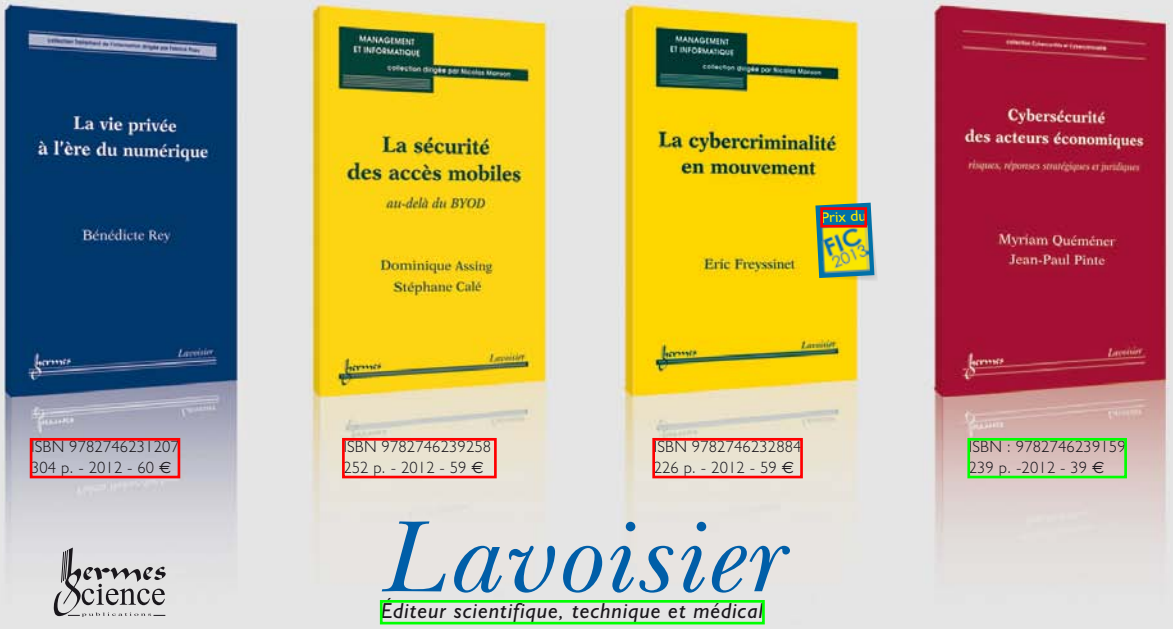
« Pensée échappée. Je la voulais écrire ; j'écris au lieu qu'elle m'est échappée. » Ce fragment, dans lequel Pascal déclare une absence de pensée, se trouve dans le recueil des pensées de Pascal ! C'est le fragment obtenu en l'inversant qui serait en droit de figurer dans un recueil de pensées : « Pas de pensée. J'écris que je n'en ai pas, et cela est une pensée. » Ne faisons pas de jaloux, inversons Descartes.

« Je ne pense pas, donc je ne suis pas. » Parfait. D'après Pascal inversé, écrire « Je ne pense pas », c'est faire acte de pensée. Or, d'après Descartes, je pense implique que je suis. Donc, quand j'écris que je ne pense pas, je suis et je ne suis pas. Que pouvions-nous espérer de mieux que ce paradoxe définitif ? ■



Cybersécurité :
enjeux et nouvelles pratiques

Disponibles chez votre libraire habituel ou sur www.lavoisier.fr



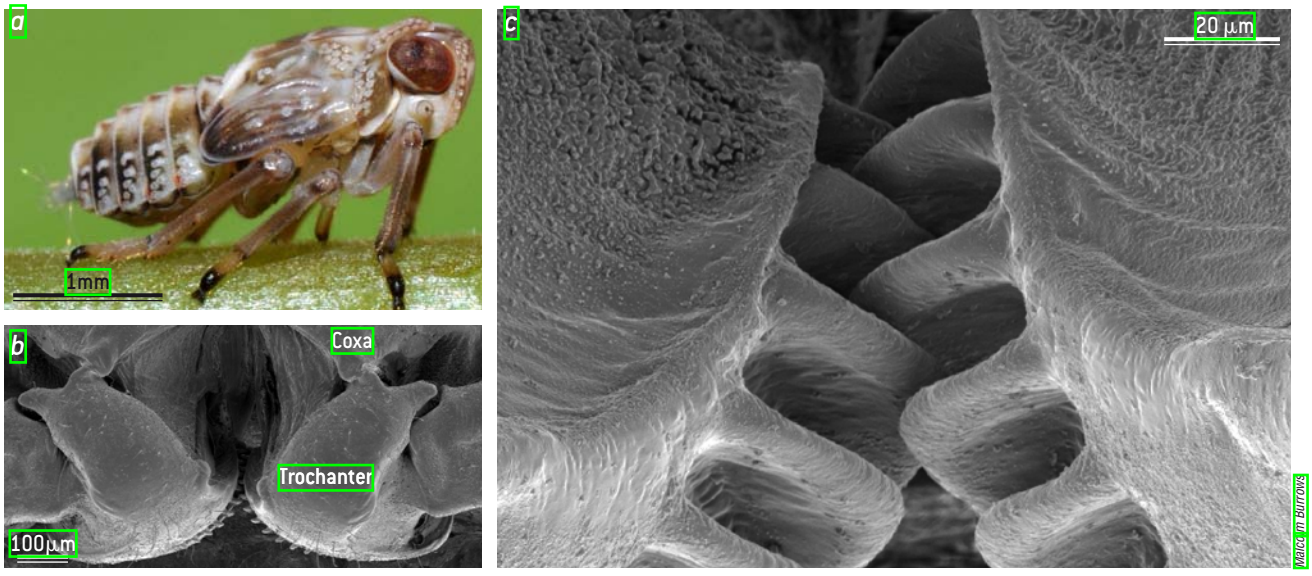
Hermes Science publications

Lavoisier
Éditeur scientifique, technique et médical

Biophysique

Les insectes ont aussi inventé les engrenages

Chez la nymphe d'une petite cigale sauteuse, des dents d'engrenage assurent la synchronisation du mouvement des deux pattes postérieures.



Chez la nymphe de la cigale bossue, *Issus coleoptratus (a)*, le mouvement des deux pattes postérieures, responsables du saut, met en jeu un mécanisme d'engrenage. Ce dernier implique le trochanter (partie de la patte précédant le fémur et située sous le corps de l'insecte). Une vue du dessous de l'insecte, prise au microscope électronique à balayage, montre les deux trochanters avec leur rangée de dents sur le côté interne *(b)*. L'imbrication des deux rangées de dents apparaît sur une autre vue, de grossissement supérieur *(c)*.

La bio-ingénierie consiste à s'inspirer du monde vivant pour concevoir des dispositifs techniques innovants. Inversement, les organismes vivants s'inspireraient-ils des inventions de l'homme ? On pourrait presque le croire : Malcolm Burrows, de l'Université de Cambridge au Royaume-Uni, et son collègue Gregory Sutton ont mis au jour un mécanisme d'engrenage impliqué dans le saut d'une petite cigale sauteuse, la « cigale bossue » *Issus coleoptratus*.

Chez les insectes sauteurs (criquets, puces, cercopes, cicadelles, etc.), le saut est réalisé par une détente rapide et puissante des deux pattes postérieures. On trouve deux configurations de ces dernières. Chez les criquets et les puces, les deux pattes bougent dans des plans verticaux distincts de chaque côté du corps de l'insecte. Chez les très bons sauteurs, tels les cercopes et les cicadelles, les bases des pattes postérieures, sous le corps de l'insecte, tournent en sens inverses et dans un même plan à peu près horizontal. Mais cette seconde configuration exige une synchronisation précise des mouvements des deux pattes : sinon, en sautant, l'insecte subirait une rotation dans le plan horizontal.

Comment cette synchronisation est-elle assurée ? Chez la cigale bossue, les influx nerveux sont envoyés aux muscles correspondants par deux paires indépendantes de neurones, ce qui ne suffit pas à synchroniser de façon assez précise, à l'échelle de quelques microsecondes, les mouvements des deux pattes. Grâce à de la vidéo ultrarapide et à un examen anatomique, M. Burrows et G. Sutton ont montré que chez la nymphe de la cigale bossue, une synchronisation précise a lieu grâce à une rangée courbe de dents présente sur le côté interne du trochanter (partie de la patte qui suit la coxa, ou hanche, et précède le fémur) de chaque patte *(voir la figure)*. Les deux trochanters, en contact sur une partie de leur côté interne, s'articulent ainsi par un engrenage qui les contraint à se mouvoir de conserve, avec un décalage inférieur à 30 microsecondes.

Ce mécanisme à engrenage ne concerne que la nymphe de la cigale bossue et d'autres espèces apparentées. Les dents disparaissent lors de la dernière mue, quand l'insecte devient adulte. Comment alors expliquer que les adultes sautent encore mieux que les nymphes ? M. Burrows avait établi en 2010 que chez eux, la synchronisation des deux pattes postérieures est assurée, au moins en partie, par le frottement entre les coxae. Le frottement pourrait être plus avantageux chez les adultes en raison de leur plus grande taille. Un autre point à souligner est que chez les nymphes, des dents cassées sont remplacées lors de la mue suivante, ce qui ne peut être le cas chez les adultes.

→ Maurice Mashaal

Science, vol. 341, pp. 1254-1256, 2013

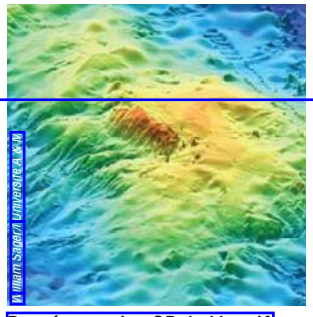
© Pour la Science - n° 433 - Novembre 2013

Géophysique

Le plus grand volcan terrestre est… sous-marin

Avec des collègues, William Sager, de l'Université A&M au Texas, vient de montrer qu'un mont sous-marin situé à quelque 1500 kilomètres à l'Est du Japon est l'un des plus grands volcans du Système solaire.

Les géophysiciens ont établi par plusieurs profils sismiques que le mont Tamu, vaste plateau de 650 kilomètres de long et de 450 kilomètres de large, est une structure d'une seule pièce. Sous les sédiments qui le recouvrent, des strates de laves alternent avec des épanchements basaltiques épais de dizaines de mètres. Ces laves sont inclinées de moins de deux degrés, ce qui montre qu'elles se sont étalées alors qu'elles étaient encore très fluides. Bref, tout indique que le mont Tamu est un immense volcan bouclier construit par des épanchements successifs, si grand qu'il ne peut être comparé qu'à l'Olympus Mons sur Mars.



Représentation 3D du Massif Tamu, le plus grand volcan connu sur Terre (650 sur 450 kilomètres).

D'après l'âge radiométrique de ses roches, il s'est formé il y a quelque 145 millions d'années à la rencontre de trois dorsales, ce qui rappelle le cas de l'Islande : à cheval sur la dorsale médio-atlantique, cette île résulte d'une remontée du matériau mantellique profond. Quoi qu'il en soit, nous avons appris qu'il est possible qu'un immense plateau sous-marin soit un volcan.

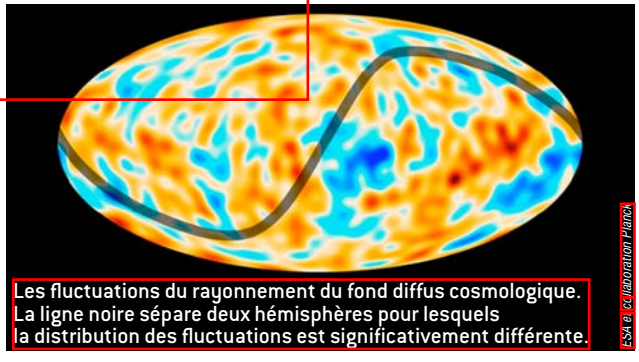
→ François Savatier

Nature Geoscience, en ligne le 5 septembre 2013

Cosmologie

L'Univers est-il hyperbolique ?

Le rayonnement du fond diffus cosmologique recèle peut-être des surprises. Les mesures du satellite *Planck* ont montré que les propriétés de ce rayonnement émis lorsque l'Univers était âgé de 380000 ans sont compatibles avec le modèle du Big Bang dans sa version la plus simple. Elles présentent néanmoins quelques anomalies. Les propriétés de la température du rayonnement devraient être identiques dans toutes les directions. Or ce n'est pas le cas des minuscules fluctuations autour de sa valeur moyenne : leur amplitude est plus grande dans un hémisphère du ciel que dans l'autre. Andrew Liddle et Marina Cortês, de l'Université d'Édimbourg, proposent une explication à ce déséquilibre : à très grande échelle, l'Univers aurait une géométrie hyperbolique.



Les fluctuations du rayonnement du fond diffus cosmologique. La ligne noire sépare deux hémisphères pour lesquels la distribution des fluctuations est significativement différente.

Les cosmologistes ont étudié un scénario d'inflation – l'expansion rapide de l'Univers alors âgé de moins d'une seconde – qui explique l'asymétrie des observations. L'inflation est contrôlée par la présence d'un champ dit scalaire, l'inflaton. Celui-ci assure que la densité de matière et le rayonnement du fond diffus sont homogènes. Mais un champ scalaire supplémentaire, le curvaton, pourrait influer légèrement sur la distribution de matière à très grande échelle.

A. Liddle et M. Cortês ont relié les propriétés du curvaton à la courbure de l'Univers. Un rayon de courbure négatif, qui correspond à une géométrie hyperbolique (comme la surface d'une selle de cheval), permettrait d'expliquer les asymétries constatées dans les fluctuations du fond diffus cosmologique. Et si ce rayon est supérieur à celui de l'Univers observable, on a localement une géométrie euclidienne, comme observé. Reste à valider ce modèle, peut-être avec les données de *Planck*.

→ Sean Bailly

Phys. Rev. Lett., vol. 111, 111302, 2013

En bref

D'anciens fleuves au Sahara

Selon une étude germano-britannique, trois grands systèmes fluviaux auraient coulé dans le Sahara, du Sud au Nord, il y a environ 100000 ans. Les chercheurs ont combiné des simulations paléoclimatiques avec des données topographiques et hydrologiques (infiltration, évaporation) pour reconstituer l'écoulement ancien des eaux de surface. Ces systèmes auraient constitué des « couloirs verts » qui favorisaient les migrations vers les régions méditerranéennes fertiles révélées par les données archéologiques.

Quand le rouge devient noir

Le pigment vermillon (du sulfure de mercure) a été très utilisé en peinture. Or il s'assombrit à la lumière. Deux hypothèses ont été avancées : soit la structure cristalline du pigment change, ce qui noircit ce dernier, soit une réaction avec les ions chlore de l'air conduit à la formation de gouttelettes métalliques de mercure. Karolien De Wael, de l'Université d'Anvers, et ses collègues ont montré que c'est le mercure qui est le vrai coupable.

Neuf espèces en une

Ce que l'on pensait être une seule espèce de crustacé, *Eurythenes gryllus*, est en fait un complexe de neuf espèces. C'est ce qu'a montré l'équipe germano-belge de Charlotte Havermans et ses collègues à partir d'analyses génétiques et morphologiques. La plupart de ces espèces (qui mesurent jusqu'à 14 centimètres) ont des aires de répartition plus ou moins restreintes et sont donc plus vulnérables qu'une seule espèce présente un peu partout. Sept des neuf espèces vivent à une profondeur supérieure à 3000 mètres, ce qui illustre la richesse de la biodiversité des abysses, longtemps sous-estimée.

En bref

Une île sur dix submergée ?

Jusqu'à 12 pour cent des îles françaises risquent d'être englouties par les eaux d'ici 2100, en raison de l'élévation du niveau des mers provoquée par le réchauffement climatique. C'est la conclusion d'une étude du CNRS, qui a analysé les conséquences d'une élévation de un à trois mètres de ce niveau, en prenant en compte le relief des îles. Quelque 300 espèces endémiques, en majorité des plantes, pourraient alors disparaître en raison de l'immersion totale de leur aire de répartition.

Orangs-outans planificateurs

On a longtemps cru la capacité de planification réservée à l'espèce humaine. Après plusieurs observations de cette capacité chez certains animaux captifs, des chercheurs suisses l'ont aussi décelée chez des orangs-outans sauvages de Sumatra. Ces singes émettent une catégorie particulière de cris, de préférence dans la direction de leurs futurs déplacements, même quand ceux-ci se produisent le jour suivant. Planifier leurs déplacements leur permettrait de les signaler aux femelles et aux rivaux, de rejoindre une source de nourriture connue, etc.

Astrochimie

Des acides aminés produits par les impacts de comètes

De la matière organique se retrouve partout dans l'Univers, du milieu interstellaire jusqu'aux planètes, en passant par les comètes. Parmi les composés carbonés qui forment la matière organique, les acides aminés sont les précurseurs des protéines – des molécules indispensables à la vie telle que nous la connaissons.

Il existe de nombreuses façons de former des acides aminés, à commencer par l'expérience de Stanley Miller qui, en 1953, en synthétisa dans une enceinte censée reproduire les conditions de la Terre primitive. Zita Martins, de l'*Imperial College* à Londres, et ses collègues ont étudié un nouveau mécanisme de synthèse des précurseurs d'acides aminés les plus simples : il serait à l'œuvre lors de la collision d'une comète avec une planète si l'un ou l'autre de ces corps est glacé.

À l'aide d'un canon, les chercheurs ont bombardé de la glace d'eau avec un projectile en métal lancé à sept kilomètres par seconde. La glace contenait, en proportions parfaitement contrôlées, les éléments typiques présents à la surface des planètes ou lunes glacées (telle Encelade, une lune de Saturne) : de l'ammoniac, du dioxyde de carbone et du méthanol. Après le choc et l'évaporation de l'eau, les résidus organiques ont été analysés. Les chercheurs ont trouvé des traces d'acides aminés, la glycine et l'alanine. D'après eux, la pression du choc et la chaleur libérée par l'impact ont rendu possible la synthèse de composés organiques tels que des composés carbonylés, puis du cyanure d'hydrogène. L'hydrolyse de ce dernier conduit à la formation de glycine.



L'impact de comètes sur la surface glacée d'Encelade (une lune de Saturne) pourrait y former des acides aminés. C'est ce que laissent penser des expériences de laboratoire.

→ S. B.

Z. Martins et al., Nature Geoscience, en ligne le 15 septembre 2013

Parasitologie

Tiques et maladie de Lyme, une coévolution

L'évolution est souvent une coévolution. Coralie Herrmann, de l'Université de Neuchâtel, en Suisse, en a étudié un exemple frappant : *Borrelia burgdorferi*, la bactérie responsable de la maladie de Lyme, engraisse son vecteur, la tique !

En Europe, c'est la tique de l'espèce *Ixodes ricinus* qui est responsable de la plupart des infections par la maladie de Lyme (jusqu'à 15 000 cas par an en France). Entre le printemps et l'automne, cet acarien se perche sur la végétation basse dans l'attente du passage d'un mammifère, sur lequel il se laisse tomber pour lui sucer le sang. Très sensible à la sécheresse, il doit régulièrement quitter son perchoir pour aller se réhydrater dans l'humus.

Entre 2010 et 2013, C. Herrmann a étudié la résistance de 1 500 tiques de l'espèce *Ixodes ricinus* placées dans des enceintes où régnaient différents taux d'humidité (13, 32, 51,5, 61 et 89 pour cent) et températures (12,5 et 25 °C). Elle a observé que certains individus demeuraient dans des endroits secs, tandis que d'autres fuyaient à la recherche d'humidité. Des tests génétiques ont révélé que les tiques les plus résistantes étaient infectées par *Borrelia burgdorferi*. Avec l'aide de ses collègues, C. Herrmann a alors mis au point un protocole pour quantifier la masse de graisse d'une tique et, à la faveur d'une nouvelle expérience, a prouvé que les réserves des tiques infectées sont 12,1 pour cent supérieures à celles des tiques saines – ce qui est très significatif chez des acariens aussi vulnérables aux conditions sèches. Manifestement, *Borrelia burgdorferi* agit sur le métabolisme de son vecteur d'une façon qui améliore sa survie et, par là, augmente ses chances d'infecter un hôte, un humain par exemple.



Cette tique *(Ixodes ricinus)* s'est fixée sous l'œil d'un chien pour le parasiter.

→ F. S.

International Journal of Parasitology, vol. 43 (6), pp. 477-483, 2013

© Pour la Science - n° 433 - Novembre 2013

PRESSES POLYTECHNIQUES ET UNIVERSITAIRES ROMANDES

Editeur scientifique de référence depuis 1980



DISPARITION PROGRAMMÉE

Yves Bolognini, Marielle Stamm

Tous les ordinateurs sans lesquels les prouesses technologiques et informatiques actuelles n'auraient jamais vu le jour. Une autre histoire de l'informatique, richement illustrée et accessible à tous.

204 p., ISBN 978-2-88074-968-2



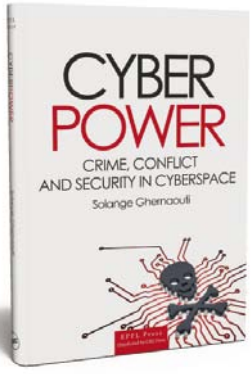
SYSTÈMES COMPLEXES

Une introduction par la pratique

Sébastien Collard, Sébastien Verel, Manuel Clergue

Les systèmes complexes sont devenus omniprésents; cet ouvrage pour non spécialistes expose comment les modéliser et les contrôler.

320 p., ISBN 978-2-88074-991-0



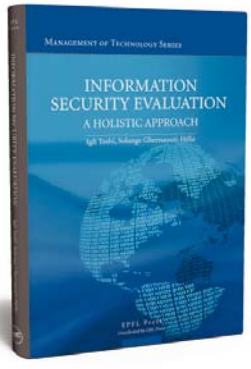
CYBERPOWER

Crime, Conflict and Security in Cyberspace

Solange Ghernaouti

This book explains the stakes of cyberpower in terms of cybercriminality, cyberterrorism, cyberconflicts and cyberwarfare. It evaluates the impacts on, and consequences for, individuals, organisations and states.

472 p., ISBN 978-2-940222-66-7



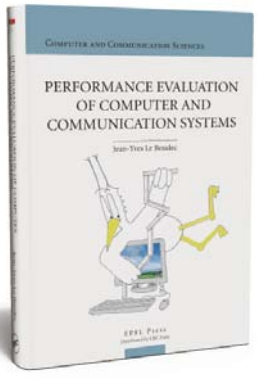
INFORMATION SECURITY EVALUATION

A Holistic Approach

Igli Tashi, Solange Ghernaouti

This book proposes a model to holistically assess all dimensions of security in order to minimize the likehood that a given threat will exploit the weakest link.

220 p., ISBN 978-2-940222-53-7



PERFORMANCE EVALUATION OF COMPUTER AND COMMUNICATION SYSTEMS

Jean-Yves Le Boudec

Performance evaluation is a critical stage of software and hardware-system development that every computer engineer and scientist should master.

420 p., ISBN 978-2-940222-40-7

NOUVEAUTÉS

ET OUVRAGES DE RÉFÉRENCE

www.ppur.org
www.epflpress.com

Ouvrages disponibles en librairie

LES FONDAMENTALES

Le forum du CNRS

En association avec Le Monde

PARIS, 14, 15 ET 16 NOVEMBRE 2013
RENDEZ-VOUS AVEC LES SCIENCES À LA SORBONNE



dépasser les frontières

Ministère de l'Enseignement supérieur et de la Recherche

île de France

Mairie de Paris

AIR LIQUIDE

SAFRAN

Fondation EADS

LVMH Recherche Parfums & Cosmétiques

casden

ABG L'intelligence

immédiats

pour la Science

Astrophysique

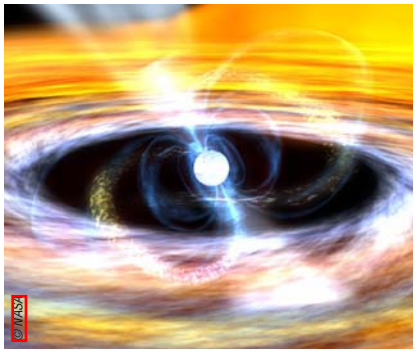
Une étoile à neutrons versatile

Étrange : l'étoile à neutrons PSR J1824-2452I émet alternativement, à quelques semaines d'intervales, des ondes radio et un rayonnement X. L'équipe dirigée par Alessandro Papitto, de l'Institut de sciences spatiales de Barcelone, explique ce comportement versatile par le fait que l'étoile appartient à un système binaire dont l'étoile compagnon est de faible masse, ce qui ferait osciller le système entre deux régimes très différents.

L'étoile à neutrons, d'une masse supérieure à celle du Soleil pour un rayon de dix kilomètres, attire la matière de son compagnon. Mais le flux de matière étant faible, le champ magnétique de l'étoile à neutrons serait assez intense pour empêcher, dans un premier temps, la matière de s'effondrer sur l'étoile. Autrement dit, la matière arrachée à son compagnon est maintenue à une certaine distance, ce qui l'empêche d'être accélérée et de chauffer. Dans ce régime, l'étoile à neutrons se comporte comme un pulsar ordinaire, qui émet des impulsions radio avec une période de l'ordre de la milliseconde. Au bout d'un certain temps, l'accumulation de matière du compagnon finit par vaincre le champ magnétique. Se forme alors un disque d'accrétion qui s'échauffe et émet un rayonnement X. Jusqu'à ce que la matière en accrétion se raréfie et que le cycle reprenne...

→ S. B.

Nature, vol. 501, pp. 517-520, 2013



L'étoile PSR J1824-2452I oscille entre un régime d'accrétion de matière, avec émission de rayons X, et un régime de pulsar radio.

250 mètres : c'est la profondeur de certains canaux détectés par satellite et par radar sous la banquise de l'Antarctique et creusés par les eaux de fonte.

Microbiologie

Des bactéries pour combattre l'obésité ?

La flore intestinale de personnes minces, transférée à des souris, protège ces dernières de l'obésité. C'est ce qu'a montré une équipe internationale dirigée par Jeffrey Gordon, de l'Université de Washington, aux États-Unis.

On savait que certains microbiotes – les bactéries de la flore intestinale – favorisent le développement de l'obésité. Les biologistes ont précisé lesquels en introduisant, chez la souris, des microbiotes de personnes obèses et minces – des jumelles, pour que les situations soient comparables. Ils ont transféré le microbiote de chaque jumelle dans l'intestin de souris stériles, dépourvues de microbiote. Les souris qui avaient reçu le microbiote de la jumelle obèse ont grossi, contrairement aux autres.

Était-ce le microbiote provenant de la personne obèse qui causait l'obésité ou, au contraire, celui de la personne mince qui protégeait de l'obésité ? Pour le savoir, les biologistes ont fait cohabiter de nouvelles souris portant chacune le microbiote d'une jumelle. En groupe, les souris ont l'habitude de manger les matières fécales de leurs congénères et, par conséquent, d'ingérer leur microbiote.

La cohabitation débutait dès la stabilisation de leur flore intestinale, cinq jours après l'inoculation des microbiotes humains. Après dix jours de cohabitation, les souris portant le microbiote de la jumelle obèse n'avaient pas grossi et leur microbiote avait été colonisé par celui de leurs congénères, provenant de la jumelle mince. Toutefois, ce n'était le cas que lorsque leur régime alimentaire était riche en fibres et pauvre en graisses. Dans le cas contraire, leur microbiote n'était pas colonisé et elles devenaient obèses. En revanche, la cohabitation est restée sans effet sur les souris ayant reçu le microbiote de la jumelle mince : leur microbiote n'a pas été colonisé par celui provenant de la jumelle obèse, et elles sont restées minces.

Qu'en conclure ? D'une part, que le microbiote de la jumelle mince protège les souris de celui de la jumelle obèse. D'autre part, que le microbiote de la jumelle mince peut « l'emporter » contre le microbiote de la jumelle obèse, à condition que la souris qui porte ce dernier suive un régime adéquat. Chez l'homme, ces résultats suggèrent que le microbiote des personnes minces les protège de l'obésité. Et qu'il sera peut-être possible, un jour, de prémunir contre l'obésité par un régime adéquat et l'injection d'un cocktail de bactéries conçu à partir du microbiote de personnes minces.

→ Marie-Neige Cordonnier

V. K. Ridaura et al. / A. W. Walker et J. Parkhill, Science, en ligne le 6 septembre 2013



En bref

Cinquième rapport du GIEC

Élaborés tous les six ou sept ans, les rapports du Groupe intergouvernemental d'experts sur l'évolution du climat (GIEC) dressent le bilan des connaissances. Le premier volet du cinquième rapport a été adopté le 27 septembre. Il confirme, voire accentue, les estimations précédentes sur l'impact des activités humaines, ainsi que sur l'ampleur du réchauffement climatique et des phénomènes associés : hausse du niveau des mers, fonte des glaces, augmentation de la fréquence des événements climatiques violents…

Pas de sexe en cas d'orage

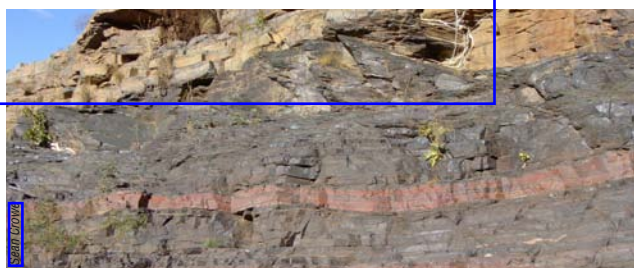
Les vents violents et les fortes pluies sont un danger pour les insectes. Ces espèces anticipent-elles les tempêtes ? Ana Cristina Pellegrino, de l'Université de São Paolo, et ses collègues ont étudié les modifications du comportement d'accouplement chez trois espèces d'insectes en cas d'orage. Les insectes se révèlent capables de détecter une baisse importante de la pression atmosphérique, un signe d'orage. De plus, les mâles réagissent moins aux phéromones des femelles et les signaux d'appel de ces dernières sont aussi réduits.

Géologie

De l'oxygène atmosphérique très ancien

Il y a trois milliards d'années, l'atmosphère contenait déjà du dioxygène, a montré Sean Crowe, de l'Université du Danemark du Sud. Les 21 pour cent d'oxygène atmosphérique actuels remontent, pense-t-on, à la « Grande oxydation ». Il s'agit d'une accumulation rapide dans l'air du dioxygène produit par la photosynthèse des cyanobactéries. Cet événement a débuté il y a 2,3 milliards d'années, lorsque le fer marin qui fixait l'oxygène s'est épuisé. Mais à quelle époque l'atmosphère a commencé à contenir de l'oxygène ?

Pour le savoir, l'équipe de S. Crowe a analysé les concentrations de chrome dans des échantillons de paléosols sud-africains vieux de trois milliards d'années. En présence d'oxygène, l'altération d'une roche provoque l'oxydation du chrome III en oxyde de chrome IV, qui est très soluble, donc emporté par le ruissellement.



Les paléosols de la formation de Pongola, en Afrique du Sud, comportent schistes verts et rouges vieux de trois milliards d'années.

Cette oxydation étant plus aisée pour le chrome 53 (l'un des quatre isotopes de l'élément), celui-ci se raréfie dans les roches lavées en présence d'oxygène. Or l'analyse isotopique a montré moins de chrome 53 dans les paléosols sud-africains que dans la croûte archéenne, ce qui prouve qu'il y avait bien de l'oxygène dans l'air quand ces sols ont été érodés…

Combien ? Assez pour oxyder le fer II : sinon, celui-ci aurait réduit le chrome IV et empêché sa solubilisation. À partir de la concentration de fer dans leurs paléosols, les chercheurs ont estimé un niveau minimum d'oxygène égal à trois dix-millièmes du niveau actuel. Une valeur plausible, car des stromatolithes, des roches calcaires construites par des communautés de bactéries où dominent les cyanobactéries, sont attestées depuis 3,5 milliards d'années.

→ F. S.

Nature, vol. 501, pp. 535-539, 2013

Archéologie

Maille à partir avec les Germains

Le champ de bataille antique le mieux conservé d'Europe se situe en Basse Saxe, près du village de Harriehausen. L'équipe de Michael Meyer, de l'Université de Berlin le fouille ; elle vient d'y retrouver une centaine de fragments bien conservés de la cotte de maille d'un soldat romain.

Des pièces d'argent retrouvées sur place fixent la date de la bataille vers 235-236 de notre ère. Les chroniqueurs de l'époque rapportent que l'empereur Maximin Ier le Thrace avait alors ordonné une vaste opération « à quelque 400 miles » de la frontière fortifiée de l'Empire, afin de punir les attaques germaniques. C'est donc au retour de cette expédition en territoire suève que plusieurs milliers de Romains, dont des légionnaires, se sont retrouvés dans une embuscade tendue entre deux collines par une troupe germaine. Au lieu de s'engager dans le piège, les soldats ont forcé le passage par la colline de gauche. De nombreuses flèches de scorpion (grande arbalète sur pied) retrouvées sur place prouvent qu'ils ont commencé par une vigoureuse préparation d'artillerie. Par ailleurs, les archéologues ont déjà retrouvé 2700 armes et autres objets, ce qui suggère qu'une fois le passage ouvert, la colonne n'a pas pu s'attarder pour ramasser son matériel…

L'abandon d'une cotte de maille confirme cette hâte. Constituée de milliers de maillons, une telle protection demandait tant de travail et coûtait si cher qu'il n'est guère pensable que des soldats l'aient abandonnée sans raison. L'apparence de ses fragments résiduels suggère en outre qu'elle a été arrachée à la va-vite, à la lisière de la bataille – sans doute par une escouade qui débarrassait l'un de ses blessés des 12 kilogrammes de sa cotte pour le porter en courant !

→ F. S.

www.roemerschlachtamharzhorn.de



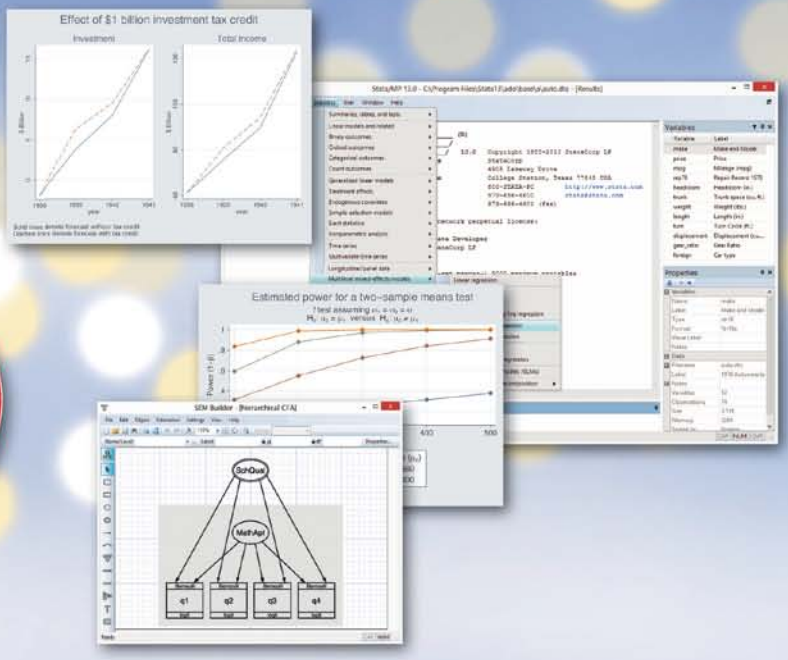
L'un des fragments de la cotte de maille romaine retrouvée sur la colline du Harzhorn, en Basse Saxe, sur le lieu d'une bataille entre Romains et Germains.

© Pour la Science - n° 433 - Novembre 2013

STATA® 13



Pour Windows, Macintosh et Linux



STATA 13 propose 77 nouvelles fonctionnalités

Notamment :

- SEM généralisée
- Estimation de l'effet de traitement
- Plus grand choix de modèles
- Modèles multi-niveaux à effets fixes et aléatoires
- Gestionnaire de projet intégré
- Estimation de modèle à effets aléatoires en données de panel

En Suisse :

SCIENTIFIC SOLUTIONS
YOUR RESEARCH STARTS WITH US

www.scientific-solutions.ch/stata

SCIENTIFIC SOLUTIONS - Distributeur Officiel • Rue du Midi 2 • CH-1009 Pully • Switzerland
Tél. : +41(0)21 711 15 20 • Fax : +41(0)21 711 15 21
info@scientific-solutions.ch • www.scientific-solutions.ch

En France et en Belgique :

RITME
NOTRE MÉTIER : FACILITER LE VÔTRE

www.ritme.com/stata

RITME - Distributeur Officiel • 34, bd Haussmann • 75009 Paris • France
Tél. : +33 (0)1 42 46 00 42 • Fax : +33 (0)1 42 46 00 33
info@ritme.com • www.ritme.com

Sauf indication contraire : logiciel et documentation en anglais, et livraison par téléchargement ou par CD-Rom. © 2013 RITME, © 2013 SCIENTIFIC SOLUTIONS. Stata est une marque déposée de StataCorp LP. Toutes les marques déposées sont la propriété de leurs sociétés respectives.

En bref

Du pollen bien plus ancien

Il n'existe pas de consensus sur la date d'apparition des premières plantes à fleurs. Les traces les plus anciennes sont sous la forme de pollen fossilisé : abondant, petit et robuste, il se conserve mieux que les feuilles et les fleurs. Les plus anciens fossiles connus dataient de 140 millions d'années. Or Peter Hochuli, de l'Université de Zurich, a découvert des fossiles de pollen provenant de plantes à fleurs qui seraient âgés de 240 millions d'années.

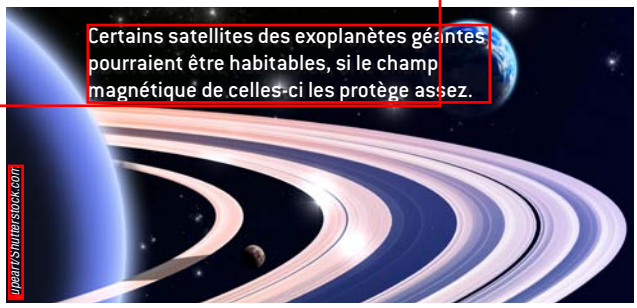
Antidécalage horaire

Soumise à un décalage horaire, l'horloge biologique interne met du temps à se synchroniser sur le nouveau cycle du jour et de la nuit. Cette horloge est coordonnée par le noyau suprachiasmatique, une aire cérébrale connectée à la rétine. À Kyoto, Yoshiaki Yamaguchi et ses collègues ont bloqué chez des souris l'activité de la vasopressine, hormone qui synchronise les neurones de cette aire. Les animaux ont récupéré plus vite d'un décalage horaire, sans doute parce que les neurones recevant le signal rétinien étaient moins liés aux autres et donc plus réactifs. Une piste antidécalage horaire ?

Planétologie

Les exolunes sont-elles habitables ?

Plus de 900 exoplanètes ont été détectées et plusieurs milliers d'autres attendent la confirmation de leur découverte. Certaines sont en orbite dans la zone habitable de leur étoile (la zone où de l'eau liquide peut exister), mais la plupart sont des géantes gazeuses, impropres à la vie. Ont-elles tout de même des satellites, ou exolunes, qui pourraient abriter la vie ? René Heller, de l'Université McMaster, au Canada, et son collègue ont voulu savoir si les lunes extrasolaires sont protégées par un champ magnétique, condition nécessaire à leur habitabilité.



Pour être habitable, une exolune ne doit être ni trop proche ni trop éloignée de sa planète mère. Trop proche, elle est soumise à des effets de marée et à un rayonnement trop intenses ; trop éloignée, elle est en dehors de la magnétosphère de la planète (la zone dominée par son champ magnétique) et n'est pas protégée des violents flux de particules chargées qui parcourent l'espace.

Les astrophysiciens ont estimé le champ magnétique des exoplanètes de taille comprise entre celle de Neptune et celle de Jupiter, à partir des données recueillies sur les planètes du Système solaire. Les plus petites ont peu de chances d'avoir une exolune habitable, car la zone protégée par leur magnétosphère est trop restreinte. En revanche, pour une planète de la taille de Jupiter, dont la magnétosphère est plus étendue, des lunes habitables pourraient exister à une distance comprise entre 5 et 20 fois le rayon de la planète.

Ces résultats sont à affiner, puisque le champ magnétique dépend notamment de la composition des planètes, mal connue. En outre, l'intérieur des magnétosphères n'est pas totalement exempt de particules chargées : certaines sont piégées à l'intérieur et leur effet sur l'habitabilité reste à évaluer. Mais l'espoir est permis !

→ Guillaume Jacquemont

R. Heller et J. Zuluaga, Astrophysical Journal Letters, à paraître

Paléontologie

Un placoderme au sourire ironique

Un fossile de poisson datant de la fin du Silurien (443,4 à 419,2 millions d'années) renverse l'ordre qui régnait dans la phylogénie des poissons. Découvert par Min Zhu, de l'Académie des sciences chinoises, et des collègues, *Entelognathus* a une mâchoire de poisson osseux, sans en être un…

Entelognathus fait partie d'un ordre éteint de poissons cuirassés remontant au début du Silurien : les placodermes. Toutefois, il est doté d'une mâchoire comportant un os maxillaire, un os prémaxillaire et un os dentaire (sans dents), comme celles des poissons osseux (truites, harengs…), tandis que sa gorge et ses branchies sont couvertes de plaques osseuses, elles aussi comparables à celles des poissons osseux. Or, jusqu'à présent, la mâchoire composite des poissons osseux était considérée comme caractéristique de leur ordre. Ce n'est plus le cas si un placoderme était pourvu d'une mâchoire comparable !

Cette mâchoire annonçait-elle les poissons osseux ? Difficile à dire à ce stade, mais une chose est claire : *Entelognathus* met du désordre dans la phylogénie des poissons. Il est si distinct des acanthodiens, ordre éteint de poissons à petites écailles et grandes épines, que ces derniers apparaissent désormais plus proches des poissons cartilagineux (raies, requins), alors qu'on les rattachait aux poissons osseux. Cela suggère que les requins dérivent des acanthodiens et relativise l'idée répandue qu'ils seraient plus primitifs que les poissons osseux. Or cette impression poussait les paléontologues à s'imaginer l'ancêtre commun des poissons osseux et des poissons cartilagineux avec une gueule de requin. *Entelognathus*, qui en a une, tout en ayant une mâchoire de poisson osseux, montre qu'il ne s'agissait que d'un préjugé…

→ F. S.

Nature, en ligne le 25 septembre 2013



Dans une mer devenue aujourd'hui une région de la Chine, le placoderme *Entelognathus* ouvre son étonnante mâchoire osseuse. En haut, le fossile de cette mâchoire.

© Pour la Science - n° 433 - Novembre 2013

Avec plus d'1 million de Sociétaires, on peut déplacer des montagnes



Quand une banque tire sa force de l'esprit coopératif, elle s'appuie sur des valeurs de solidarité, d'écoute et de confiance. Créée par des enseignants, la CASDEN s'engage ainsi auprès de plus d'un million de Sociétaires à réinvestir leur épargne dans le financement des projets de chacun.

Rejoignez-nous sur **casden.fr** ou contactez-nous au **0826 824 400***

*Accueil téléphonique ouvert de 8h30 à 18h30 du lundi au vendredi (0,15€ TTC/min en France métropolitaine).

BANQUE POPULAIRE

L'offre CASDEN est disponible en Délégations Départementales et également dans le Réseau Banque Populaire.

casden
BANQUE POPULAIRE

CASDEN, la banque coopérative de l'éducation, de la recherche et de la culture

CASDEN Banque Populaire | Société Anonyme Coopérative de Banque Populaire à capital variable | Siège social : 91 Cours des Roches | 77186 Noisiel | Siret n° 784 275 778 00842 - RCS Meaux. | Immatriculation ORIAS n° 07 027 138
BPCE – Société anonyme à directoire et conseil de surveillance au capital de 1 467 226 960 € | Siège social : 50 avenue Pierre Mendès France – 75201 Paris Cedex 13 | RCS PARIS n° 493 455 042 | Immatriculation ORIAS n° 08 045 100
BROCA & WERNICKE | Illustration : Kiki Lloret

Neurophysiologie

Entendre avec la bouche

Les grenouilles de Gardiner *(Sechellophryne gardineri)*, petits batraciens endémiques des îles de Mahé et de Silhouette, dans les Seychelles, n'ont pas de tympan. Pourtant, elles entendent : elles répondent à des coassements diffusés par un haut-parleur. Renaud Boistel, de l'Université de Poitiers, et ses collègues du CNRS ont montré que la vibration sonore est transmise à l'oreille interne (qui la traduit en signaux nerveux) *via* la bouche de l'animal.

Les chercheurs ont reconstitué une vue tridimensionnelle du squelette et des tissus mous de la grenouille grâce à des techniques d'imagerie par rayons X. Ils ont ainsi invalidé plusieurs voies de transmission envisagées et montré que la configuration des tissus séparant la bouche de l'oreille interne est favorable à la propagation des vibrations sonores. Ils ont ensuite validé la capacité de ces tissus à transmettre le son, grâce à une simulation sur ordinateur du comportement vibratoire de la tête de la grenouille.

→ G. J.

R. Boistel et al., PNAS, en ligne le 2 septembre 2013

Énergie

Du biocarburant à pandas ?

La compétition avec les usages alimentaires freinant le développement des biocarburants, on cherche à exploiter la lignocellulose de la partie non consommable des plantes (paille, résidus agricoles, etc.). L'équipe d'Ashli Brown, de l'Université du Wisconsin-Madison, a identifié des bactéries intestinales de panda qui pourraient aider à convertir ce composé en carburant.

La transformation de la lignocellulose requiert des procédés coûteux et énergivores. Or le panda se nourrit de bambous, qu'il digère vite : sa flore intestinale doit donc transformer efficacement ce composé. Les chercheurs, qui y traquent des micro-organismes exploitables depuis plusieurs années, ont annoncé la découverte de plus de 40 bactéries potentiellement intéressantes.

→ G. J.

246e conférence nationale de la Société américaine de chimie, www.acs.org/content/acs/en/meetings.html

Biologie

Des cornes peu coûteuses

Le scarabée-rhinocéros japonais *Trypoxylus dichotomus* est un grand coléoptère asiatique dont le mâle présente deux cornes remarquables par leur taille *(voir la photographie)*. Ces attributs, mi-armes, mi-ornements, lui servent à lutter contre d'autres mâles dans la compétition pour s'attirer la faveur des femelles. Une étude d'Erin McCullough et Douglas Emlen, de l'Université du Montana, aux États-Unis, indique que le développement de ces grandes cornes n'est pas associé à un surcoût pour l'organisme qui les porte.

En comparant de nombreux individus, les deux biologistes ont montré que le développement des cornes ne détourne pas de ressources au détriment d'autres parties du corps (ailes, yeux, pattes antérieures, organes génitaux), et qu'il n'y a pas d'impact sur l'immunité de l'insecte, ni sur sa probabilité de survie dans la nature.

Ce résultat surprend : il va à l'encontre de ce que supposent généralement les théoriciens de la sélection sexuelle. Selon eux, l'existence d'un surcoût dû à de tels attributs est nécessaire pour que ces derniers soient des indicateurs fiables de la qualité du mâle, donc pour qu'ils aient une importance aux yeux des femelles et des autres mâles.

→ M. M.

E. L. McCullough et D. J. Emlen, Animal Behaviour, en ligne le 26 septembre 2013

Un *Trypoxylus dichotomus* mâle



© JM

DERNIÈRE minute...

VOIR LES GÈNES EN MOUVEMENT

Visualiser les mouvements des gènes dans des cellules vivantes : telle est la prouesse réalisée par Yusuke Miyanari et deux collègues, à Strasbourg. En ciblant une séquence spécifique d'un gène à l'aide d'une protéine de synthèse couplée à une protéine fluorescente, les biologistes ont observé *in vivo* la dynamique du gène au fil des divisions cellulaires. Transposable à toute séquence génétique, cette technique devrait aider à comprendre le rôle de la dynamique du génome dans l'expression des gènes.

NITRURE DE BORE : UN BON ANTIOXYDANT

Le nitrure de bore (BN), dans sa structure cristalline hexagonale, forme des feuillets d'épaisseur atomique. Or Zheng Liu et ses collègues, à l'Université Rice, aux États-Unis, viennent de montrer que des films d'épaisseur nanométrique de ce matériau procurent une excellente protection contre l'oxydation, et ce jusqu'à des températures élevées. Ainsi, du nickel recouvert d'un tel film, que l'on peut fabriquer par dépôt en phase vapeur, résiste à l'oxydation jusqu'à des températures de 1 100 °C. D'innombrables applications se profilent...

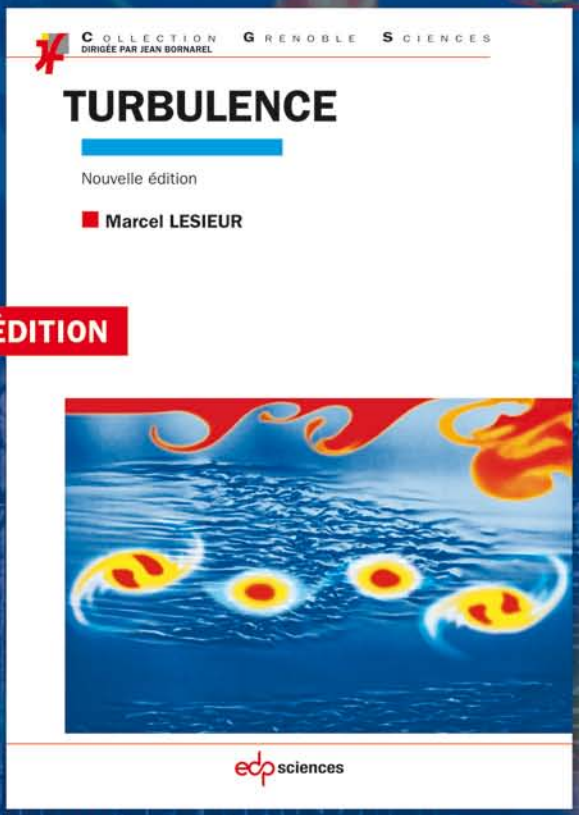
Retrouvez plus d'actualités et toutes les références sur www.pourlascience.fr

© Pour la Science - n° 433 - Novembre 2013

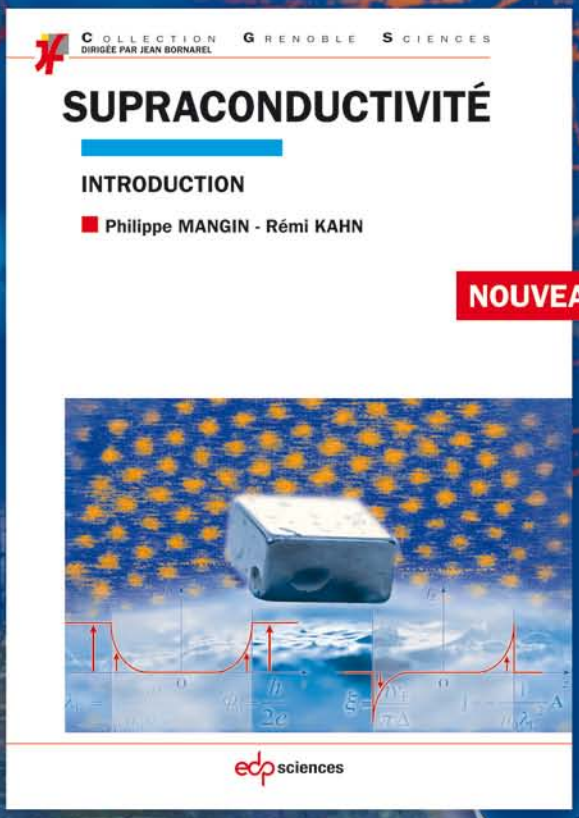
GRENOBLE SCIENCES

ET LES SCIENCES DE LA MATIÈRE

Comment simuler la turbulence pour la comprendre. Un livre d'initiation au monde fascinant de la turbulence dans les fluides.



RÉÉDITION



NOUVEAUTÉ

La supraconductivité, une propriété de la matière qui fait rêver. Le manuel de référence de niveau Master.

ET D'AUTRES OUVRAGES À SUCCÈS



GRENOBLE SCIENCES

http://www.grenoble-sciences.com

Grenoble Sciences, Université Joseph Fourier, Bât. B de Phitem,
230 rue de la Physique – BP 53, 38041 Grenoble cedex 9
Tél. (33)4 76 51 46 95 – Fax. (33)4 76 51 45 79
Email : Grenoble.Sciences@ujf-grenoble.fr

Université Joseph Fourier GRENOBLE

Région RhôneAlpes

POINT DE VUE

ISET : un test génétique prénatal à encadrer

En ouvrant la voie à une caractérisation génétique du fœtus dès le début de la grossesse, la méthode de diagnostic prénatal ISET soulève de nouvelles questions éthiques.

Floriane BLANC

Breveté sous l'acronyme ISET *(Isolation by Size of Epithelial Tumor/Trophoblastic Cells)*, un nouveau procédé mis au point par l'équipe de Patrizia Paterlini-Bréchot à l'Hôpital Necker-Enfants malades, à Paris, est susceptible de bouleverser le suivi de grossesse. Validée cliniquement, la technique permet, grâce à une simple prise de sang de la femme enceinte et dès la cinquième semaine d'aménorrhée (absence de règles), de communiquer aux parents potentiels, en quatre jours environ, le diagnostic génétique du fœtus, pour l'heure concernant l'amyotrophie spinale et la mucoviscidose. Le test ISET est fondé sur l'étude du génome des cellules fœtales qui circulent dans le sang maternel. Cette méthode offre la possibilité de développer des tests diagnostiques non invasifs aussi fiables que l'amniocentèse pour toute pathologie reposant sur une anomalie chromosomique ou présentant un facteur génétique, et ce à un coût équivalent, voire inférieur, aux examens actuels.

L'ISET constitue un progrès dans le suivi de grossesse. La prochaine déclinaison envisagée du test, notamment, porte sur les trisomies, dont le dispositif actuel de dépistage est insatisfaisant à plusieurs titres. Ce dernier comporte deux étapes. La première est effectuée entre la 11e et la 14e semaine d'aménorrhée. De l'ordre du tri statistique, elle croise détection de marqueurs sanguins et mesure par échographie. de l'épaisseur d'une zone précise située sur la nuque du fœtus. Elle se révèle anxiogène, car elle place dans le groupe à risque des femmes enceintes de fœtus non porteurs d'une trisomie. En outre, elle ne dépiste pas certains cas, ainsi dévoilés seulement à la naissance. Lors de la seconde étape, les femmes incluses dans le groupe à risque se voient proposer une amniocentèse. Prélèvement du liquide qui entoure le fœtus, le liquide amniotique, la procédure provoque entre 0,5 à 1 pour cent de fausses-couches, selon un rapport publié par la Haute autorité de santé en 2007. Fiable et non invasif, l'ISET remplacerait utilement cette méthode.

POUR LA PREMIÈRE FOIS,
à l'issue d'un diagnostic prénatal réalisable dans un laboratoire médical, les parents pourraient se retrouver seuls à décider de poursuivre la grossesse ou de l'interrompre.

Nonobstant, le nouveau test renouvelle la question éthique du dépistage prénatal. En proposant la recherche de particularités génétiques, sans risque et par un acte – la prise de sang – que toute femme enceinte accomplit des dizaines de fois au cours de sa grossesse, et ce avant la fin du délai légal d'avortement, il donne la possibilité aux parents de choisir seuls les caractéristiques génétiques de leur progéniture.

Certes, la démarche générique d'interrompre ou de poursuivre une grossesse à la suite de la découverte d'une particularité génétique existe déjà. Mais, jusqu'à présent, elle était encadrée par les conseils génétiques et donc accompagnée par différents spécialistes du corps médical. Le nouveau test, lui, pourra être fait dès le début de la grossesse, bien avant le délai légal d'avortement. En d'autres termes, pour la première fois, à l'issue d'un diagnostic prénatal réalisable dans un laboratoire médical, les parents pourraient se retrouver seuls face au diagnostic, seuls à décider de poursuivre la grossesse ou de l'interrompre.

Plus encore, les perspectives de développement de l'ISET promettent que, tôt ou tard, tout aspirant à la parentalité pourra s'assurer que le fœtus n'est porteur d'aucune caractéristique génétique listée par la médecine comme facteur ou origine d'une quelconque particularité ou pathologie, qu'elle soit physique, physiologique ou mentale. Ne faut-il pas craindre que les futurs parents dressent seuls la liste des risques génétiques qu'ils n'entendent pas faire prendre à leur descendance ?

Car sous la pression des évolutions croisées de la médecine et de la société, tout parent d'aujourd'hui considère de son devoir que l'enfant soit en bonne santé. Mais qu'est-ce qu'un enfant en bonne santé ? L'environnement, familial, social, scolaire, de l'enfant n'est-il pas essentiel ? Soyons clairs : l'absence d'anomalie génétique n'est pas une garantie de bonne santé. Évidemment, tout parent aspire à ce que son enfant soit à l'abri de toute maladie grave, douloureuse ou invalidante, et ce d'autant plus que les structures et les mesures d'accompagnement du handicap sont en France largement insuffisantes.

© Pour la Science - n° 433 - Novembre 2013

Dès lors, quel futur parent ne sera pas tenté par ce test ? Mais, sans accompagnement, où chacun placera-t-il le curseur ? Quelles caractéristiques génétiques cibler ? Les maladies incurables, les maladies invalidantes, ou tout risque potentiel ?

Promis à la commercialisation cette année, l'ISET nécessite un accompagnement éthique et politique. Éthique, parce qu'aux sirènes de la médecine prédictive, proposant de supprimer le risque génétique, doit répondre la prise en compte des progrès de la médecine curative. Politique, parce que les futurs parents doivent être accompagnés dans leur choix par des conseils pluridisciplinaires qui ne soient pas uniquement composés de membres du corps médical. Sur cette question, leur réflexion nécessite que les faits leur soient exposés de façon claire et objective, dans un cadre qui sera le fruit d'une réflexion collective.

Dès maintenant, l'ISET exhorte donc une réflexion éthique, collective et pratique sur son usage. Il conviendrait que les instances politiques organisent rapidement une première commission pour évaluer l'opportunité et les conditions d'une telle commercialisation. Comment l'ISET s'insère-t-il dans le corpus actuel des lois de bioéthique ? Quelles extensions ou dispositions spécifiques faut-il mettre en place pour que son usage soit strictement encadré ? Cette commission, comme celles qui devront suivre, nécessairement réparties sur le territoire pour évaluer opinions et perspectives sur la question, ne devra pas être composée des seuls membres du corps médical, car l'avenir génétique de la communauté n'est pas seulement une question d'ordre médical. C'est avant tout une question de société.

L'ISET nous place aujourd'hui devant un choix, tant individuel que collectif. Quelle société voulons-nous construire ? Que signifierait ériger la génétique en critère absolu de la norme biomédicale ? Quelle construction sociale ce schéma engendrerait-il ? Sur ces questions bioéthiques, c'est toute la société qui doit être invitée à réfléchir. ■

Floriane BLANC est enseignante-chercheure contractuelle en épistémologie des sciences médicales.

Réagissez en direct à cet article sur www.pourlascience.fr

DÉVELOPPEMENT DURABLE

Énergies renouvelables : le coût de l'intermittence

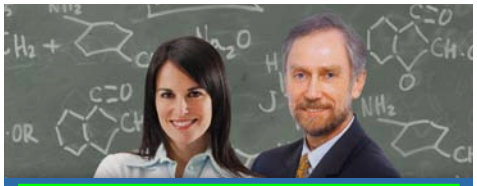
Le système électrique devra être adapté pour gérer l'intermittence de la production éolienne et solaire. Quel en sera le coût ?

Renaud CRASSOUS et Fabien ROQUES

Selon le Paquet énergie-climat adopté en 2008 par le Parlement européen, la France devra produire 23 pour cent de l'énergie totale qu'elle consomme par des sources renouvelables d'ici 2020. Pour y parvenir, l'effort portera en grande partie sur la production électrique, où la part de l'éolien et du solaire devra passer de 3,5 pour cent (en 2012) à 11 pour cent. Or, du fait des variations météorologiques locales et de l'alternance du jour et de la nuit, la puissance injectée sur le réseau par ces moyens de production peut changer vite et souvent. À l'échelle d'une région ou d'un pays, les variations instantanées sont lissées par un effet dit de foisonnement (les dispositifs sont assez nombreux pour que leurs variations instantanées se compensent en partie), mais sur une échelle de quelques heures, les variations restent importantes. Une telle proportion d'énergies intermittentes nécessitera alors d'importantes adaptations du système électrique, qui doit en permanence équilibrer l'énergie produite et celle soutirée par les consommateurs, en maintenant la fréquence et la tension dans un intervalle étroit. Le coût résultant sera-t-il acceptable ?

Le premier type d'adaptation visera à sécuriser la fourniture : il faudra toujours être capable de répondre à la demande, même s'il n'y a ni vent ni soleil. Plusieurs actions complémentaires seront menées : l'importation d'électricité depuis les pays voisins, la modulation de la consommation (par les réseaux dits intelligents, qui commencent à être déployés) et la construction de centrales électriques supplémentaires. Celles-ci ne serviront que peu de temps dans l'année et seront d'autant plus nombreuses que l'on a de moyens de production intermittents (ces derniers se substituant à ceux capables de garantir la fourniture d'électricité quelles que soient les conditions). Pour 10 à 20 pour cent d'éolien et de solaire, le surcoût associé, surtout dû à la construction des centrales supplémentaires, est estimé entre 5 et 15 euros par mégawattheure de production intermittente.

Les fluctuations de la production éolienne et solaire s'ajouteront à celles de la demande,



Connectez-vous au service d'offres d'emplois scientifiques sur

POUR LA SCIENCE

en partenariat avec naturejobs



- Chaque semaine, plus de **8000 offres d'emploi dans le secteur des sciences.**
- En un clic **la sélection des postes à pourvoir en France** et dans les pays francophones ou voisins.
- Une équipe Naturejobs spécialisée, **au service des recruteurs francophones.**

Contact : Muriel Lestringuez
E-mail : m.lestringuez@nature.com
Tél. : +44 20 7843 4994

POUR LA SCIENCE .fr

Le site de référence de l'actualité scientifique

si bien que le reste du parc devra faire face à des variations de production accrues et un peu plus imprévisibles. Cela implique une utilisation plus fréquente des moyens de production les plus flexibles (telles les turbines à gaz), dont le fonctionnement coûte plus cher. Les estimations du surcoût restent modérées, de l'ordre de deux à cinq euros par mégawattheure de production intermittente.

Le dernier type d'adaptation concerne les réseaux, qui devront être renforcés et étendus. Il s'agira de raccorder de multiples sources d'énergie parfois éloignées des centres de consommation (telles des éoliennes marines). Dans certains cas, d'importantes redistributions seront nécessaires : s'il y a du vent seulement dans le Sud du pays, par exemple, on doit pouvoir acheminer une partie suffisante de la production vers le Nord. Enfin, pour éviter de trop grandes variations de tension et de fréquence, il faudra équiper le réseau de moyens de régulation supplémentaires.

L'idée que des énergies renouvelables locales permettraient d'éviter des dépenses liées au réseau électrique est donc erronée, sauf dans certains cas où production et consommation varient de conserve (comme en Californie, où un plus fort ensoleillement entraîne l'augmentation à la fois de la production photovoltaïque et de la consommation électrique, conditionnée par les climatiseurs).

Le chiffrage des coûts associés au renforcement des réseaux est encore peu fiable. En France, les investissements nécessaires d'ici 2020 représenteront entre 3 et 10 euros par mégawattheure intermittent. En revanche, en Allemagne, ils seront de 30 à 50 euros par mégawattheure sur la même période, pour un objectif de 25 pour cent de la production électrique assuré par les énergies solaire et éolienne. Un tel surcoût ne peut être extrapolé directement aux autres pays, car chacun a ses spécificités, mais il pourrait indiquer des effets de seuil, mal identifiés pour l'instant : à partir d'un certain niveau d'intermittence, le coût du renforcement des réseaux augmenterait notablement.

Au total, l'insertion des productions solaire et éolienne dans le système électrique, à hauteur de 10 à 15 pour cent de la production totale, pourrait occasionner des surcoûts moyens de l'ordre de 10 à 30 euros par mégawattheure intermittent. En comparaison, le prix de détail de l'électricité est de l'ordre de 140 euros par mégawattheure en France. Les surcoûts subiront des variations propres à chaque pays, en fonction des capacités hydrauliques, de l'état du réseau initial, des interconnexions avec les pays voisins, etc.

Au-delà de 20 pour cent d'énergies renouvelables intermittentes, le coût pourrait augmenter rapidement, même si les estimations sont peu précises. Aujourd'hui, seul le Danemark est au-delà de cette proportion, puisque l'électricité d'origine éolienne y représente entre 30 et 40 pour cent de la production totale, mais c'est un cas particulier : la gestion de l'intermittence est assurée grâce à ses multiples connexions avec ses voisins, notamment la Suède et la Norvège. Ces pays ont de nombreux barrages hydrauliques et des stations de pompage, qui fournissent un moyen de stockage massif de l'électricité : elles pompent l'eau en altitude lors des périodes de surproduction et la font redescendre à travers des turbines hydroélectriques quand les besoins augmentent.

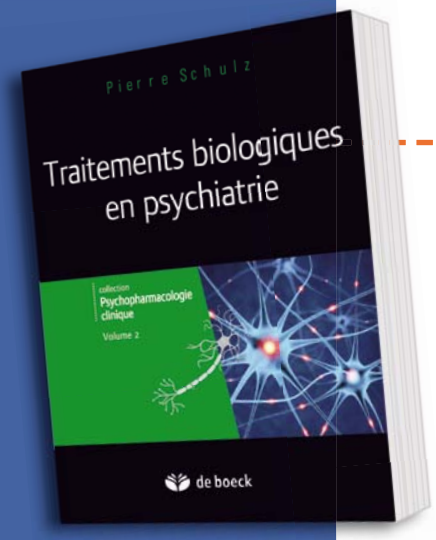
DES ÉNERGIES RENOUVELABLES locales ne permettent pas forcément d'éviter des dépenses liées au réseau.

L'intermittence de l'éolien et du solaire représente donc un nouveau défi technique et économique. D'ici 2020, il devrait être gérable en optimisant le déploiement de ces énergies (par exemple, en concentrant les éoliennes là où le réseau est déjà assez solide pour gérer l'intermittence sans être trop renforcé). Au-delà, il faudra réaliser des ruptures technologiques, par exemple sur le stockage ou sur les réseaux intelligents (qui permettront de mieux moduler la demande), pour limiter les surcoûts. ■

Renaud CRASSOUS est ingénieur et économiste à la Direction de la stratégie et de la prospective d'EDF. Fabien ROQUES est économiste au sein du cabinet de conseil Compass Lexecon.

© Pour la Science - n° 433 - Novembre 2013

L'EXCELLENCE EN SCIENCES



TRAITEMENTS BIOLOGIQUES EN PSYCHIATRIE
Pierre Schulz
Novembre 2013 • 800 pages • 62 €

■ Cet ouvrage propose une synthèse des connaissances contemporaines sur les traitements biologiques en psychiatrie, avec une sélection d'informations en neurosciences cliniques et des recommandations pour la pratique médicale.



CHIMIE ORGANIQUE
Jonathan Clayden, Nick Greeves, Stuart Warren
Traduction André Pousse
Août 2013 • 1280 pages • 120€

■ Un manuel de chimie organique rigoureux et captivant, qui permet à l'étudiant de comprendre comment on passe des principes fondamentaux aux synthèses les plus innovantes.



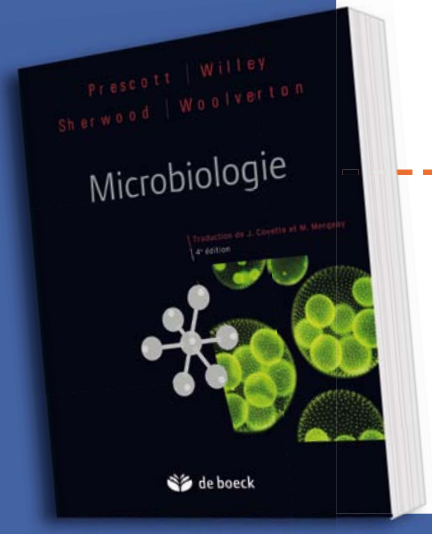
BIOSTATISTIQUE
Harvey J. Motulsky
Traduction : Michèle Dramaix-Wilmet
Septembre 2013 • 512 pages • 49 €

■ Un texte clair, accessible à tous, qui explique de manière facilement compréhensible les méthodes statistiques et détaille l'interprétation des résultats issus des analyses.



QUANTIQUE, FONDEMENTS ET APPLICATIONS
Avec 250 exercices et problèmes résolus
José-Philippe Perez, Robert Carles, Olivier Pujol
Septembre 2013 • 1120 pages • 54 €

■ Cet ouvrage met l'accent sur les fondements scientifiques et sur les applications concrètes et quantitatives de la physique quantique, en 20 leçons structurées, illustrées par de nombreux exemples et prolongées par 250 exercices et problèmes résolus.



MICROBIOLOGIE
Joanne M. Willey, Linda M. Sherwood, Christopher J. Woolverton
Traduction Jacques Coyette, Max Mergeay
Octobre 2013 • 1200 pages • 120 €

■ *Microbiologie* est la traduction de la 8e édition du *Prescott's Microbiology*, l'ouvrage américain de référence sur la microbiologie, décrite dans ses aspects fondamentaux, médicaux, écologiques, alimentaires et industriels.

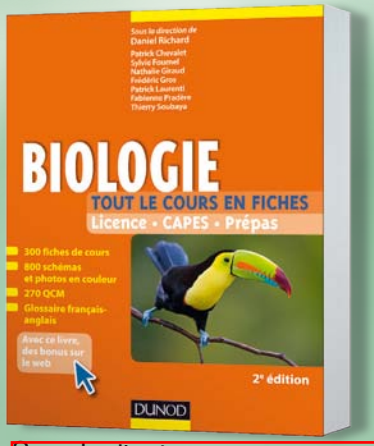
En librairie et sur **www.deboeck.fr**

de boeck

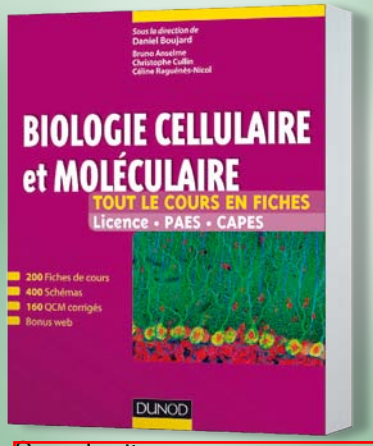
La rentrée des sciences chez DUNOD

Collection **Tout le cours en fiches**

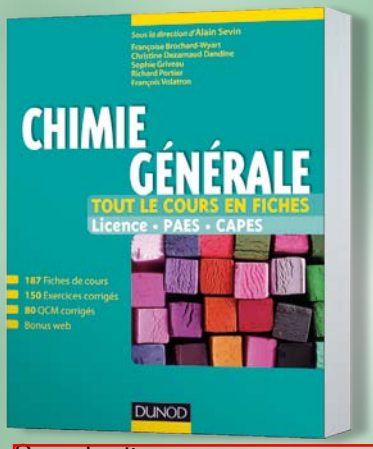
- Des fiches de cours synthétiques pour retenir l'essentiel
- Des QCM pour s'évaluer
- Des exercices et des sujets de synthèse pour s'entraîner
- Des focus pour compléter ses connaissances
- Des ressources numériques complémentaires



Sous la dir. de **D. Richard**, ***et al.***
9782100582235 • **45 €**



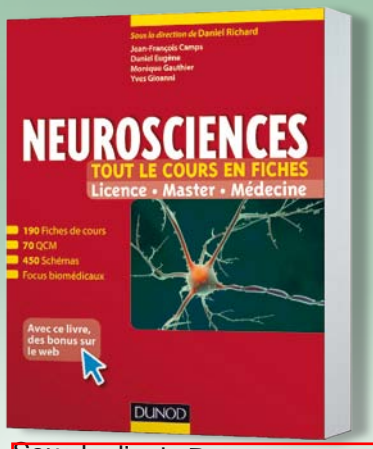
Sous la dir. de **D. Boujard**, ***et al.***
9782100564255 • **35 €**



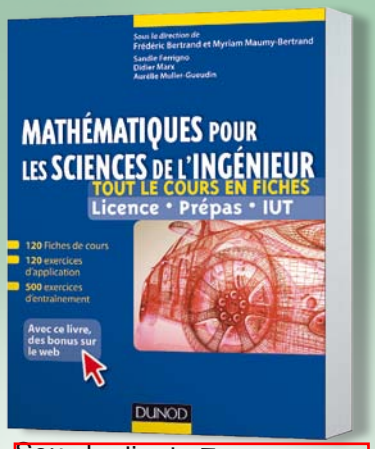
Sous la dir. de **A. Sevin**, ***et al.***
9782100570201 • **35 €**



Sous la dir. de **J. Maddaluno**, ***et al.***
9782100581641 • **35 €**



Sous la dir. de **D. Richard**, ***et al.***
9782100585045 • **49 €**

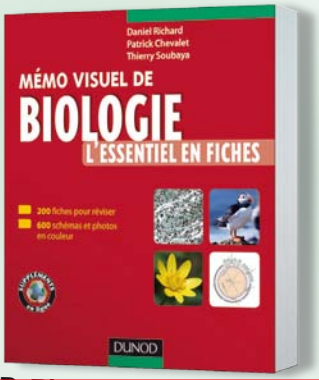


Sous la dir. de **F. Bertrand, M. Maumy Bertrand**, ***et al.***
9782100570614 • **29 €**

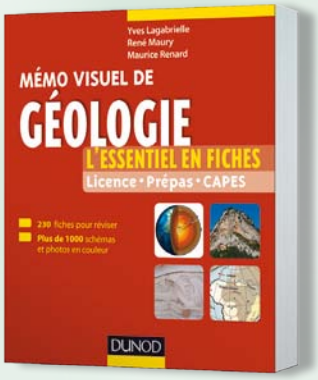
Collection **Mémo visuels**



Pour réviser d'un seul coup d'œil



D. Richard, P. Chevalet, T. Soubaya
9782100564118 • **22,90 €**



Y. Lagabrielle, R. Maury, M. Renard
9782100584994 • **22,90 €**

www.dunod.com

DUNOD
ÉDITEUR DE SAVOIRS

VRAI OU FAUX

Certains autistes sont-ils surdoués ?

Oui. Certains autistes ont un quotient intellectuel très élevé et d'autres une ou plusieurs capacités surdéveloppées.

Isabelle Soulières



Dans le film *Rain Man*, Raymond est un autiste doté de capacités exceptionnelles : il mémorise de multiples noms et numéros de téléphone dans l'annuaire, il estime quasi instantanément le nombre d'allumettes tombées par terre… Ce film a répandu l'idée que certains autistes sont surdoués. Est-ce bien le cas ?

La plupart des psychologues considèrent qu'une personne est surdouée si elle a un quotient intellectuel supérieur à 130, la moyenne étant de 100. Ce critère est satisfait par certains individus présentant les symptômes de l'autisme : difficultés dans les interactions sociales (sujets de conversation inappropriés, difficulté à interpréter les signes non verbaux et les sous-entendus), comportements répétitifs, intérêts particuliers (dynasties chinoises, histoire d'une période spécifique, etc.) poussés à l'obsession.

De façon générale, seuls 25 à 50 pour cent des autistes présentent une déficience intellectuelle. Cette proportion pourrait même être inférieure, car plusieurs études ont montré que les tests déterminant le quotient intellectuel (QI) ne sont pas adaptés à l'intelligence des autistes.

La proportion de personnes ayant un QI supérieur à 130 est probablement identique chez les autistes et dans la population générale, soit environ deux pour cent. Outre ces cas d'intelligence « générale », de nombreuses personnes autistes ont une ou plusieurs capacités surdéveloppées par rapport au reste de la population. L'artiste britannique Stephen Wiltshire, par exemple, peut dessiner une ville de mémoire avec une précision exceptionnelle, après l'avoir vue une seule fois. Certains pianistes peuvent reproduire une mélodie complexe après une écoute unique. D'autres autistes sont capables de faire du calcul de calendrier (donner le jour de la semaine correspondant à une date future ou passée, parfois distante de plusieurs centaines d'années) en moins d'une seconde.

Des capacités spéciales chez 15 à 30 pour cent des personnes autistes

On retrouve ces capacités spéciales chez des personnes autistes un peu partout dans le monde, dans des cultures, des langues et des systèmes d'éducation différents. Il s'agit surtout d'habiletés mathématiques (calcul mental complexe, calcul de calendrier, factorisation de nombres), mnésiques (mémorisation de listes, de plans, de dates), musicales (reconnaître une note sans note de référence, composition), graphiques (dessin en deux ou trois dimensions) et de lecture (certains autistes apprennent à lire seuls dès l'âge de deux ou trois ans).

On estime qu'environ 15 à 30 pour cent des personnes autistes présentent au moins une capacité spéciale. Cette fréquence élevée pourrait s'expliquer en partie par leur façon de se concentrer sur des intérêts particuliers : quand un tel intérêt, qui se traduit par une pratique incessante, se combine à une aptitude particulière, la capacité correspondante est portée à un niveau très élevé.

L'entraînement n'est pas seul en cause. Diverses études ont révélé que les autistes traitent l'information d'une façon différente, aussi efficace, sinon plus. Prenons le cas d'un adulte autiste surdoué pour le dessin en trois dimensions, désigné par E.C. dans les publications scientifiques : quand il dessine une pièce d'instrument de musique, il commence par de petits détails, qu'il relie ensuite pour former la pièce entière, dans des proportions parfaites. En revanche, les dessinateurs professionnels dessinent d'abord la forme globale de la pièce, pour ensuite ajouter les détails. De façon générale, les autistes ont un biais local, c'est-à-dire qu'ils se concentrent plus sur les détails et les éléments constitutifs que sur la structure globale.

Quelles qu'en soient les causes, les capacités spéciales de certains autistes renseignent sur le type d'informations qu'ils traitent de façon optimale. Cela pourrait permettre d'adapter les méthodes d'apprentissage. ■

Isabelle SOULIÈRES est professeure au Département de psychologie de l'Université du Québec, à Montréal (Canada), et chercheuse à l'Institut universitaire en santé mentale de Montréal.

ENTRETIEN





La difficile transmission… d'un autoapprentissage

Entretien avec Philippe Descola et l'un de ses élèves, Emmanuel de Vienne, autour de la méthode de travail des ethnologues.

Claude Lévi-Strauss (1908-2009) a rénové l'anthropologie en proposant une nouvelle méthode pour étudier les sociétés humaines, montrant notamment qu'il n'existe pas de peuples « primitifs ». L'avoir eu comme directeur de thèse n'a certainement pas été sans conséquences sur la carrière de Philippe Descola, professeur au Collège de France et distingué par une médaille d'or du CNRS en 2012 pour l'ensemble de son œuvre. Dans la leçon inaugurale qu'il prononça au Collège de France en 2001, Philippe Descola rappelait ce que lui ont apporté ses études ethnographiques : « Ainsi, [...] plus encore que la reconnaissance due à qui vous enseigne et enrichit votre expérience, ce dont je suis redevable à mes compagnons amérindiens c'est de m'avoir permis, en bouleversant mes évidences par l'assurance tranquille avec laquelle ils adhéraient aux leurs, de m'interroger en retour sur ce que j'avais tenu jusque-là, plus ou moins consciemment, pour des vérités incontestables ; m'incitant ainsi à renouer avec cette vertu fugace de l'étonnement, source du questionnement philosophique et moteur des progrès scientifiques [...]. »

Sans doute a-t-il transmis à ses élèves cette « vertu de l'étonnement ». L'un d'eux, Emmanuel de Vienne, est maître de conférences à l'Université Paris Ouest-Nanterre-La Défense. Si Philippe Descola refuse d'admettre qu'il a fondé une « école », il se réjouit qu'une nouvelle génération d'anthropologues réalise des travaux de qualité.

POUR LA SCIENCE

De 1976 à 1979, vous avez vécu parmi les Achuar, une population jivaro d'Amazonie, à la frontière du Pérou et de l'Équateur. Comment procède l'anthropologue ?

➜ **PHILIPPE DESCOLA :** L'anthropologie occupe une place singulière au sein des sciences. Elle requiert une formation durant laquelle chacun se nourrit du savoir ethnographique accumulé au fil des ans par ses prédécesseurs et apprend les méthodes d'enquête également issues de l'expérience des générations précédentes. Jusque-là, cela pourrait ressembler à tout apprentissage. La principale différence, c'est l'enquête ethnographique sur le terrain. Contrairement à l'étudiant en biologie ou en physique qui interagit au quotidien avec son directeur de thèse, l'ethnographe est seul. L'apprentissage du terrain est un autoapprentissage. On est loin des expériences contrôlées réalisées en laboratoire. L'ethnographe fait sur lui-même l'expérience du rapport à autrui, éprouve sur lui-même la pertinence des questions qu'il se pose et des réponses que les personnes qu'il étudie apportent. L'atelier de l'ethnographe, c'est lui-même, ce n'est pas la paillasse. Dès lors, il est difficile de transmettre ce savoir-faire. Seules se partagent les expériences ethnographiques rapportées dans les livres, qui décrivent les résultats acquis par d'autres et les conditions dans lesquelles ils ont été recueillis.

PLS

Dès lors, deux ethnologues proposent-ils la même analyse d'une même société ?

➜ **PHILIPPE DESCOLA :** C'est une question ouverte. Si vous regardez des ethnographies successives, par exemple ce qui a été écrit sur la société jivaro, vous constatez que les descriptions sont très similaires. Ce qui varie, ce sont les interprétations, parce que les ethnologues acquièrent progressivement des connaissances plus fines et maîtrisent de mieux en mieux l'outil linguistique. Une différence notable entre le travail de ma génération et celui de la génération d'Emmanuel, c'est que nous utilisions la langue vernaculaire comme un outil de communication, alors que l'usage même de la langue est devenu pour eux un objet d'étude.

➜ **EMMANUEL DE VIENNE :** En effet, deux ethnologues n'observent pas nécessairement les mêmes choses. Et il est vrai que les chercheurs de ma génération utilisent davantage la linguistique pour aborder des thèmes classiques de l'ethnologie amazonienne. Cela induit une attention nouvelle aux conditions de l'énonciation, aux genres de discours, à l'emploi de certaines ressources grammaticales. Pierre Déléage, qui fut aussi étudiant de Philippe Descola, a montré l'intérêt de cette approche pour analyser le chamanisme chez les Sharanahua d'Amazonie péruvienne.

PLS

L'ethnographe décide-t-il seul de son sujet ou en accord avec son directeur de thèse ?

➜ **PHILIPPE DESCOLA :** Durant ses années d'études, le futur ethnologue baigne dans une atmosphère riche et intellectuellement exigeante.

© Pour la Science - n° 433 - Novembre 2013

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.

© Hannah Gordon | Fotolia.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

© Pour la Science - n° 433 - Novembre 2013

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.

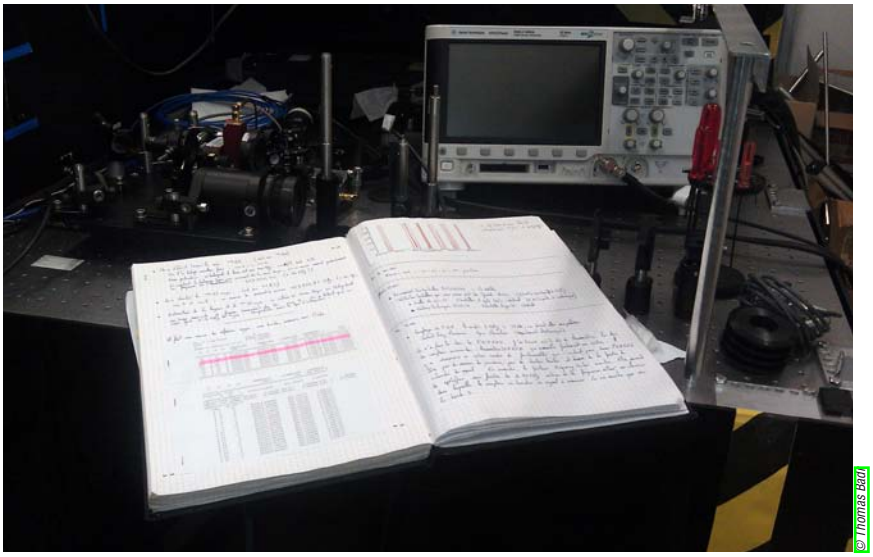
Kilooctet (ko) : 1 000 octets (environ une page d'un livre).

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 Bédé

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

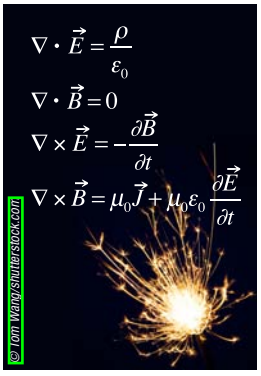
En fait, cette idée est légèrement antérieure au déluge de données. Elle est sans doute apparue quand nous avons pris conscience que les lois qui régissent l'évolution de l'atmosphère ou le fonctionnement d'une cellule, par exemple, ne pouvaient pas s'exprimer en quelques lignes.

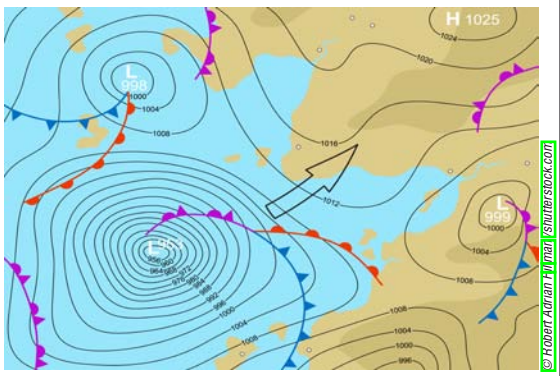
Lorsqu'une théorie s'exprime par des énoncés ou des algorithmes qui font des centaines de pages, le lien entre prévision et explication se distend. Ainsi, nous avons pris l'habitude d'accepter que les prévisions météorologiques indiquent le temps du lendemain avec une probabilité de succès assez élevée, 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.



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 ?

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-

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.

© 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.

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

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. ■



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.

© Tupikov Valeriy/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.

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.

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. ■

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



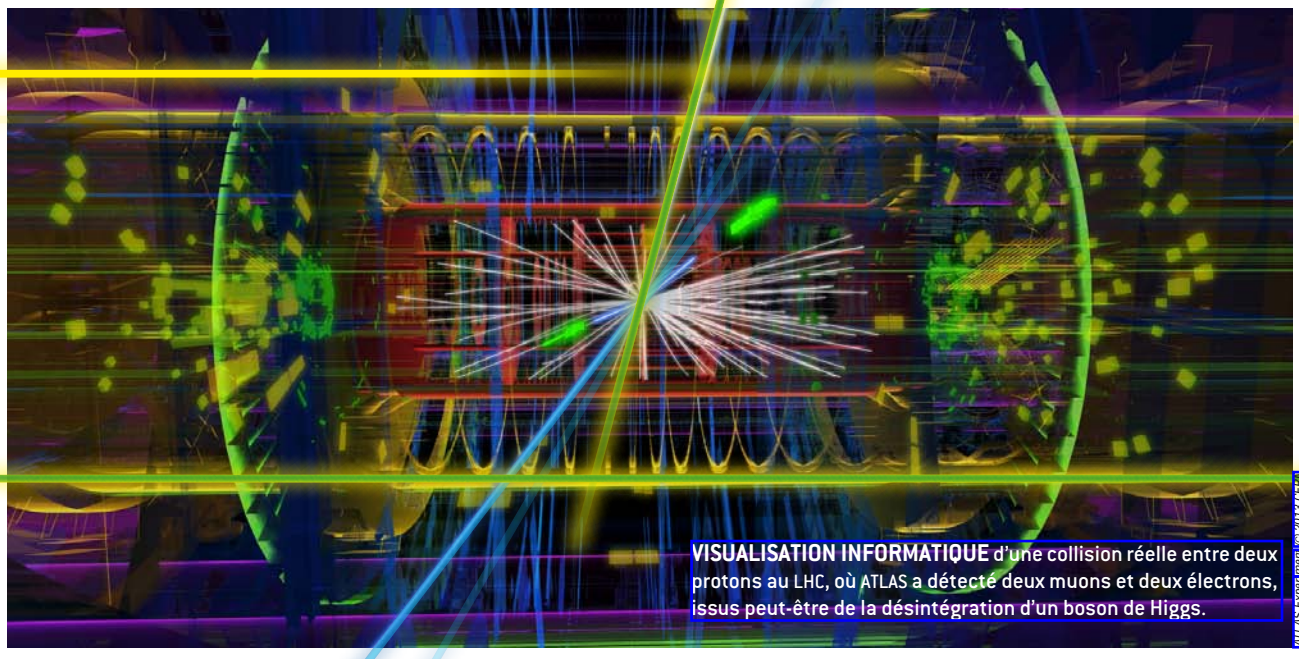
Faïrouz MALEK est physicienne au Laboratoire de physique subatomique et de cosmologie (LPSC) du CNRS et de l'Université Joseph Fourier, à Grenoble. Elle participe à l'expérience ATLAS du LHC et est responsable scientifique du projet LCG-France, la partie française de la grille de calcul du LHC.

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.

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

© iSAKSO/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.

© Pour la Science - n° 433 - Novembre 2013

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

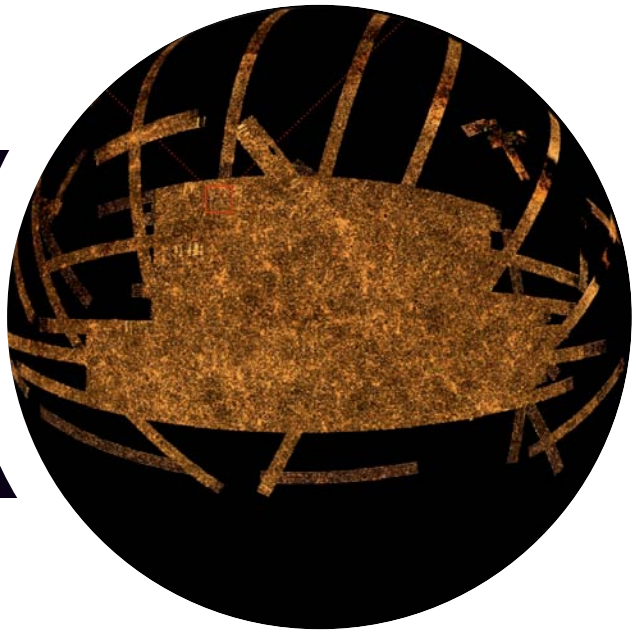
Ces informations individuelles, mises en lien d'une étoile à une autre, permettent de révéler en détail la structure spirale de la Voie lactée, de découvrir de nouvelles classes d'étoiles, d'étudier les déplacements des astres à l'intérieur de la Galaxie, etc. L'ensemble des renseignements ainsi obtenus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

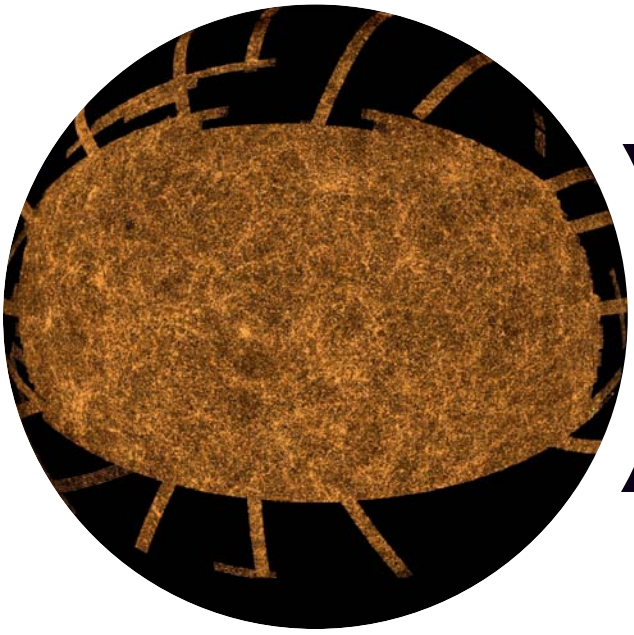
Un deuxième volet de SDSS est le relevé BOSS *(Baryon Oscillation Spectroscopic Survey)*, qui concerne l'Univers lointain. Son objectif est d'éclairer le mystère de l'« énergie sombre », entité de nature inconnue qui serait responsable de l'accélération de l'expansion de l'Univers, et d'aider à améliorer 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

© Illustration SDSS-III





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. ■

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.

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

© Pour la Science - n° 433 - Novembre 2013



Francis Daumas et Marion Massol

Au cœur d'un centre d'archivage

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

Leur quantité croît de façon exponentielle, comme les volumes de données produites.

Des centres d'archivage ont alors été développés dans plusieurs pays depuis une quinzaine d'années. Ils ont d'abord été l'œuvre d'organismes publics, tels des agences spatiales, des instituts de recherche, des bibliothèques nationales ou des services administratifs. Un marché privé s'est aussi mis en place, mais avec des qualités de service inégales et des garanties diverses sur la pérennité des données, assurée plus souvent à court ou moyen terme qu'à long terme.

En outre, de grandes infrastructures ont été créées, à des échelles nationales ou internationales. Par exemple, les projets européens PRACE (*Partnership for advanced computing in Europe*, ou Partenariat pour le calcul avancé en Europe) et *Eudat* (*European data infrastructure*, ou Infrastructure de données européenne) visent à se doter respectivement d'importants moyens de calcul (et de stockage des données associées) et d'une plate-forme collaborative de partage et de conservation des données scientifiques.

Plus de huit millions de fichiers archivés

En France, le pionnier de l'archivage numérique est le CNES (Centre national d'études spatiales), qui archive depuis 1998 des images prises par des satellites et de grands observatoires, ainsi que diverses données de calcul. Le CINES, quant à lui, dispose d'un service d'archivage numérique opérationnel depuis 2006. Il a une capacité de stockage de plusieurs pétaoctets (10^{15} octets), soit l'équivalent d'une pile de CD d'environ dix kilomètres de hauteur. Plus de huit millions de fichiers sont aujourd'hui archivés, occupant une centaine de téraoctets (10^{12} octets). Le CINES a reçu un agrément (une certification) délivré par le SIAF (Service interministériel des archives de France) pour les archives dites intermédiaires, c'est-à-dire les documents qui ne sont pas nécessaires à l'activité quotidienne des services, mais qui doivent être conservés un certain temps pour des raisons administratives ou juridiques. La Commission européenne lui a aussi délivré une accréditation validant la qualité de son système d'archivage électronique.

Le CINES est chargé d'archiver toutes les thèses de doctorat françaises au format numérique. Il participe au projet de Bibliothèque scientifique numérique nationale, ou BSN *(voir l'article de Frédéric Blin dans ce numéro)*, où il pilote les questions liées à l'archivage du patrimoine scientifique public. Il héberge une partie des données de nombreux projets et organismes : la Très grande infrastructure de recherche (TGIR) *HumaNum* du CNRS (composée de moyens humains et numériques destinés aux sciences humaines), plusieurs bibliothèques universitaires, divers organismes de recherche (INSERM, IRSTEA, etc.), la cour des comptes... Il héberge aussi la première plate-forme numérique française de cours en ligne ouverts et massifs (CLOM, ou MOOC en anglais), dont le lancement a été annoncé début octobre 2013. En outre, il mène des actions de sensibilisation et de formation aux problématiques de l'archivage auprès de la communauté scientifique.

Outre l'archivage, le CINES propose des moyens de calcul numérique intensif à l'ensemble de la communauté scientifique. Le centre héberge notamment l'un des plus puissants supercalculateurs européens, financé par la société civile GENCI (Grand équipement national de calcul intensif) et utilisé pour des simulations complexes dans de nombreux domaines (climatologie, mécanique des fluides, physique des plasmas, astrophysique, géophysique, biologie, etc.). Archivage et calcul sont complémentaires, car les données servent d'entrées aux calculs qui, eux-mêmes, produisent de grandes quantités de données, dont une partie doit être conservée. Cependant, les volumes de données étant colossaux, on ne leur applique pas certaines procédures d'archivage à long terme (telle la copie du document en cinq exemplaires, sur laquelle nous reviendrons).

Une cinquantaine d'ingénieurs et de techniciens travaillent au CINES. Une douzaine d'entre eux se consacrent aux projets d'archivage, qui font aussi parfois appel à des intervenants extérieurs. Les profils sont variés : archivistes (spécialistes de la gestion des documents), informaticiens, juristes (garants du respect de la réglementation en vigueur), etc.

Des infrastructures sécurisées

Le centre s'est doté d'une infrastructure et d'une plate-forme logicielle adaptées à l'archivage. Les données y sont accessibles en permanence *via* le réseau Internet. L'alimentation électrique des équipements liés à l'archivage est redondante, afin de

L'ESSENTIEL

- D'énormes quantités de données numériques sont stockées dans de grands centres d'archivage dotés d'infrastructures sécurisées.
- Leur préservation à long terme pose plusieurs difficultés : détérioration et évolution des supports, formats de fichiers qui se périment, etc.
- On y répond le plus souvent par une gestion active des archives, incluant la migration des documents vers de nouveaux supports et l'actualisation de leur format.

© Pour la Science - n° 433 - Novembre 2013

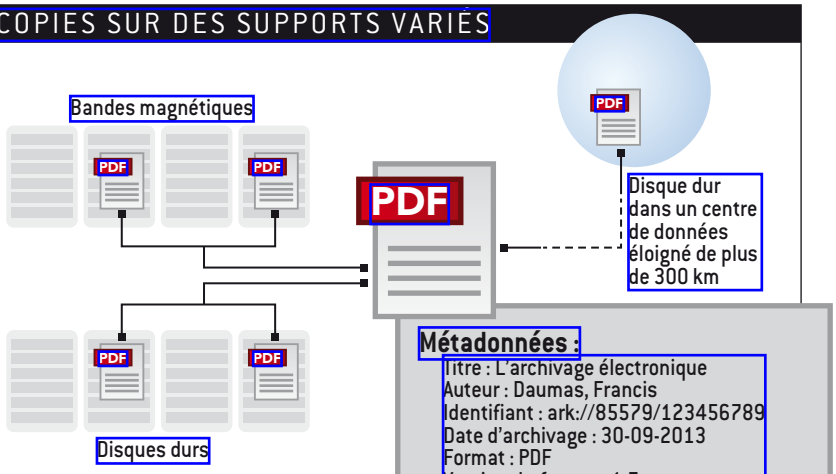
prévenir les coupures de courant : elle est garantie par deux lignes électriques d'ERDF indépendantes et par des groupes électrogènes ; en outre, divers dispositifs de protection sont installés. Les conditions climatiques (température, humidité) des salles hébergeant les machines sont contrôlées. L'accès au site est sécurisé (filtrage des visiteurs, accès restreint à certaines zones), de même que les connexions *via* le réseau (grâce à des pare-feux, des protocoles sécurisés, des antivirus). En outre, toutes les données informatiques du centre (archives et systèmes) sont sauvegardées quotidiennement. Cette sauvegarde s'effectue de façon incrémentale, c'est-à-dire qu'on n'enregistre que les changements.

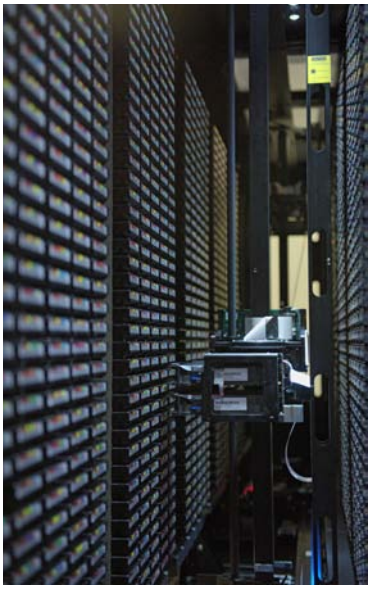
La plate-forme d'archivage du CINES, conforme aux normes internationales du domaine, est fondée sur trois « serveurs logiques », respectivement dits de transfert, de stockage et d'accès. Le premier reçoit les documents à archiver, le deuxième les conserve et le troisième permet leur consultation. Décrivons-les plus en détail.

Un paquet d'information à archiver contient l'objet à conserver et un ensemble de métadonnées (des données permettant de comprendre d'autres données). Ces dernières sont élaborées en partie par l'organisme dit versant, qui fournit les documents à archiver (c'est en général l'organisme qui les produit) et en partie par l'archiveur. Elles décrivent le contexte (organisme producteur, organisme versant), le document lui-même (titre, auteur, format), les paramètres de gestion (localisation physique, dispositif de stockage, etc.). Un autre type de métadonnées, qui ne sont pas associées à un document précis mais à la plate-forme d'archivage, décrit les informations dites de représentations, essentielles à la lecture de l'archive : les caractéristiques des formats, les structures des métadonnées, la description de la plate-forme d'archivage et des normes utilisées…

PLUSIEURS COPIES SUR DES SUPPORTS VARIÉS

Les supports physiques sont vulnérables et les estimations de leur durée de vie varient selon les études. Pour parer cette menace, cinq copies du document à archiver sont réalisées sur des types de supports différents : deux sur des bandes magnétiques et trois sur des disques durs (dont une située à plus de 300 kilomètres des autres, afin de préserver le document d'une catastrophe majeure, tel un séisme). Chaque document est accompagné d'un ensemble de métadonnées qui permettent de l'identifier, de le localiser et de faire évoluer son format.







Les robothèques gèrent de façon automatique un ensemble de bandes magnétiques *(ci-dessus, une robothèque vue de l'extérieur et, ci-contre, à gauche, vue de l'intérieur)*. Deux copies du document y sont enregistrées. Lorsqu'une archive est demandée, un bras robot prélève la bande où elle est stockée et la place dans un dispositif de lecture.



Les baies de stockage renferment de multiples disques durs. Deux copies du document y sont stockées, dans des salles distinctes.



© CINES

© Pour la Science - n° 433 - Novembre 2013

Le serveur de transfert contrôle les documents reçus, en vérifiant notamment la conformité des métadonnées et des formats de fichiers. Il peut aussi effectuer diverses transformations, telle la conversion d'un format non libre (par exemple *Word*) en un autre (tel *Open office*). Il enrichit ensuite les métadonnées par des paramètres d'archivage, tel un identifiant international unique (une succession de lettres et de chiffres caractéristique du document). Enfin, il transmet le document au serveur de stockage.

Ce dernier en réalise cinq copies. Deux d'entre elles sont enregistrées dans deux baies de stockage distinctes (des armoires comprenant un certain nombre de disques durs), situées dans deux salles distantes. Deux autres sont inscrites sur des bandes magnétiques gérées par des dispositifs automatisés nommés robothèques. Ces robothèques, de deux pétaoctets chacune, sont aussi dans deux salles différentes. Une dernière copie est réalisée à plus de 300 kilomètres du site, dans un autre centre informatique, pour garantir les données contre les risques de catastrophe naturelle majeure.

Enfin, le serveur d'accès reçoit les demandes de consultation des archives, vérifie les autorisations de leurs émetteurs et leur transmet ou non les documents. Il peut également effectuer diverses modifications, comme flouter des visages pour préserver l'anonymat des personnes lors de l'envoi d'une photographie.

Des métadonnées bien renseignées répondent aux problèmes de localisation et de référencement des documents soulevés au début de cet article. Qu'en est-il de l'obsolescence et de la détérioration des supports et des formats ? Deux options se présentent : conserver en état de marche tous les dispositifs matériels et logiciels permettant la lecture de l'archive, afin de les réutiliser plus tard ; effectuer une migration des supports et des formats à mesure que les techniques évoluent (là encore, les métadonnées sont essentielles). Le CINES a choisi la seconde option.

Ainsi, les documents sont transférés vers les types de supports les plus récents – par exemple, aucune donnée n'est conservée sur disquette aujourd'hui. En outre, la durée de vie des supports est prise en compte pour assurer un renouvellement adapté. Les documents stockés sur des disques durs sont ainsi transférés tous les cinq ans, et ceux stockés sur bande magnétique tous les dix ans.

Les migrations concernent aussi les formats (PDF, JPEG…). La plate-forme du CINES accepte une quinzaine de formats, couvrant les différents types de données. Le choix est fait en privilégiant ceux dont les caractéristiques sont publiées et disponibles, et dont on pense que les versions successives seront compatibles avec les documents anciens. Une procédure de migration se déclenche lorsqu'un format est obsolète. Elle consiste en général à adapter un document aux versions successives d'un même format, mais il arrive aussi qu'un format disparaisse (tel celui des fichiers créés par le logiciel de bureautique *Claris*, abandonné au début des années 2000) et qu'il faille en changer. Elle se déroule en étroite collaboration avec l'organisme versant des données archivées, qui valide la nouvelle forme du document.

ENVIRON 40 RISQUES SONT IDENTIFIÉS
aujourd'hui, tels qu'un incendie ou un format de fichier qui se périme.

La gestion des archives à long terme nécessite également une « démarche qualité » – un ensemble de bonnes pratiques techniques et organisationnelles. Le respect des normes (assurant notamment l'interopérabilité avec d'autres systèmes d'archivage ou de collecte de données), la vérification du bon encodage du document selon le format choisi, les copies multiples et le contrôle de leur état participent à cette démarche. Au niveau organisationnel, tous les processus mis en œuvre (par exemple, l'ensemble des actions à réaliser pour faire migrer un document d'un format à un autre) sont répertoriés et conservés par écrit. Cela permet de transmettre aux générations suivantes le savoir-faire et l'historique du service, mais aussi de mieux analyser et d'améliorer les procédures. En outre, les risques potentiels et les actions à mener pour y parer sont actualisés tous les six mois. Environ 40 risques sont identifiés, tels qu'un incendie ou un format de fichier qui se périme.

Offrir un service d'archivage numérique ne se résume donc pas à acquérir des dispositifs de stockage et à y copier des données. La pérennisation de documents nécessite une gestion active et une surveillance régulière, aussi bien des logiciels que du matériel. ■

LES AUTEURS





Francis DAUMAS est directeur du CINES (Centre informatique national de l'enseignement supérieur), à Montpellier.

Marion MASSOL est responsable du Département d'archivage du CINES.

SUR LE WEB

Groupe PIN (pérennisation de l'information numérique) de l'association Aristote : http://pin.association-aristote.fr

BIBLIOGRAPHIE

Collectif, **Archivage numérique pérenne,** *La Gazette du CINES*, février 2013 (www.cines.fr).

J.-C. Hourcade *et al.*, **Longévité de l'information numérique : Les données que nous voulons garder vont-elles s'effacer ?,** EDP Sciences, 2010.

Norme iso 14 721, ou OAIS (*Open archival information system*, soit Système ouvert d'archivage d'information). www.iso.org

L'information scientifique de référence **maintenant sur tablette et smartphone !**



Téléchargez gratuitement
l'application sur App Store et Google Play.
Le 1er numéro de *Pour la Science* est offert !

Disponible sur App Store

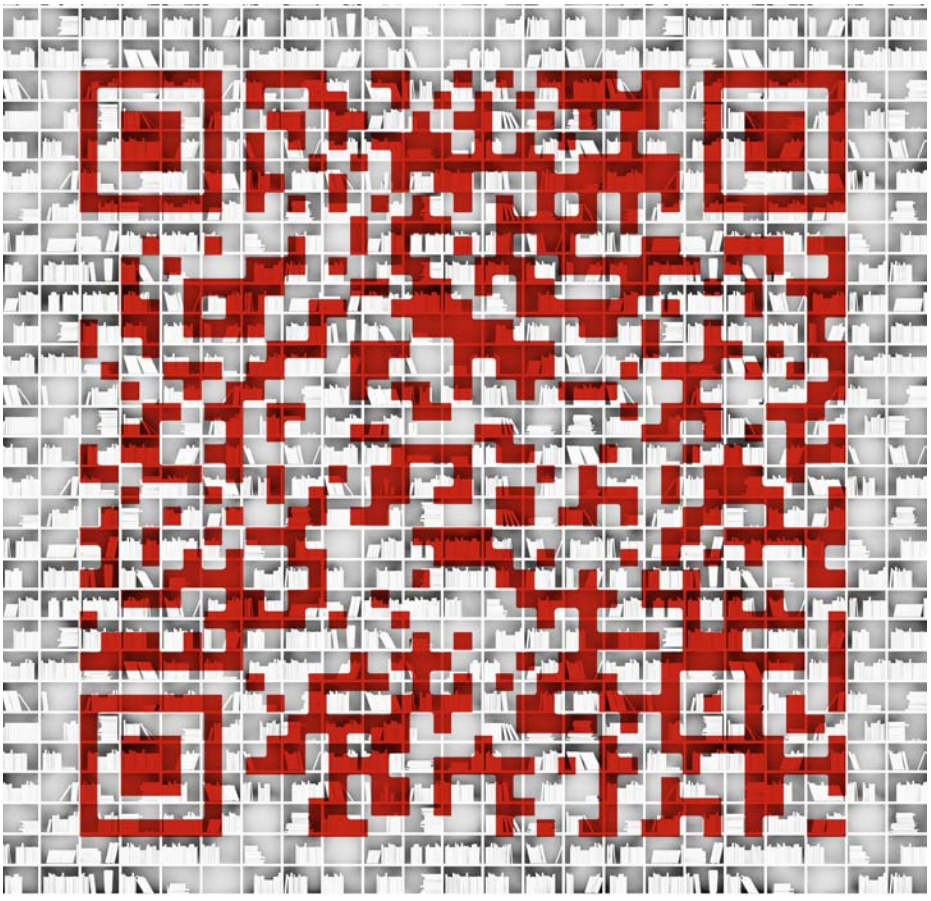
DISPONIBLE SUR Google play

Flashez ce QR code avec votre mobile ou votre tablette pour télécharger immédiatement l'application.

Avec l'application « Pour la Science », retrouvez dès leur sortie le mensuel *Pour la Science* à 5,49 € et son hors-série trimestriel *Dossier Pour la Science* à 5,99 € en version numérique optimisée pour tablette : lecture intuitive, sommaire interactif, contenus enrichis, feuilletage hors connexion, etc.

&

Découvrez aussi l'application « Cerveau & Psycho » pour retrouver le magazine *Cerveau & Psycho* et son hors-série *L'Essentiel Cerveau & Psycho*



© Zastolskiy Victoria /Shutterstock.com/Pour la Science

Le savoir numérisé

Nouvelles pratiques, nouveaux enjeux

Bibliothèques numériques, fonds numérisés, encyclopédies et dictionnaires en ligne, portails d'archives scientifiques en libre accès... En quelques années, Internet a permis un développement extraordinaire de l'accès au savoir, offrant à tout un chacun la possibilité de consulter, en quelques clics, des manuscrits de Descartes ou les dernières recherches en physique, ou encore de découvrir par le menu l'histoire du transistor.

Les bouleversements introduits par ces nouveaux modes de stockage et de diffusion de la connaissance sont cependant bien plus profonds. Qu'il s'agisse d'ouvrages numériques natifs, de sites Web ou de documents numérisés, les œuvres et les connaissances sous forme numérique s'accumulent. Comment les préserver, alors qu'elles sont stockées sur des supports dont la pérennité ne dépasse pas quelques dizaines d'années ? Et comment concilier ces nouvelles pratiques avec le droit d'auteur ? Les bibliothèques sont au cœur de ces réflexions *(voir* La numérisation des bibliothèques*, page 48)*. Le modèle économique de diffusion du savoir est aussi ébranlé par le succès, d'une part, de l'encyclopédie en ligne gratuite *Wikipédia* et, d'autre part, du libre accès aux publications scientifiques : si certains éditeurs d'encyclopédies semblent tirer leur épingle du jeu *(voir* Dictionnaires & encyclopédies à l'ère numérique*, page 52)*, la situation est tendue entre les chercheurs partisans du libre accès et les éditeurs de revues scientifiques *(voir* La science accessible à tous*, page 54)*.

Mais le changement le plus drastique concerne les contenus eux-mêmes. Quelle fiabilité accorder aux entrées des encyclopédies et dictionnaires en ligne, qui ne sont pas estampillés par une autorité intellectuelle bien définie, ou aux articles scientifiques proposés en accès libre par leurs auteurs avant publication ? Plus que jamais, il est indispensable d'aborder ses lectures avec un œil critique. L'accès libre au savoir est à ce prix.

Marie-Neige Cordonnier

La numérisation des bibliothèques

Frédéric Blin

Depuis une vingtaine d'années, les bibliothèques du monde entier numérisent leurs collections et les rendent accessibles sur Internet. Ce nouveau rôle bouleverse les modes d'accès au savoir.

On trouve tout sur Internet, on n'a plus besoin de bibliothèques : cette phrase est devenue si banale qu'elle ne choque plus, ou si peu. Et pourtant. Peut-on tout trouver sur Internet ? Les centaines de millions de livres, journaux, revues, images, objets conservés par nos bibliothèques ont-ils été déjà tous numérisés ? Et ce qui existe sous forme numérique est-il accessible sur Internet, où que l'on se trouve ? Quel est le rôle des bibliothèques désormais, dans ce monde où l'information et le savoir sont de plus en plus exclusivement numériques ?

Au cœur des réflexions professionnelles des bibliothécaires, ces questions devraient intéresser un public bien plus large, notamment au sein des communautés universitaires et scientifiques. Le numérique a entraîné un bouleversement des modes d'accès aux savoirs. Toutefois, la facilité relative avec laquelle cet accès se fait masque divers dangers. Tout d'abord, les savoirs et informations non numérisés, délaissés, risquent de disparaître ou d'être oubliés. Ensuite, la médiation entre les contenus informatifs et les utilisateurs de ces contenus, réalisée par les bibliothécaires, est remise en cause par la communication directe sur Internet, mais aussi par le contrôle que d'autres acteurs – ceux qui tirent un profit économique de l'information – exercent de plus en plus sur la possibilité même d'y accéder. Enfin, le déluge d'informations et de données, qui s'accroît chaque jour, soulève la question de nos capacités à y naviguer de façon pertinente et à tout conserver ou, du moins, à préserver ce qui sera utile aux générations futures, en particulier pour les besoins scientifiques et mémoriels.

Du projet *Gutenberg* à *Google Books*

La première initiative de numérisation de livres, le projet *Gutenberg* piloté par l'Américain Michael Hart, de l'Université de l'Illinois, aux États-Unis, remonte à 1971. Ce projet, qui se poursuit, repose non pas sur des bibliothèques, mais sur les contributions de volontaires. Ceux-ci numérisent des ouvrages du domaine public et participent à leur transformation en textes numériques. Les principales bibliothèques du monde, quant à elles, ont commencé à numériser leurs collections au début des années 1990, pour deux raisons principales. D'une part, il s'agissait d'assurer une meilleure préservation des documents : la numérisation était alors vue comme un moyen de remplacer à terme le microfilm. D'autre part, l'objectif était de faciliter l'accès du public à des collections fragiles et précieuses.

En France, *Gallica*, la bibliothèque numérique de la Bibliothèque nationale, a ainsi ouvert en 1997, donnant alors accès à une sélection de ses trésors les plus précieux. D'autres grandes bibliothèques lui ont emboîté le pas dans les années suivantes : la

© Shutterstock

Bibliothèque interuniversitaire de médecine avec *Medic@*, le Conservatoire national des arts et métiers avec le Conservatoire numérique des arts et métiers (CNUM), la Bibliothèque municipale de Lyon… La numérisation des fonds anciens est devenue une activité quotidienne, si ce n'est banale, des bibliothèques françaises, et les ouvertures de bibliothèques numériques modernes se sont multipliées au cours des dernières années : *NumeLyo* à Lyon, *Rosalis* à Toulouse, *Numistral* à Strasbourg, ou les bibliothèques numériques régionales d'Aquitaine, de Bretagne, de Lorraine, du Limousin, ou encore *Manioc* en Antilles-Guyane… Même les plus petites mettent en valeur leurs fonds anciens grâce à des outils partagés, telle la bibliothèque numérique *e-corpus*, développée par le Centre de conservation du livre d'Arles, qui réunit aujourd'hui les fonds de plus de 250 établissements.

La France n'est à cet égard ni en avance ni en retard par rapport aux autres pays comparables. Depuis quelques mois, l'Allemagne dispose de sa bibliothèque numérique nationale, la *Deutsche Digitale Bibliothek*. Nourri en partie de programmes financés par la Fondation allemande pour la recherche et des réalisations des trois principaux centres de numérisation du pays, à Göttingen, Munich et Berlin, ce projet ambitieux a pour objectif la numérisation de l'ensemble des publications du pays antérieures à 1800. Dans le monde hispanophone, la bibliothèque numérique *Miguel de Cervantes*, hébergée à l'Université d'Alicante, fait figure de référence depuis 1999.

En Chine, c'est tout un réseau national qui se construit à partir de la Bibliothèque numérique nationale ouverte en 2007. D'ici 2015, ce réseau reliera les plates-formes de bibliothèques numériques développées localement par les universités ou bibliothèques publiques, donnant à chacune d'elles accès non seulement au patrimoine numérisé, mais aussi à des ressources et à des services numériques et multimédia.

En Afrique, plusieurs projets de soutien au développement de bibliothèques numériques ont vu le jour depuis une dizaine d'années. Parmi eux, le projet *African Digital Libraries Support Network*, lancé en 2010 avec le soutien de la fondation internationale EIFL *(Electronic Information for Libraries)*, a permis de créer un réseau de centres d'expertise dans huit pays africains dont l'objectif est l'accompagnement de projets de constructions de bibliothèques numériques.

Citons enfin la toute récente Bibliothèque numérique publique d'Amérique *(Digital Public Library of America)*, dont la mission est de donner accès aux considérables patrimoines des bibliothèques, universités, archives et musées des États-Unis, afin d'éduquer et d'informer, et de transmettre ce patrimoine à l'ensemble des citoyens américains.

Dans ce paysage, *Google Books* occupe une place particulière. Le lancement, en 2004, de ce projet visant à numériser et à diffuser l'ensemble des livres du monde a marqué un tournant majeur dans l'histoire des bibliothèques numériques. Pour la première fois, une entreprise privée disposait du potentiel de contrôle sur le patrimoine documentaire mondial, concurrençant directement les bibliothèques. Celles-ci ont dû se positionner, à l'instar de la Bibliothèque nationale de France : son passage d'une logique de numérisation sélective à une logique de numérisation systématique s'explique par la volonté de constituer une masse documentaire conséquente qui lui permettrait de devenir le plus légitime vecteur du patrimoine français numérisé.

Lexique

Humanités numériques : **Cette nouvelle science utilise les méthodes et les outils du numérique en sciences humaines et sociales.**

BSN : La Bibliothèque scientifique numérique coordonne en France les initiatives favorisant l'accès à l'information scientifique.

ARROW : Ce dispositif européen vise à simplifier la gestion des droits dans les projets de numérisation.

OAIS : Ce modèle de gestion, d'archivage et de préservation des données numériques est la norme en vigueur.



Les bibliothèques réagissent

La bibliothèque numérique *Europeana*, quant à elle, a été créée en 2008 pour redonner aux États européens l'initiative sur la mise en valeur de leur patrimoine. Grâce notamment au fort soutien de la Commission européenne aux programmes de numérisation collaboratifs visant à l'enrichir, elle compte plus de 20 millions de documents et d'objets. Aux États-Unis, *Hathi Trust*, construit par les bibliothèques ayant participé aux débuts de *Google Books* à partir du fonds d'ouvrages numérisés par *Google*, est devenu, en volume de titres numérisés, le plus proche concurrent de *Google*.

Malgré la richesse et la multiplicité des initiatives, les dizaines voire les centaines de millions de documents déjà numérisés ne représentent qu'une petite partie des richesses documentaires présentes dans les bibliothèques. Ainsi, en 2012, le projet *Enumerate*, financé par la Commission européenne, a estimé que seuls 20 pour cent des collections des institutions culturelles européennes étaient déjà numérisés.

L'AUTEUR



Frédéric BLIN est directeur de la conservation et du patrimoine de la Bibliothèque nationale et universitaire de Strasbourg.

En réalité, si l'on considère les établissements n'ayant pas répondu à l'enquête, cette proportion doit être bien inférieure…

Dans le monde scientifique aussi, la numérisation est devenue courante depuis une petite quinzaine d'années. En France, aux programmes de numérisation de revues scientifiques, tels que NUMDAM en mathématiques et PERSEE pour les sciences humaines et sociales, ont succédé les projets soumis à l'Agence nationale pour la recherche, puis aux Investissements d'excellence. Une science – les humanités numériques – est en fort développement aux États-Unis, en Europe et partout dans le monde, désignant l'utilisation d'outils informatiques avancés pour exploiter des contenus numériques de manière innovante. En France, la Bibliothèque virtuelle humaniste (BVH), pilotée par le Centre d'études supérieures de la Renaissance de Tours, et *Biblissima*, de l'Institut de recherche et d'histoire des textes du CNRS, sont deux des projets les plus emblématiques de cette nouvelle science, qui a aujourd'hui ses propres Très grandes infrastructures de recherche (TGIR) : *HumaNum* en France et *Dariah* en Europe.

La Bibliothèque scientifique numérique (BSN), TGIR lancée en 2008 par le ministère de l'Enseignement supérieur et de la Recherche, vise à structurer le paysage et à coordonner les initiatives liées à l'information scientifique et technique à l'échelle nationale, en particulier celles mettant en œuvre la numérisation du patrimoine scientifique national. L'appel à projets lancé en avril 2013 par son segment 5 (BSN5-Numérisation du patrimoine scientifique) a ainsi fait remonter près de 75 réponses – des projets issus de disciplines aussi diverses que la paléontologie, la chimie, l'histoire de l'art ou encore les sciences économiques, témoignant non seulement de la richesse des initiatives, mais aussi des potentialités infinies que la numérisation ouvre à la recherche.

Le paradoxe du droit d'auteur

La plupart de ces projets se sont heurtés, à des degrés divers, à la question du droit d'auteur. Courant en général sur 70 années après la mort de l'auteur, ce droit contraint les porteurs de projets à s'assurer de l'accord des ayants droit avant de numériser une œuvre encore protégée. La recherche de ces ayants droit peut se révéler longue et difficile, et donc coûteuse. Deux solutions sont alors possibles : soit numériser, attendre que les ayants droit se manifestent et, si besoin, retirer ensuite les ouvrages concernés, soit se contenter de numériser des ouvrages libres de droit. La première alternative a été le choix de la société *Google*. Elle l'a entraînée dans des procédures juridiques avec les éditeurs de multiples pays. La seconde est celle adoptée par les bibliothèques.

Celles-ci ne disposant pas des ressources pour effectuer les recherches des ayants droit, un registre, ARROW (*Accessible Regis-*

Numériser un document

La numérisation d'un document se déroule en quatre étapes principales : la prise de vue (ou l'enregistrement en cas de matériel audiovisuel), la structuration, l'enrichissement et la description du document numérique, l'archivage des fichiers et, enfin, la mise à disposition du public.

La prise de vue est l'étape la plus simple. Dans le cas d'un document imprimé, une résolution de numérisation à 300 pixels par pouce est suffisante pour la majorité des usages, tout en limitant le poids relatif des fichiers. Mais le matériel utilisé détermine aussi la qualité de l'image et la bonne conservation du document physique. Des modèles de scanners patrimoniaux ont ainsi été développés en partenariat avec des bibliothèques pour garantir une conservation optimale de l'original.

La deuxième étape peut être assez complexe, selon les moyens et les objectifs du projet. Le premier type d'enrichissement est la reconnaissance optique de caractères (OCR), pour transformer une simple image en un texte reconnaissable par un logiciel informatique. Les logiciels de reconnaissance les plus performants atteignent aujourd'hui plus de 99 pour cent de réussite, pour des textes imprimés des XIX^{e} et XX^{e} siècles en caractères latins. Des projets de recherche sont en cours pour les documents plus anciens (projet européen IMPACT) ou les manuscrits, beaucoup plus difficiles à traiter.

La structuration informatique des documents permet d'identifier des parties des ouvrages, offrant par exemple un accès direct aux chapitres. Divers schémas de structuration existent. Les plus courants pour l'imprimé sont le METS et la TEI, ou l'EAD pour les manuscrits et archives. Enfin, la création de métadonnées permet d'identifier et de repérer le document dans l'océan de données du Web. Le standard de métadonnées minimal aujourd'hui, en bibliothèque, est le *Dublin Core*.

L'archivage doit être pensé en amont du projet : la résolution des images doit optimiser le poids des fichiers par rapport aux usages attendus, les métadonnées doivent être assez précises pour que l'on puisse retracer l'historique et l'identité des fichiers numériques, le cycle de vie des archives doit être aux normes, et le coût de la conservation doit être pris en compte.

Enfin, la diffusion doit répondre aux attentes des utilisateurs en s'accompagnant de services et de données telles que les conditions d'utilisation. Le statut (*Creative Commons*, licence ouverte, etc.) sous lequel l'institution décide de placer ses images et ses données numérisées conditionnera les usages qui pourront en être faits.



Une page de l'*Autobiographie du bienheureux Henri Suso*, un livre manuscrit sur parchemin du XIV^{e} siècle. Des projets de recherche visent à mettre au point des logiciels capables de reconnaître les textes de tels manuscrits anciens.

© BNUS Strasbourg

© Pour la Science - n° 433 - Novembre 2013

tries of Rights information and Orphan Works towards Europeana), financé par la Commission européenne, a été démarré en 2008 pour faciliter cette recherche. Cependant, ARROW ne peut à lui seul identifier l'ensemble des œuvres sous droits et leurs ayants droit légitimes. Aussi, en raison de la réglementation sur le droit d'auteur, il est plus facile aujourd'hui de trouver sur Internet un ouvrage de 1860 que de 1960. Un pan entier de la littérature, celle du XX^{e} siècle, est en grande part inaccessible en ligne, alors même que la valeur commerciale des documents concernés est faible, voire nulle, et que ceux-ci sont souvent indisponibles – dans le commerce, sous format papier voire même en format électronique – ou orphelins – sans ayant droit connu.

Le programme de numérisation des œuvres indisponibles lancé par l'État en 2013 *via* la Bibliothèque nationale de France vise justement à numériser 500 000 œuvres du XX^{e} siècle pour combler ce vide, mais les modalités – notamment de financement – de ce programme sont très critiquées par une partie importante des cercles professionnels. Une loi sur les œuvres indisponibles est en préparation au niveau européen, bénéficiant du soutien des associations de bibliothèques. Celles-ci sont aussi moteur d'un projet de traité international sur les exceptions et les limitations au droit d'auteur en discussion au sein de l'Organisation mondiale de la propriété intellectuelle.

De nouvelles archives à préserver

La pérennité des documents numérisés est aussi au cœur des réflexions. Une erreur fondamentale serait de croire que la préservation de la version physique originale d'un document numérisé serait inutile. D'une part, l'original a une valeur patrimoniale en soi et, d'autre part, la durée de la pérennité des données numériques n'est aujourd'hui de l'ordre que d'une trentaine d'années. Numériser pour faire de la place sur les rayonnages est pour l'heure un danger réel pour la transmission du patrimoine et de la mémoire. Les exemples de pertes des données conservées sur des CD-ROM sont légion, sans parler de la qualité technique des numérisations réalisées il y a quelques années, insuffisante par rapport aux besoins d'aujourd'hui.

Quelques chiffres

130 millions de livres ont été publiés dans le monde depuis l'invention de l'imprimerie, selon *Google*, dont 20 % sont dans le domaine public.

20 % des collections des établissements culturels européens étaient numérisés en 2012, selon l'enquête *Enumerate*.

30 millions de documents sont accessibles *via Google Books*, 26 millions *via Europana*, 11 millions *via Hathi Trust* et 2,2 millions *via Gallica*.

5 000 euros : c'est le coût annuel moyen de la préservation à long terme d'un téraoctet de données numérisées.

BIBLIOGRAPHIE

A. Jacquesson, **Google Livres et le futur des bibliothèques numériques. Historique du projet, techniques, documentaires, alternatives et controverses,** Cercle de la librairie, 2012.

Th. Claerr et I. Westeel (sous la dir.), **Manuel de la numérisation,** Cercle de la librairie, 2012.

J.-Ph. Accart et A. Rivier, **Mémento de l'information numérique,** Cercle de la librairie, 2012.

D. Baker et W. Evans (éd.), **Digital Library Economics. An academic perspective,** Chandos Publishing, 2009.

Bibliothèque scientifique numérique : bibliothequescientifique-numerique.fr

Gallica : gallica.bnf.fr

Les bibliothèques ne se sont intéressées à la préservation à long terme des documents numérisés qu'assez récemment, à partir de 2000 pour les plus avancées, et la norme en vigueur – l'OAIS (Système ouvert d'archivage d'information) – n'a été définie qu'en 2002. Si les bibliothèques nationales ont développé leurs propres infrastructures d'archivage (SPAR pour la Bibliothèque nationale de France, E-Depot pour la Bibliothèque royale des Pays-Bas, etc.), les autres bibliothèques recourent à des prestations extérieures d'archivage pour leurs propres collections, soit auprès de leur bibliothèque nationale, soit auprès d'autres organismes, privés ou publics.

En France, dès 2008, le Centre informatique national de l'enseignement supérieur (CINES) a été missionné pour l'archivage pérenne des bibliothèques universitaires par le ministère de l'Enseignement supérieur et de la Recherche *(voir l'article de F. Daumas et M. Massol dans ce numéro)*. Le CINES est notamment au cœur du dispositif envisagé par la Bibliothèque scientifique numérique pour archiver les contenus numériques produits par l'enseignement supérieur national.

Néanmoins, avec un coût estimé aujourd'hui à 5 000 euros par téraoctet et par an, l'archivage numérique représente une charge financière considérable pour les établissements. Ainsi, des politiques documentaires spécifiques ont été définies par les bibliothèques, déterminant des priorités parmi les natures de données à préserver. Toutefois, la question de la préservation à long terme des données numériques reste largement ignorée des porteurs de projets de numérisation. Et l'ampleur des coûts fait peser un risque fort sur la pérennité des résultats de ces projets.

Avec le numérique, les bibliothèques – notamment les bibliothèques universitaires et de recherche – ont été bouleversées en profondeur. Pourtant, leurs missions fondamentales – préserver le patrimoine documentaire, rendre accessible le savoir, accompagner leurs usagers dans la découverte et l'usage de ce savoir noyé dans l'immensité – restent intactes : le numérique leur permet d'accomplir ces missions avec davantage encore de pertinence et d'efficacité. Les professionnels de l'information numérique sont devenus, et deviendront encore davantage, des partenaires incontournables des projets de recherche, ainsi que de l'enseignement à distance, amené à révolutionner l'université dans les prochaines années. ■

Dictionnaires & encyclopédies à l'ère numérique

Rémi Mathis



L'encyclopédie en ligne *Wikipédia* et le *Wiktionnaire* modifient en profondeur les définitions mêmes de l'encyclopédie et du dictionnaire.

Alors que le livre numérique se développe et se diffuse à une vitesse mesurée, les encyclopédies ont connu, sous l'effet du numérique, des bouleversements majeurs. Et pourraient en subir de nouveaux. L'intérêt d'un support permettant d'accompagner les textes d'un grand nombre d'images, voire de films et de sons, a vite été compris des éditeurs. Entre 1985 et 1995, ils proposent tour à tour des versions sur CD-ROM de leur encyclopédie. L'avantage est évident : plus besoin de manipuler de lourds volumes, ni de les stocker chez soi, possibilité de rechercher un terme ou un concept dans le contenu, ou de rebondir d'un article à l'autre au sein d'une navigation simple et efficace, sans compter l'extrême réduction du coût de fabrication de l'objet en lui-même.

Jusqu'alors, le marché de l'encyclopédie était bien balisé, avec des marques de référence dans chaque pays : *Britannica* en Angleterre, *Universalis* en France, *Brockhaus* en Allemagne. Le changement de support rebat les cartes en attirant sur le marché de nouveaux venus : dès 1993, le géant de l'informatique *Microsoft* propose sa propre encyclopédie multimédia, *Encarta*, disponible en sept langues. L'entreprise se fonde sur des technologies développées pour d'autres usages (par exemple son logiciel *MapPoint* pour les cartes) afin de rendre la consultation plus efficace et agréable. Elle change aussi complètement le modèle intellectuel de l'encyclopédie.

Une encyclopédie était alors un produit éditorial de luxe, au coût élevé, dont les articles étaient rédigés par les meilleurs spécialistes de la discipline. *Encarta* a une approche différente : *Microsoft* achète des droits non exclusifs de diffusion à des encyclopédies existantes (*Funk&Wagnalls*) et rachète d'autres encyclopédies (*Collier*, *New Merit Scholar*). Le contenu est traduit, mais varie selon les pays, car des partenariats spéciaux sont signés : la version en néerlandais reprend, par exemple, des contenus de l'encyclopédie locale *Winkler Prins*.

Le résultat est un corpus d'environ 40 000 articles (plus ou moins selon les langues), mélanges de textes divers qui ne sont pas signés et ne font pas référence à des sources externes. Le texte est d'abord

L'AUTEUR



Rémi MATHIS est conservateur au Département des estampes de la Bibliothèque nationale de France. Il est aussi président de *Wikimédia France*.

BIBLIOGRAPHIE

R. Mathis, **Wikipédia. Une somme originale de copies,** *Medium*, n° 32-33, pp. 356-377, 2012.

R. Mathis, **Wikipédia, espace fluide, espace à parcourir,** *Revue de la BNU*, n° 6, pp. 52-61, 2012.

R. Mathis, **Défiances et production. Les bibliothèques françaises et *Wikipédia*,** *Bulletin des bibliothèques de France*, vol. 1, pp. 10-13, 2011.

© Pour la Science - n° 433 - Novembre 2013

© Ursa Lye Stu, Dshutterstock.com



diffusé sur CD-ROM, puis sur Internet. Ce nouvel acteur, qui distribue son encyclopédie selon des canaux nouveaux, est un concurrent redoutable pour les encyclopédies traditionnelles : le chiffre d'affaires de *Britannica* est divisé par deux entre 1990 et 1996.

Créée en 2001, l'encyclopédie en ligne *Wikipédia* prend vite une place proéminente sur le marché. Le changement est radical, car son modèle se fonde sur des habitudes venues d'Internet et non de l'édition. L'idée est de bâtir une encyclopédie de façon collaborative selon les méthodes du logiciel libre : chacun peut apporter sa pierre à l'édifice selon des règles mises au point par la communauté. Il n'y a pas de rédacteur en chef, tous les contributeurs sont à égalité et résolvent les éventuels conflits par la discussion et la recherche de consensus.

La croissance de *Wikipédia* est exponentielle. Dès 2005, la version anglophone propose 500 000 articles – le double l'année suivante. La version francophone, elle, atteint le demi-million d'articles en 2007… Plusieurs centaines de versions linguistiques se développent (285 en 2013), dans des langues qui n'avaient jamais connu d'encyclopédie auparavant, telles que le wolof ou des langues indiennes parlées par plusieurs dizaines de millions de personnes.

Le modèle étonne et est parfois remis en cause : des erreurs grossières sont relevées par des journalistes et des polémiques naissent sur la fiabilité de *Wikipédia*. Surtout, *Wikipédia* place son lecteur dans une situation nouvelle. Il ne lui est plus demandé de croire ce qui est écrit en partant du principe qu'un spécialiste a peu de chance de faire des erreurs, mais il lui est fourni un texte, de qualité variable, en transformation perpétuelle. Au lecteur d'adapter sa lecture à ce qu'il trouve et de faire preuve d'esprit critique. Autre différence : *Wikipédia* met en valeur la « neutralité de point de vue ». Ses contributeurs ne prennent pas parti sur des polémiques ou des débats ; ils se contentent de rapporter (sources à l'appui) les différents points de vue jugés pertinents par la communauté des spécialistes.

En 2009, face au succès public de *Wikipédia*, *Microsoft* ferme *Encarta*, et les éditeurs s'adaptent au lectorat numérique. *Britannica* met un terme à la publication sur papier en mars 2012, *Universalis* en 2013. Ils mettent l'accent sur le fait que les articles sont écrits par des spécialistes, ce qui est censé être un gage de qualité. Surtout, ils cherchent de nouveaux modèles économiques compatibles avec l'existence de *Wikipédia* : un abonnement à *Universalis* ne vaut que 50 euros par an (contre 5 000 euros sur papier, 3 000 pour la dernière édition).

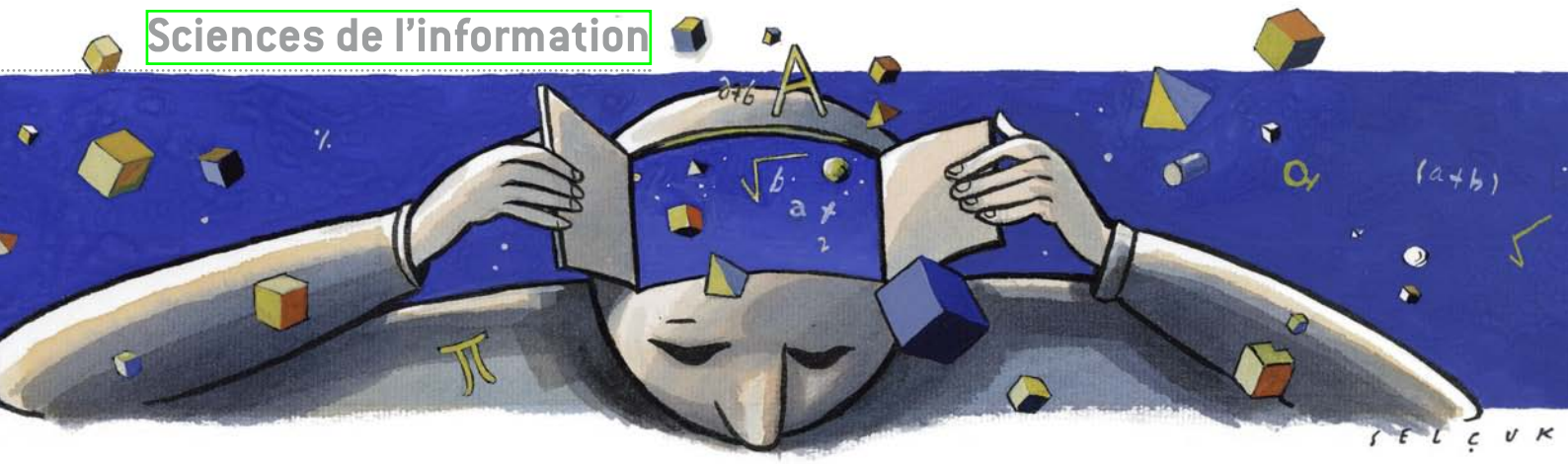
Chez *Britannica*, afin de souffrir la comparaison avec la capacité de *Wikipédia* à être à jour en temps réel, les articles sont repris et rédigés par des employés (non experts). Surtout, la vente de l'encyclopédie ne représente plus que 15 pour cent du chiffre d'affaires de la société, le reste étant constitué de services liés à la formation en ligne et au marché de l'éducation. Déficitaire dans les années 1990, *Britannica* fait des bénéfices depuis 2003. Le passage au numérique aura servi de catalyseur pour réinventer un modèle et se tourner vers les besoins des lecteurs.

La définition même des publications évolue. Une encyclopédie traitait autrefois du savoir légitime, à dominante littéraire et historique. Les possibilités offertes par le numérique – telle l'absence de contrainte de place – a ouvert l'encyclopédie à une autre catégorie de savoir, technique, scientifique ou populaire. Certains y retrouvent l'esprit de l'*Encyclopédie* de Diderot et d'Alembert, dont les auteurs allaient jusque dans les ateliers des artisans. D'autres, tenants de la tradition, se désolent de voir s'effacer le caractère légitimant de l'encyclopédie – et donc son rôle de guide et de juge.

Quoi qu'il en soit, les canons sont délaissés et le relativisme gagne du terrain : tout est bon à savoir, pensent de plus en plus de lecteurs, et ce n'est pas à une entreprise privée – fût-elle travailler des experts – de nous expliquer ce qui doit être su et ce qui est sans intérêt. Les dictionnaires subissent la même évolution. Si *Larousse* et le *Robert* continuent à décider quels mots peuvent être utilisés, le *Wiktionnaire*, cousin de *Wikipédia*, référence tous les mots, du moment qu'ils sont attestés dans des publications d'auteurs. La légitimité d'une instance chargée de faire le tri dans l'usage des mots s'affaiblit.

Le paysage évolue à une rapidité telle qu'il est difficile d'anticiper les dix prochaines années. L'accès mobile vient encore rebattre les cartes, et les modèles éditoriaux et économiques sont mal assurés. *Wikipédia* est de plus en plus au centre d'Internet par les données sémantiques qu'elle propose… mais peut s'effondrer en quelques années si les contributeurs se désintéressent de la rédaction et de la maintenance du projet. L'avenir dépendra des relations entre les industries d'Internet, les lecteurs – dont les pratiques évoluent sans cesse – et les éditeurs des outils de diffusion du savoir. Si *Wikipédia* révèle que le lecteur est encore à la recherche d'une connaissance organisée et centralisée, le développement du Web sémantique – un mouvement qui vise à faire émerger de nouvelles connaissances de celles déjà présentes sur Internet – et son utilisation dans les moteurs de recherche pourraient faire disparaître la notion d'encyclopédie ou de dictionnaire en tant qu'œuvre en soi. Autant dire que des surprises peuvent vite arriver. ■

LE WEB SÉMANTIQUE POURRAIT
faire disparaître la notion d'encyclopédie ou de dictionnaire en tant qu'œuvre en soi.



La science accessible à tous

Christine Berthaud et Agnès Magron

Bien que certaines revues scientifiques demeurent payantes, de plus en plus d'articles de recherche sont en accès libre sur Internet. Les enjeux de ce mouvement sont tant scientifiques qu'économiques.

Connaissez-vous Jack Andraka ? Cet Américain de 15 ans a mis au point un test de dépistage du cancer du pancréas simple et peu coûteux. Après la mort d'un proche atteint de cette maladie, l'adolescent a inventé tout seul sa méthode révolutionnaire en faisant des recherches sur Internet et en lisant des études médicales libres d'accès. La performance est surprenante. Toutefois, certains ne s'étonneront pas, à l'heure du tout gratuit sur le Web, qu'un internaute ait pu compulser gratuitement de la littérature savante. Or cette situation a été obtenue de haute lutte : elle est le fruit d'un mouvement international – le libre accès – né au sein d'une communauté scientifique déterminée à se réapproprier la diffusion des résultats qu'elle produit.

L'édition scientifique représente un marché lucratif, notamment celui des sciences dites exactes (chimie, médecine, biologie, physique), évalué à 17,3 milliards d'euros en 2011. Publier est un enjeu majeur pour les chercheurs : c'est par ce moyen qu'ils font connaître leurs travaux, sont reconnus par les collègues de leur discipline, progressent dans leur carrière, obtiennent des financements. Ils ne perçoivent en général pas de droits d'auteur pour leurs publications. Ce sont aussi les chercheurs qui expertisent les textes soumis aux éditeurs : ce processus d'évaluation, nommé revue par les pairs, est indissociable de la publication scientifique. Il permet de valider les résultats présentés dans le texte, en particulier d'un point de vue méthodologique. Ce travail de relecture n'est pas rétribué par les éditeurs : être considéré comme « expert » constitue en quelque sorte une rémunération symbolique. Le support de diffusion le plus utilisé est la revue, même si, en sciences humaines et sociales, les livres restent importants.

L'édition scientifique : un marché contrôlé

La plupart des lecteurs sont des chercheurs, mais ce sont en majorité (68 à 75 pour cent) les bibliothèques universitaires ou de recherche qui acquièrent les ouvrages et financent les abonnements aux revues. Les prix sont si prohibitifs (en moyenne 4450 dollars l'abonnement annuel à une revue en chimie ou 2322 dollars en astronomie) que seuls des budgets institutionnels peuvent les supporter. Et les prix ne cessent d'augmenter, surtout depuis l'arrivée du numérique.

Les éditeurs ont justifié un temps ces augmentations par les investissements qu'ils devaient faire pour s'adapter et numériser des archives (qu'ils revendent après d'ailleurs). Ils ont surtout affaire à une clientèle captive : une bibliothèque se doit de proposer à ses usagers les revues incontournables de leurs disciplines. Le numérique a renforcé cette dimension : les revues sont proposées dans des bouquets composés par des éditeurs, et les bibliothèques doivent négocier des licences d'accès pour leurs utilisateurs.

Qui vend ? Le marché de l'édition scientifique est dominé par des entreprises privées, certaines cotées en Bourse. Quatre grandes sociétés contrôlent le marché : *Reed-Elsevier* (qui affiche un portefeuille de 2000 titres de revues), *Springer-Verlag*, *Blackwell-Riley* et *Thompson Reuters*. Cette situation de monopole cristallise les critiques de la communauté scientifique : elle y voit un détournement par des opérateurs privés de fonds publics finançant la recherche aussi bien en amont (salaires des chercheurs) qu'en aval

© Pour la Science - n° 433 - Novembre 2013

© CNSD - Centre pour la communication scientifique directe

(subventions des bibliothèques). De plus, la pression financière exercée sur les bibliothèques menace la diffusion des résultats.

Le passage au numérique a permis de s'affranchir des contraintes de la distribution papier. L'appropriation forte et rapide de ce nouveau média par certaines communautés de chercheurs a favorisé l'émergence de réalisations qui ont bouleversé l'accès à la littérature scientifique.

De *ArXiv* au libre accès

Les physiciens ayant besoin de partager vite les résultats de leurs recherches, l'un d'eux, Paul Ginsparg, a créé en 1991, au Laboratoire américain Los Alamos de l'Université de Californie, aux États-Unis, le premier serveur d'archives ouvertes, *hep-th (High Energy Physics - Theory)*, aujourd'hui connu sous le nom d'*ArXiv* (prononcé *arkaïv*). Ce serveur, sur lequel les chercheurs déposent leurs articles (publiés ou non) sous la forme de fichiers, a servi de modèle à d'autres projets répondant à la définition de ce que, un peu plus tard, on a nommé l'*Open Access*, le libre accès aux publications scientifiques.

Ce nouveau concept a été formalisé en 2002 par une déclaration, la *Budapest Open Access Initiative* (BOAI), qui définit deux stratégies pour mettre en accès libre les publications des chercheurs : la voie verte, c'est-à-dire la mise en ligne des écrits sur le Web par leurs auteurs, et la voie dorée – la publication dans des revues électroniques en accès ouvert. Dès 2003, les grandes institutions de la recherche et de l'enseignement supérieur s'engagent à soutenir les politiques d'accès ouvert, notamment en incitant leurs chercheurs à déposer leurs publications sur des serveurs d'archives ouvertes au sein des bibliothèques universitaires et des institutions. En juillet 2012, la Commission européenne a demandé aux États membres de définir une politique intitulée « Pour un meilleur accès aux informations scientifiques : dynamiser les avantages des investissements publics dans le domaine de la recherche ».

En France, la voie dorée est représentée notamment par le Centre pour l'édition électronique ouverte (CLEO), qui gère Openedition, un portail de revues en sciences humaines et sociales. Pour la voie verte, le CNRS a créé en 2000 le Centre pour la communication scientifique directe (CCSD), dont la principale mission est la gestion de l'archive ouverte HAL (Hyper articles en ligne) interconnectée aux grands serveurs internationaux tel *Pubmed Central*, de l'Institut américain de la santé, proposant de nombreuses archives en sciences de la vie.

Contrairement à *ArXiv* dont elle s'est pourtant inspirée, HAL rassemble des publications issues de tous les domaines scientifiques, déposées par les auteurs eux-mêmes dans plus de 60 pour cent des cas ou par des professionnels de la documentation. Ces dépôts sont faits dans le respect du droit de chaque acteur de la publication. Cette plate-forme, dont l'illustration page ci-contre est l'emblème, est utilisée par l'ensemble des établissements français, universitaires ou de recherche. Faisant partie de la mise en œuvre de la politique demandée par la Commission européenne, elle est aussi soutenue par le ministère de l'Enseignement supérieur et de la Recherche.

Ce projet permet de prendre en compte des aspects techniques lourds, telle la possibilité de garantir l'accessibilité aux textes par des liens URL solides et d'apporter des solutions au problème de la pérennité à long terme des données par des mises à jour régulières des fichiers déposés.

Les éditeurs considèrent le libre accès comme une menace pour leur modèle économique. Ils tentent de différer le dépôt dans une archive ouverte en imposant aux auteurs un embargo de six mois à deux ans après la date de publication, mais leur proposent aussi de payer, avant publication, pour que leur article soit en libre accès. Ce modèle est dit hybride, car si les revues sont disponibles sur abonnement, le téléchargement de certains articles est gratuit. Les revues en libre accès bousculent les modèles économiques. Certaines reportent vers l'auteur les coûts de production de la revue. Le financement n'intervient donc plus en aval de la publication sous forme d'abonnements, mais en amont grâce au paiement par l'auteur d'une somme forfaitaire pour chaque article accepté. En général, cette somme est subventionnée et prise en charge par son institution ou par l'organisme de financement de la recherche.

Aujourd'hui, les enjeux du libre accès et de la pérennité se généralisent aux données produites au cours de la recherche, tels les résultats d'expériences ou les données statistiques. Cela concourt à la validation des résultats en facilitant leur reproduction par d'autres équipes, ainsi qu'au libre accès de l'ensemble des recherches financées par des fonds publics. ■

HAL en chiffres

2 800 nouveaux fichiers y sont déposés par mois.

236 000 fichiers y étaient hébergés en septembre 2013.

1,5 téraoctet de données y est stocké.

286 000 auteurs y ont déposé au moins un article.

2 500 collections (institutionnelles, thématiques…) y ont été créées.

À ÉCOUTER

Le jeudi 31 octobre de 14h à 15h, Christine Berthaud reviendra sur les enjeux du libre accès dans la partie Actualités de l'émission **La marche des sciences**, sur *France Culture*. www.franceculture.fr

LES AUTEURS





Christine BERTHAUD est directrice du Centre pour la communication scientifique directe (CCSD, CNRS-UPS2275), à Villeurbanne.

Agnès MAGRON est gestionnaire de communauté au sein du CCSD.

SUR LE WEB

ArXiv, archive ouverte en physique et en mathématiques : arxiv.org

HAL, archive ouverte couvrant tous les domaines scientifiques : hal.archives-ouvertes.fr

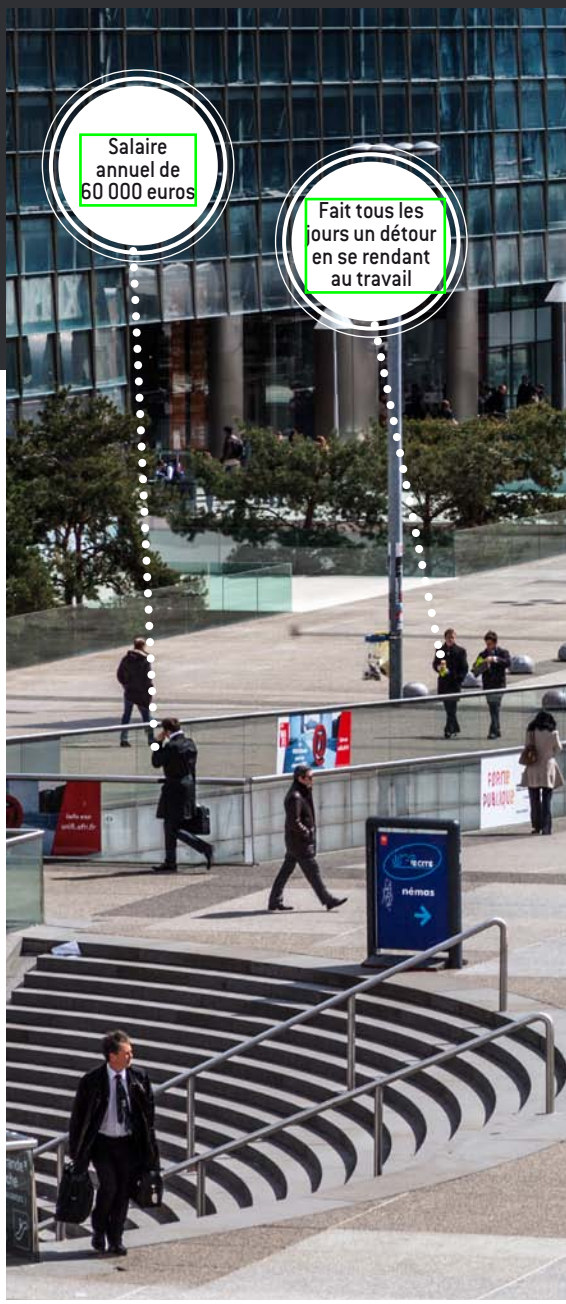
Le CCSD, concepteur de HAL et chargé de sa gestion : ccsd.cnrs.fr

Le Centre pour l'édition électronique ouverte : cleo.openedition.org

Orienter la société grâce aux **données massives**

Les traces numériques que chacun de nous laisse derrière lui chaque jour pourraient devenir un cauchemar pour notre vie privée… ou servir à construire un monde plus sain et plus prospère.

Alex Pentland



Dès le milieu du XIXe siècle, la croissance urbaine rapide, encouragée par la révolution industrielle, avait engendré des problèmes sociaux et environnementaux majeurs. Les villes y ont répondu en construisant des systèmes centralisés pour fournir de l'eau, de la nourriture et de l'énergie, favoriser le commerce, faciliter les transports, maintenir l'ordre et donner un accès aux soins. Aujourd'hui, ces solutions plus que centenaires sont de moins en moins adaptées. Nos grandes villes sont surpeuplées et encombrées par les bouchons. Et nous sommes confrontés à de nouveaux défis, en particulier pour nourrir et loger une population mondiale qui devrait croître de deux milliards d'habitants, tout en limitant au mieux l'impact du réchauffement climatique.

Ces problèmes typiques du XXIe siècle appellent une réflexion nouvelle. Mais de nombreux économistes et sociologues abordent encore les questions sociales avec des concepts d'un autre âge, tels que les marchés et les classes, dans des modèles simplifiés qui réduisent les interactions sociales à des règles générales et ignorent les comportements individuels. Il faut examiner la situation plus en profondeur et prendre en compte les détails les plus fins des interactions sociales.

Les technologies numériques nous le permettent aujourd'hui. Elles offrent une vue unique sur les milliards d'interactions individuelles par lesquelles les gens échangent des idées, de l'argent, des biens ou des informations. Le Laboratoire de dynamique humaine que je dirige au MIT (l'Institut de technologie du Massachusetts) explore cette masse de données pour tenter de faire émerger les lois qui régissent ces interactions. Par cette approche, nous commençons à pouvoir prédire certains phénomènes tels que les chutes des cours de la Bourse, les résultats d'élections politiques ou les épidémies de grippe. L'analyse des données sociales peut nous aider à bâtir des systèmes financiers stables, des institutions politiques qui fonctionnent, des systèmes de soins efficaces et économiques, et bien d'autres choses. Mais nous devons d'abord prendre la pleine mesure du poten-

1. LES TRACES NUMÉRIQUES que nous laissons derrière nous permettent de retracer et de prédire nos comportements individuels.

© Pour la Science - n° 433 - Novembre 2013

tiel des données massives, le « *Big Data* », et construire un cadre juridique pour les exploiter à bon escient. La capacité à suivre, prédire et même influer sur le comportement des individus et des groupes n'est pas sans rappeler le feu prométhéen : elle peut servir au meilleur comme au pire.

Le pouvoir de prédiction des miettes numériques

Pendant que nous vaquons à nos occupations quotidiennes, tels des Petits Poucets, nous semons dernière nous des miettes virtuelles. Les personnes que nous appelons au téléphone, les endroits où nous allons, les aliments que nous mangeons ou les produits que nous achetons laissent des traces numériques. Ces petits détails en disent bien plus long sur notre vie que ce que nous choisissons d'en révéler nous-mêmes. Nos publications sur les réseaux sociaux ne livrent que des informations que nous avons choisi de divulguer à d'autres personnes, sélectionnées sur la base des critères du moment. Les traces numériques, en revanche, enregistrent la réalité de notre comportement.

Nous sommes des animaux sociaux, et notre comportement n'est jamais aussi unique que nous aimerions le croire. Les gens que vous appelez, à qui vous envoyez des textos et avec qui vous passez du temps (même les habitants de votre quartier, que vous reconnaissez sans les avoir directement rencontrés) vous ressemblent probablement à de nombreux égards. Mon équipe de recherche a montré qu'il est possible de prédire votre risque de développer un diabète simplement en examinant la liste des restaurants que vous fréquentez et des gens avec qui vous sortez. Les mêmes données révèlent le genre de vêtements que vous avez tendance à acheter ou votre capacité à rembourser un emprunt. Et parce que



© Shutterstock.com / Radu Razvan

L'ESSENTIEL

- L'organisation de la société s'appuie sur des principes datant du début de la révolution industrielle. Pour répondre aux défis du XXIe siècle, une organisation sociale plus efficace est nécessaire.
- L'analyse de la masse de données numériques que nous laissons derrière nous au quotidien fait émerger de nouvelles idées pour s'attaquer aux problèmes sociaux.
- Mais nous devons d'abord donner aux individus plus de contrôle sur leurs informations avant d'exploiter celles-ci.

notre comportement change quand nous sentons que nous tombons malade – nous n'allons pas aux mêmes endroits, nous n'achetons pas les mêmes choses, nous n'appelons pas les mêmes personnes et nous faisons des requêtes différentes sur Internet –, il est possible grâce à l'analyse de données de dresser une carte qui prédit heure par heure l'endroit où vous risquez le plus d'attraper la grippe, par exemple.

Parmi les motifs qui se dégagent des données massives, ceux qui éclairent le plus le fonctionnement de la société sont ceux qui concernent le flux d'idées et d'informations entre les personnes. Nous pouvons mesurer et visualiser ce flux en étudiant les interactions sociales (conversations en face à face, appels téléphoniques, messages sur les réseaux sociaux...), les comportements d'achat des individus (*via* les données des cartes bancaires) ou les déplacements (*via* les données de géolocalisation par GPS). Le flux d'idées est central dans la compréhension de la société, non seulement parce qu'être informé en temps utile est crucial pour construire des systèmes efficaces, mais aussi parce que le partage et l'association des idées sont au fondement de l'innovation. En général, les communautés coupées du reste de la société stagnent.

L'une des découvertes les plus surprenantes que nous avons faites est que les motifs formés par le flux d'idées sont directement reliés à l'augmentation de la productivité et à l'innovation. Les groupes et les organisations dont les membres interagissent et tissent des liens en dehors du groupe ont globalement une productivité plus élevée, une plus grande créativité, et leurs membres vivent même plus longtemps et en meilleure santé. Cette règle s'observe d'ailleurs chez toutes les espèces sociales. L'échange d'information semble essentiel à la bonne santé de toutes les sociétés.

On peut se représenter les entreprises ou les organisations comme des machines à idées. Ces machines agrègent et diffusent des idées, principalement par des interactions entre individus. Deux grandeurs permettent de mesurer la valeur d'un flux d'idées. La première est l'engagement, c'est-à-dire la proportion d'échanges d'individu à individu au sein du groupe. La relation entre l'engagement et la productivité est simple : un niveau d'engagement élevé prédit une productivité élevée du groupe, quel que soit (ou presque) le sujet sur lequel travaille le groupe ou le type de personnalité de ses membres. Le second facteur est l'exploration, c'est-à-dire la capacité des membres du groupe à établir de nouvelles connexions avec des personnes extérieures et à rapporter au sein du groupe de nouvelles idées provenant de l'extérieur. L'exploration est un bon indicateur à la fois de l'innovation et de la production créative.

Dans des expériences de terrain menées dans des entreprises, mes étudiants et moi avons mesuré le niveau d'engagement et d'exploration en équipant les employés de badges sociométriques, des dispositifs qui enregistrent les interactions en face à face. Nous avons découvert qu'en augmentant l'engagement au sein d'un groupe, on peut améliorer la productivité de façon spectaculaire tout en réduisant le stress. Par exemple, après avoir appris qu'un centre d'appels organisait les pauses de ses employés de telle façon qu'une seule personne prenne sa pause à la fois, j'ai convaincu le directeur de prévoir des pauses café avec plusieurs personnes en même temps, afin d'augmenter l'engagement entre les employés. Ce simple changement s'est traduit par un gain de productivité de 11 millions d'euros par an.

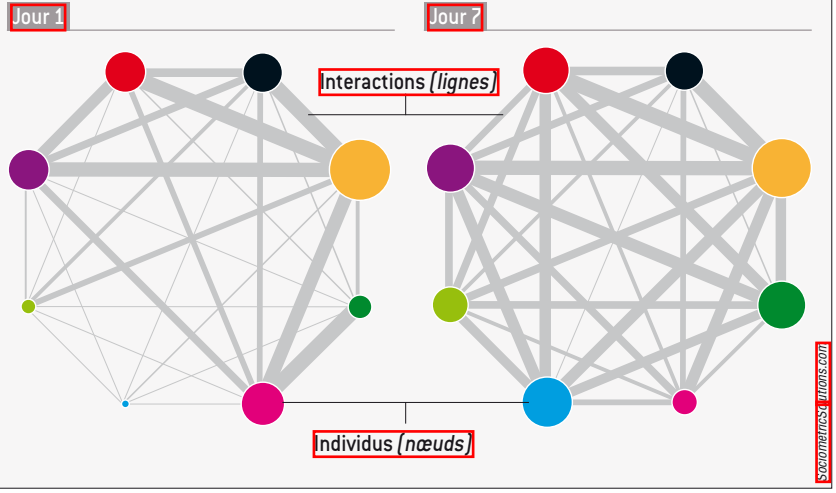
L'AUTEUR



Alex PENTLAND est directeur du Laboratoire de dynamique humaine du MIT, l'Institut de technologie du Massachusetts, aux États-Unis, et codirige les projets du Forum économique mondial en matière de *Big Data* et données personnelles.

Dynamique de groupe

En suivant les interactions sociales à l'aide de badges d'identification équipés de capteurs ou grâce aux données de téléphonie mobile, on peut mettre en évidence les dynamiques de groupe, ce qui permet de visualiser comment les membres d'une équipe collaborent. Les diagrammes ci-dessous illustrent la dynamique d'un groupe de réflexion de huit personnes étudié par les auteurs. La taille des cercles représente l'ampleur de la communication d'un membre avec le reste du groupe. À la fin de chaque journée, le diagramme de leurs interactions était présenté aux membres du groupe. En utilisant ces diagrammes, le groupe a pu identifier les points faibles et les corriger. L'équipe est ainsi devenue plus interactive et donc plus productive.



© Pour la Science - n° 433 - Novembre 2013

Les canaux de communication « riches », en particulier les interactions en face à face, ont beaucoup plus de poids que les canaux de communication électroniques. En d'autres termes, les courriels ne pourront jamais complètement remplacer les réunions et les conversations.

Nous avons également découvert que des phases successives d'exploration et d'engagement (les membres du groupe échangent entre eux, puis vont à la recherche d'informations nouvelles, les transmettent au groupe, et ainsi de suite) sont systématiquement associées à une meilleure production créative. Mes collègues ont observé ce schéma oscillant dans les interactions en face à face au sein d'organismes de recherche, et ont utilisé ces mesures pour identifier avec précision les journées les plus créatives des chercheurs. La même approche fonctionne avec les équipes virtuelles, dont les membres sont éparpillés dans divers laboratoires.

Des motifs similaires dans le flux d'information permettent de prédire la productivité globale d'une ville ou d'une région entière. L'engagement dans la communauté et l'exploration hors de celle-ci sont même reliés à des grandeurs sociales telles que l'espérance de vie, le taux de criminalité ou la mortalité infantile. Les quartiers qui sont des ghettos de l'information s'en sortent aussi mal que les ghettos physiques, alors que les quartiers dont les habitants sont engagés les uns avec les autres et connectés aux communautés environnantes sont plus prospères, et leurs habitants en meilleure santé.

Optimiser le flux d'idées

L'utilisation des données massives pour diagnostiquer les problèmes et prédire les succès est une chose. Mais ce qui est passionnant, c'est que l'on peut aussi utiliser ces données pour concevoir des organisations (entreprises, institutions politiques, collectivités) qui fonctionnent mieux que celles existant.

Ce potentiel s'illustre facilement pour les entreprises. En mesurant le flux d'idées, il est en général possible d'identifier des changements simples qui améliorent la productivité et la créativité. Le service de publicité d'une banque allemande avait par exemple des difficultés à lancer des campagnes publicitaires efficaces pour les nouveaux produits, et la banque souhaitait comprendre pourquoi. Lorsque nous avons étudié la situation avec des badges d'identification sociométriques, nous nous sommes aperçus que les différents services échangeaient beaucoup de courriers électroniques, mais que presque personne ne parlait aux employés du service commercial. La raison en était simple : il était situé à un autre étage. Inévitablement, le service de la publicité concevait des campagnes que le service commercial n'était pas en mesure de mettre en œuvre. Quand la direction a vu notre diagramme représentant cette rupture du flux d'information, elle a pris conscience qu'il fallait déménager le service commercial au même étage que les autres départements, ce qui régla le problème.

Augmenter l'engagement au sein d'un groupe n'est cependant pas une solution miracle. En fait, le faire sans développer l'exploration peut même être une source de problèmes. Par exemple, quand j'ai mesuré avec le postdoctorant Yaniv Altschuler le flux d'information au sein du réseau social d'investisseurs *eToro*, nous avons constaté qu'à partir d'un certain point, les gens deviennent tellement interconnectés que le flux d'idées est dominé par les boucles de rétroaction. Certes, tout le monde échange des idées, mais ce sont les mêmes qui reviennent sans cesse. Les traders sont dans une sorte de chambre résonnante. Ce travail en circuit fermé expliquerait l'apparition de bulles financières, comme la bulle Internet du début des années 2000, provoquée par des investissements déconnectés de la réalité économique.

Néanmoins, nous avons montré qu'il est possible d'influer sur le flux d'idées au sein d'une communauté grâce à de petites incitations. Des « coups de pouce » peuvent inciter des personnes isolées à s'engager davantage avec les autres ; ou encourager des personnes prisonnières de la façon de penser du groupe à s'informer hors de leurs cercles habituels. Dans une expérience impliquant 2,7 millions de petits investisseurs du réseau *eToro*, nous avons « régulé » la communauté en donnant à certains investisseurs des coupons de réduction qui les encourageaient à explorer les idées d'un panel d'investisseurs plus diversifié. Résultat : le comportement de l'ensemble du groupe est resté dans les limites raisonnables de la « sagesse du plus grand nombre ». Le plus remarquable est que nous avons réussi à augmenter la rentabilité globale de plus de six pour cent alors que nous n'avions octroyé les incitations qu'à un petit nombre d'investisseurs.

Le développement d'un flux d'idées peut également contribuer à résoudre la « tragédie des biens communs », bien connue en économie, c'est-à-dire une situation où l'intérêt pour chacun de profiter le plus possible d'une ressource commune dépasse le prix que chacun doit payer pour utiliser cette ressource. Le secteur des assurances santé en est un bon exemple. Les personnes qui négligent de faire de l'exercice, de surveiller leur alimentation, de prendre



Wen Dong et Alex Pentland

2. EN EXPLOITANT LES DONNÉES DES TÉLÉPHONES PORTABLES, on peut dresser la carte des zones où les individus sont le plus susceptibles d'attraper la grippe *(ici un exemple sur le campus de l'Institut de technologie du Massachusetts, probabilité croissante du rouge au violet)*.

leurs médicaments ou qui ont des comportements à risque engendrent des frais médicaux élevés, qui sont répercutés par les assurances sur le prix des cotisations payées par chacun. Ce surcoût individuel est cependant trop faible pour que quiconque agisse afin de résoudre le problème.

La solution habituelle est d'identifier les individus qui abusent du bien commun et de leur infliger des pénalités (ou leur accorder des incitations) pour les pousser à mieux se comporter. Cette approche est onéreuse et porte rarement ses fruits. En revanche, le doctorant Ankur Mani et moi avons montré qu'en encourageant un engagement accru entre les individus, on parvient à réduire le problème. La clé est de proposer de petites incitations financières à ceux qui interagissent le plus avec les contrevenants, en les récompensant eux plutôt que le contrevenant pour leur comportement plus vertueux. Dans les situations réelles (par exemple, pour pousser les gens à faire des économies d'énergie), nous avons montré que l'approche fondée sur la pression sociale est jusqu'à quatre fois plus efficace que les méthodes habituelles.

Une nouvelle donne pour les données

Cette même approche peut être utilisée pour la mobilisation sociale (dans les situations d'urgence ou quand un effort coordonné est nécessaire pour atteindre un but commun). En 2009, à l'occasion des 40 ans d'Internet, l'Agence américaine de recherche pour la défense (DARPA) a mené une expérience pour voir comment les réseaux sociaux et Internet pouvaient servir à une mobilisation d'urgence à l'échelle des États-Unis. L'agence offrait une récompense de 40 000 dollars à l'équipe qui trouverait le plus rapidement dix gros ballons rouges disséminés à travers le pays. Quelque 4 000 équipes se sont inscrites. Elles ont presque toutes adopté l'approche la plus simple, à savoir offrir une récompense à quiconque leur signalerait la présence d'un ballon. Mon équipe a choisi une stratégie différente. Nous avons décidé de partager le prix entre les personnes qui verraient un ballon et ceux qui auraient utilisé leurs réseaux sociaux pour recruter ces personnes. Cette stratégie, similaire à celle de la pression sociale évoquée précédemment, encourageait les gens à utiliser leurs réseaux sociaux autant que possible. Nous avons gagné la compétition en localisant tous les ballons en seulement neuf heures.

Pour mettre en pratique ces idées et améliorer le fonctionnement des systèmes sociaux en exploitant les données numériques personnelles, il est d'abord nécessaire d'établir ce que j'appelle une « nouvelle donne pour les données » : garantir que les données nécessaires soient facilement disponibles tout en protégeant la vie privée des citoyens. Selon moi, la clé est de traiter les données personnelles comme un bien : les individus auraient un droit de propriété sur les données qui les concernent. Que signifie « posséder » ses propres données ? En 2007, j'ai suggéré une analogie avec les principes de la propriété, de l'utilisation et de la cession dans le droit anglais :

Vous avez le droit de posséder les données qui portent sur vous. Indépendamment de l'entité qui recueille les données, celles-ci vous appartiennent, et vous pouvez y accéder à tout moment.

Vous avez le droit au contrôle total de l'utilisation de vos données. Cette utilisation ne peut se faire que par accord actif, et les termes doivent en être clairement expliqués. Si vous n'êtes pas satisfait de la façon dont une entreprise utilise vos données, vous pouvez lui en retirer l'accès.

Vous avez le droit de vous débarrasser de vos données ou de les distribuer. Vous pouvez faire détruire vos données ou les faire redéployer ailleurs.

Des laboratoires vivants

Ces cinq dernières années, au Forum économique mondial, j'ai contribué à animer le débat sur ces principes entre responsables politiques, dirigeants d'entreprises multinationales et groupes de défense des citoyens. Les réglementations sont en train d'évoluer. Aux États-Unis, la nouvelle charte de respect de la vie privée des consommateurs, et dans l'Union européenne, le projet de directive sur la protection des données personnelles, accordent ainsi aux individus un meilleur contrôle de leurs données, tout en encourageant une plus grande ouverture.

Pour la première fois dans l'histoire, nous pouvons observer la société avec une précision suffisante pour envisager de construire des systèmes sociaux plus performants que ceux ayant existé jusqu'à présent. Les données massives promettent

BIBLIOGRAPHIE

A. Pentland, **Society nervous system : building effective governements, energy and public health**, *Computer*, vol. 45(1), pp. 31-38, 2012.

A. Madan *et al.*, **Sensing the health state of a community**, *IEEE Pervasive Computing*, vol. 11(4), pp. 36-45, 2012.

S. Moturu *et al.*, **Using social sensing to understand the links between sleep, mood, and sociability**, *Proc. of IEEE SocialCom*, pp. 208-214, 2011.

A. Madan *et al.*, **Pervasive sensing to model political opinions in face-to-face networks**, *Proc. of the 9th international conference on Pervasive computing*, pp. 214-231, 2011.

A. Madan *et al.*, **Social sensing : Obesity, unhealthy eating and exercise in face-to-face networks**, *Proc. of ACM Wireless Health*, pp. 104-110, 2010.

A. Madan *et al.*, **A social sensing to model epidemiological behavior change**, *Proc. of ACM Ubicomp*, pp. 291-300, 2010.

A. Pentland et *al.*, **Life in the network : the coming age of computational social science**, *Science*, vol. 323, pp. 721-723, 2009.

Laboratoire de dynamique humaine au MIT : http://hd.media.mit.edu/

© Pour la Science - n° 433 - Novembre 2013

de transformer nos vies autant que l'invention de l'écriture ou d'Internet.

Bien sûr, cette transition vers une société orientée par les données est un défi. Dans un monde de données illimitées, la méthode scientifique à laquelle nous sommes habitués ne fonctionne plus : il y a tellement de connexions et de corrélations potentielles entre les données que les outils statistiques classiques fournissent souvent des résultats absurdes. L'approche scientifique classique donne de bons résultats quand l'hypothèse est claire et que les données sont conçues pour répondre à la question. Mais dans la complexité inextricable des systèmes sociaux à grande échelle, il existe souvent des milliers d'hypothèses raisonnables, et il est impossible d'ajuster les données à toutes ces hypothèses en même temps.

Ainsi, à l'avenir, nous devrons analyser les données du monde réel pour faire émerger les lois sous-jacentes plus systématiquement et plus en amont que jusqu'ici. Nous devons construire des « laboratoires vivants » où nous pourrons tester nos idées pour construire des sociétés orientées par les données.

Un exemple d'un tel laboratoire vivant est la Cité des données ouvertes que nous venons d'inaugurer à Trente, en Italie, avec la coopération des élus, de *Telecom Italia*, de *Telefonica*, de la fondation Bruno Kessler et de l'Institut pour la conception dirigée par les données (IDDD). Le but de ce projet est de démultiplier le flux d'idées au sein de la ville de Trente. Des logiciels tel que notre système openPDS (*Personal Data Store*), qui implémente les principes de la nouvelle donne pour les données, permettent aux individus de partager en toute sécurité des données personnelles (telles des informations relatives à leur santé ou à leur famille) en contrôlant où va l'information et quelle utilisation en est faite. Par exemple, une des applications d'openPDS encourage le partage entre les familles des bonnes pratiques concernant l'éducation des jeunes enfants. Comment les familles dépensent-elles leur argent ? Quels sont la fréquence de leurs sorties et leur niveau de socialisation ? Une fois que l'individu a donné son accord, ses données peuvent être rassemblées, rendues anonymes et partagées automatiquement en toute sécurité avec d'autres familles.

POUR LA PREMIÈRE FOIS DANS L'HISTOIRE nous pouvons observer la société avec une précision suffisante pour envisager de construire des systèmes sociaux plus performants.

Nous pensons que de telles expériences montreront que les avantages potentiels d'une société orientée par les données valent la peine de faire ces efforts et de prendre ces risques. Imaginez : nous pourrions anticiper les krachs financiers, détecter et prévenir les épidémies, utiliser au mieux les ressources naturelles et encourager la créativité pour prospérer. Ce rêve pourrait devenir bientôt une réalité... si nous évitons les écueils avec prudence. ■

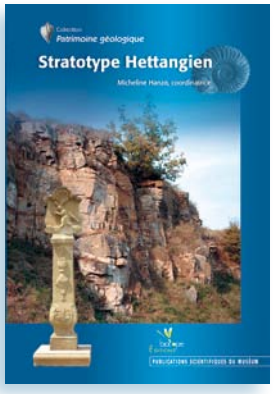
Muséum national d'Histoire naturelle

Le stratotype Hettangien

Coordonné par Micheline **Hanzo**

Le temps du géologue est divisé en étages définis par les fossiles qu'ils contiennent. Afin d'établir des références universelles, certains sites où les couches géologiques correspondant à ces étages affleurent ont été qualifiés de « stratotypes ».
La collection *Patrimoine géologique* a pour objet de publier des synthèses sur chacun des stratotypes situés en France.

Le stratotype de l'Hettangien est donc une référence géologique internationale. Il est de ce fait aussi la célébrité d'Hettange-Grande, localité de Moselle où il a été défini.
Suite à bien des controverses, c'est Eugène Renevier qui, en 1864, en fut l'initiateur, séduit par les travaux de son prédécesseur Olry Terquem, sur des niveaux particulièrement riches en fossiles.
Ce stratotype historique méritait d'être protégé ! Il obtint le statut de Réserve naturelle nationale d'Hettange-Grande en 1985. *Bildstock*, pavés, pierres à four, etc., témoignent de l'exploitation du matériau gréseux du site stratotypique.
Bien documenté, richement illustré, cet ouvrage constitue une synthèse des connaissances acquises sur l'Hettangien au niveau de son stratotype, et de la Lorraine belgo-luxembourgeoise, en évoquant plus brièvement d'autres sites du même âge en France.



317 p. • 232 figs • 35,90 € TTC • format 16,5 x 24 cm
ISBN 978-2-85653-688-9

ouvrage en vente à la librairie des Publications Scientifiques face à la Grande Galerie de l'Évolution

PUBLICATIONS SCIENTIFIQUES DU MUSÉUM

Commandes et renseignements :
Muséum national d'Histoire naturelle
Publications Scientifiques
Case postale 41 • 57 rue Cuvier • 75231 Paris cedex 05
Tél. : 01 40 79 48 05 • Fax : 01 40 79 38 40
diff.pub@mnhn.fr • http://www.mnhn.fr/pubsci

Comment préserver l'anonymat

Tristan Allard, Benjamin Nguyen et Philippe Pucheral

Presque tous nos comportements laissent des empreintes numériques. Les informaticiens conçoivent et développent des techniques pour que l'analyse de ces gisements de données ne compromettent pas la vie privée.

Santé, déplacements, achats, appels téléphoniques, réseaux sociaux, recherches d'information : toutes ces facettes de nos vies laissent des empreintes numériques sur Internet ou sur divers appareils. Elles sont stockées et classées dans des bases de données gigantesques, telles celles des moteurs de recherche, qui gardent l'historique des requêtes associées à une adresse IP (qui identifie un appareil connecté à Internet) pendant plusieurs années. Ces masses d'informations représentent une manne sans précédent pour les sociétés humaines *(voir l'article d'Alex Pentland dans ce numéro)*. Dans le domaine de la santé, par exemple, on peut analyser les caractéristiques individuelles de chaque patient, afin de mieux le prendre en charge, ou réaliser des études de santé publique : l'analyse du nombre de fois où le mot « grippe » est tapé dans un moteur de recherche dans une certaine région (reconstituée approximativement avec les adresses IP) renseigne ainsi sur l'importance locale de l'épidémie. De même, les données de mobilité pourraient être enregistrées par les systèmes de géolocalisation (GPS) et aider à optimiser l'aménagement urbain. Si l'analyse de données sur les individus est pratiquée depuis des millénaires, notamment lors des recensements démographiques, la quantité et la diversité des informations disponibles aujourd'hui en multiplient l'intérêt potentiel.

Cette manne d'informations fait cependant planer une menace, elle aussi sans précédent, sur la vie privée des individus. La plupart du temps, les données ne sont pas publiques, mais elles circulent tout de même entre différents acteurs : des ensembles de données médicales sont transmis à des organismes de recherche, les sites en ligne peuvent vendre (légalement ou non) une partie des données personnelles de leurs utilisateurs, etc. Certaines données sont même accessibles à tous sur Internet : le moteur de recherche de *Facebook* (uniquement disponible en anglais pour l'instant) permet, par exemple, de trouver les utilisateurs qui aiment tel sport, ont tel âge, habitent telle région... Autre exemple, le site *Netflix*, qui recommande des films en fonction des goûts de chacun, a publié des données de ses utilisateurs à l'occasion d'un concours visant à optimiser les algorithmes qui déterminent les films à recommander.

Un compromis entre utilité et protection

Or de nombreuses études montrent qu'il est souvent possible d'identifier la personne associée à un jeu de données, même quand ce jeu ne contient pas son nom, ni ses coordonnées. Des pans entiers de la vie privée peuvent être dévoilés, avec

© Pour la Science - n° 433 - Novembre 2013



Pour la Science © Shutterstock.com/Kritchanut

des préjudices multiples : discrimination au crédit bancaire ou à l'assurance selon l'état de santé, discrimination à l'emploi selon l'orientation sexuelle ou le groupe ethnique, etc. Protéger les données personnelles sans les rendre inexploitables est donc un enjeu majeur à l'ère du tout numérique. La diffusion de ces données se révèle être un art de l'équilibre, où l'on recherche le meilleur compromis entre utilité des données et protection des individus. C'est cet « art » que nous examinerons ici.

Le souci de la protection des données s'est accru pendant la seconde moitié du XX^{e} siècle, à mesure que l'informatique se généralisait. Il a fait naître un domaine de recherche spécifique dans les années 1970, visant à garantir un anonymat plus ou moins complet aux titulaires des données. Un ensemble de statistiques sur des données personnelles peut révéler beaucoup d'informations sur un individu précis. Par exemple, si un recensement liste le revenu moyen par habitant avec un découpage du territoire en carrés de 200 mètres de côté, le propriétaire d'un grand domaine risque d'être seul dans la bande de territoire considérée ; un individu qui le sait peut donc déduire son revenu de l'analyse des statistiques. On voit ici que la quantité de renseignements extractibles de données publiées dépend de ce que l'on sait par ailleurs. Le développement des ordinateurs ayant rendu les informations disponibles plus nombreuses et plus accessibles, le danger s'est multiplié.

Le phénomène s'est encore accéléré depuis le début des années 2000, avec l'essor d'Internet, qui permet de consulter et de croiser instantanément ou presque de multiples sources d'information (réseaux sociaux, annuaires, listes de diplômés, etc.). Lors de la publication de données, il est devenu crucial de prendre en compte les connaissances annexes accessibles à un

L'ESSENTIEL

- **De nombreuses données personnelles sont publiées sur Internet ou transmises à divers organismes en indiquant leurs titulaires par un pseudonyme.**
- **Cela ne suffit pas à garantir l'anonymat des données, d'où une menace sur la vie privée.**
- **De nombreuses techniques sont développées pour protéger les données sensibles sans les rendre inexploitables.**



1. LES REQUÊTES TAPÉES DANS UN MOTEUR DE RECHERCHES renseignent sur l'identité de celui qui les soumet. Ainsi, supposons que l'on sache que l'individu 6208 s'intéresse à l'arbre généalogique des Dupont et qu'il a cherché un coiffeur et un boulanger rue Férou. En croisant ces requêtes avec des données disponibles sur Internet, telles les pages blanches, on découvre qu'un certain Marc Dupont habite rue Férou. On peut alors avoir accès aux requêtes de cet individu et s'apercevoir, par exemple, qu'il s'intéresse à la dépression (sans doute en souffre-t-il) ou qu'il a consulté des pages où l'on propose des rencontres discrètes. Sa vie privée est menacée.

individu mal intentionné. En outre, les analystes se contentent de moins en moins de statistiques sur les données personnelles et demandent d'accéder directement à celles-ci, afin d'augmenter la précision de leurs études.

La conjonction de ces deux facteurs a exacerbé la nécessité d'une protection robuste. Au cours des dix dernières années, de nombreux chercheurs ont étudié la façon de publier des données tout en préservant la vie privée des individus concernés. Ils ont développé des méthodes dites d'assainissement des données, qui consistent à dégrader leur précision de façon contrôlée, afin de réduire la probabilité que l'on puisse réidentifier la personne correspondante ou accéder à des informations sensibles la concernant.

Les faiblesses de la pseudonymisation

Comment rendre anonyme un jeu de données ? Le moyen le plus ancien et le plus intuitif est de remplacer le nom de l'individu concerné par un pseudonyme ou un nombre – on parle de pseudonymisation. C'est la méthode qu'a utilisée l'entreprise AOL, en août 2006, lorsqu'elle a mis en ligne 20 millions de requêtes adressées à son moteur de recherche par 650 000 utilisateurs sur une période de trois mois (officiellement pour les rendre disponibles aux chercheurs de divers domaines). Les noms des utilisateurs étaient remplacés par des nombres aléatoires, mais les mots clefs des recherches étaient laissés en clair. Quelques jours plus tard, des journalistes du *New York Times* ont identifié la personne portant le pseudonyme 4417749. Elle avait tapé des centaines de requêtes, tels « paysagiste à Lilburn, Géorgie », « individus dont le nom de famille est Arnold » ou « hommes célibataires de 60 ans ». Les journalistes ont croisé ces mots clefs avec des données disponibles sur le Web et ont peu à peu reconstitué l'identité probable de l'individu correspondant, une veuve de 62 ans habitant Lilburn, qu'ils ont fini par retrouver.

De nombreux cas similaires illustrent l'échec de la pseudonymisation à rendre les données anonymes. Remplacer les seuls champs directement identifiants, tels que le nom et le prénom ou le numéro de sécurité sociale, ne prémunit pas contre une réidentification ultérieure *(voir la figure 2)*. Pourtant, la pseudonymisation de données reste aux yeux de la loi une forme acceptable d'anonymisation, sans doute parce que ses failles sont mal comprises par les législateurs.

Prévenir les croisements de données

Au début des années 2000, Latanya Sweeney, de l'Université Carnegie Mellon, aux États-Unis, a proposé une méthode, nommée *k*-anonymat, pour prévenir les réidentifications *via* des croisements de données. Elle a d'abord montré qu'il était parfois possible de retrouver le titulaire d'un jeu de données médicales pseudonymisé à partir des champs {Sexe, Date de naissance, Code postal}. En effet, ces champs sont aussi présents dans d'autres jeux de données comprenant le nom du titulaire *(voir la figure 2)*. En outre, la combinaison des renseignements correspondants est

LES AUTEURS


Tristan ALLARD est postdoctorant à l'Inria (Institut national de recherche en informatique et en automatique).


Benjamin NGUYEN est chercheur à l'Inria et maître de conférences à l'Université de Versailles Saint-Quentin-en-Yvelines.


Philippe PUCHERAL est chercheur à l'Inria et professeur à l'Université de Versailles Saint-Quentin-en-Yvelines.

© Pour la Science - n° 433 - Novembre 2013

unique pour la plupart des gens, selon plusieurs études récentes – qui portaient sur les États-Unis, mais c'est probablement aussi le cas ailleurs.

Pour empêcher les réidentifications, L. Sweeney a proposé de scinder les champs du jeu de données en deux catégories : les champs quasi identifiants (QID, pour *Quasi identifiers data*), dont la combinaison de valeurs peut être unique, et les champs sensibles (SD, pour *Sensitive data*), qui doivent rester privés (par exemple, un diagnostic médical). Les valeurs des quasi-identifiants d'un individu sont ensuite rendues identiques à celles d'au moins $(k - 1)$ autres individus dans le jeu de données, pour un entier k convenablement choisi. En d'autres termes, le k-anonymat dissimule chaque individu dans une foule d'au moins k personnes impossibles à distinguer par leurs quasi-identifiants. Tout croisement avec une autre source de données aura ainsi une précision inférieure à $1/k$.

Les algorithmes du k-anonymat se fondent sur le principe de généralisation, qui consiste à remplacer les valeurs précises des quasi-identifiants par des ensembles de valeurs. Ceux-ci peuvent être des intervalles numériques ou des catégories. Par exemple, l'intervalle d'âges [15, 20] (c'est-à-dire de 15 à 20 ans) englobe plus d'individus qu'un âge précis. De même, la catégorie « France » englobe plus d'individus que la catégorie « Bourgogne ».

Il serait simple d'élaborer un algorithme qui parcourre les quasi-identifiants et forme un ensemble les incluant tous, tel {âge de 0 à 150 ans, habitant sur la Terre}. Mais les données seraient trop dégradées (c'est-à-dire trop imprécises) pour être utiles aux analystes. On a alors développé des algorithmes dits par partitionnement : ils forment des ensembles d'au moins k individus en regroupant les quasi-identifiants voisins, qu'ils remplacent ensuite par un intervalle ou une catégorie les incluant tous. Le plus utilisé est l'algorithme dit de Mondrian *(voir l'encadré page 66)*.

Cependant, le k-anonymat est loin de résoudre tous les problèmes. Considérons la situation suivante : un attaquant dispose d'un jeu de données k-anonyme et recherche la valeur sensible (tel le diagnostic médical) d'un individu cible dont il connaît le quasi-identifiant. Il peut retrouver le groupe auquel appartient sa cible, et donc l'ensemble des valeurs sensibles de ce groupe. La protection apportée par le k-anonymat est d'autant plus faible que le nombre de ces valeurs est petit. Dans le cas extrême, supposons que les k individus du groupe partagent la même valeur sensible, par exemple qu'ils aient tous un même cancer. L'attaquant sait alors que sa cible souffre de ce cancer : la valeur sensible n'est pas protégée.

Brouiller les données confidentielles

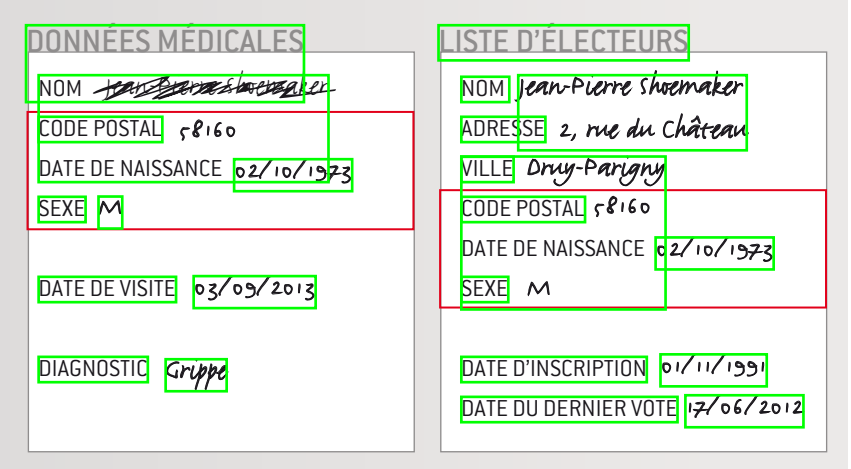
De multiples techniques ont succédé au k-anonymat. Elles cherchent à imposer une diversité minimale aux valeurs sensibles de chaque groupe, afin de limiter ce que peut apprendre un attaquant muni de divers renseignements sur sa cible. Certaines tentent d'estimer ces renseignements, afin de prendre en compte des attaques potentielles réalistes.

Ainsi, le modèle de Confidentialité bayésienne optimale, proposé en 2006 par Ashwin Machanavajjhala, de l'Université Cornell à New York, et ses collègues, quantifie la connaissance préalable qu'a l'attaquant de sa cible par la probabilité de trouver la valeur convoitée avant l'observation du jeu de données : si l'attaquant recherche le diagnostic médical d'une personne nommée Dupont, qu'il ne connaît de sa cible que son nom et qu'il sait que dix pour cent des Dupont ont un cancer, il a une chance sur dix de ne pas se tromper en affirmant que sa cible a un cancer. Cette connaissance est dite *a priori*. Après la découverte des données, l'attaquant a plus de chances de trouver la donnée qu'il cherche ; la connaissance de sa cible qui en résulte, connaissance dite *a posteriori*, est quantifiée par la probabilité de lui associer la bonne valeur sensible.

C'est ici que la théorie bayésienne intervient, car elle permet de calculer des probabilités conditionnelles : les probabilités de déduire telle ou telle information d'un jeu de données sachant qu'on a telle ou telle connaissance préalable. La confidentialité des données publiées est caractérisée par la différence entre connaissance *a posteriori* et connaissance *a priori*, autrement dit par la connaissance supplémentaire sur la cible apportée par les données. Assurer un niveau adéquat de protection revient à limiter cette valeur.

Cependant, cette technique est difficile à mettre en œuvre, car elle suppose de connaître exactement ce que l'attaquant sait de sa cible. C'est d'autant plus difficile qu'on ignore souvent qui sera le ou les attaquants – ces derniers pouvant même avoir des connaissances *a priori* différentes.

La l-diversité, méthode conçue la même année et par les mêmes chercheurs, est davantage applicable. Elle consiste à imposer la présence de l valeurs sensibles distinctes dans chaque groupe *(voir l'encadré page 66)*. Un attaquant qui ne connaît de sa cible que le groupe auquel elle appartient ne peut trouver sa valeur sensible avec une probabilité supérieure à $1/l$.



2. LE TITULAIRE D'UN DOSSIER MÉDICAL dépourvu de coordonnées précises *(à gauche)* peut souvent être retrouvé en croisant plusieurs jeux de données (de tels dossiers circulent parfois entre les hôpitaux et divers instituts de recherche ou de statistique). En effet, la plupart du temps, la combinaison {Code postal, Date de naissance, Sexe} *(encadrée en rouge)* est unique. Il suffit alors de trouver un autre jeu de données qui la contient et qui renferme aussi les coordonnées de la personne pour identifier cette dernière. C'est le cas, par exemple, d'une liste électorale *(à droite)*, disponible sur simple demande à la mairie.

De nombreuses autres techniques ont été élaborées au cours des années suivantes, afin de s'adapter à la diversité des données, des attaques et des attaquants. La « *t*-proximité » vise à créer des groupes au sein desquels la distribution des données est à peu près la même que dans la population globale. La « (*c*, *k*)-sûreté » prend en compte la capacité de l'attaquant à formuler *k* déductions logiques du type : « Si Adrien a la grippe, alors Bettie aussi. » La « 3D-confidentialité » considère que l'attaquant connaît *a priori* trois types de données : *l* valeurs sensibles que sa cible n'a pas, *k* valeurs sensibles que d'autres individus ont et *m* déductions logiques entre individus. La « *m*-invariance » protège contre les comparaisons entre les publications successives d'un même jeu de données, telles certaines informations statistiques sur un hôpital, dont les variations reflètent les arrivées, les départs et les évolutions des patients.

L'application de ces techniques commence souvent par celle de l'algorithme de Mondrian. Il sert à construire des groupes *k*-anonymes, au sein desquels on vérifie la conformité de la distribution des données sensibles vis-à-vis du modèle choisi *(voir l'encadré ci-dessous)*.

Volontairement ou non, tous ces modèles sont des exemples du paradigme de non-information, formulé en 1977 par le statisticien suédois Tore Dalenius : selon ce paradigme, un jeu de données est d'autant plus confidentiel qu'il renseigne peu l'attaquant sur sa cible. En 2006, Cynthia Dwork, alors chercheuse à *Microsoft*, a proposé une nouvelle façon de définir la confidentialité, qualifiée de confidentialité différentielle : un jeu de données serait confidentiel s'il n'est presque pas modifié par l'ajout des données d'un individu, quelles qu'elles soient.

L'assainissement consiste alors à perturber les données de telle sorte que celles de chaque individu se trouvent noyées dans la perturbation (et non plus dans la foule, comme dans le *k*-anonymat). On introduit, par exemple, un grand nombre de fausses données (typiquement 100 fois plus que de vraies), afin que le jeu de données qui contient la contribution d'un individu particulier ressemble beaucoup à celui qui ne la contient pas. Certains algorithmes suppriment aussi de vraies données. La protection est assurée par l'impossibilité de distinguer les fausses données des vraies dans le jeu de données assaini, dont on ne peut même pas garantir qu'il contient la contribution d'un individu précis.

Une vision alternative de la confidentialité

Les fausses données sont élaborées de façon à ce qu'on puisse toujours extraire certaines informations. Prenons l'exemple d'un jeu de 100 dossiers médicaux, à partir duquel on souhaite savoir combien de personnes ont la grippe. On perturbe ce jeu de données en y introduisant 1 000 faux dossiers, chacun se voyant attribué de façon équiprobable une maladie parmi quatre (dont la grippe).

COMMENT RENDRE ANONYME UN JEU DE DONNÉES ?

Pour rendre le titulaire d'un jeu de données impossible à identifier, on se contente souvent de remplacer les champs tels que le nom ou le numéro de sécurité sociale par un pseudonyme. Cela n'empêche pas toujours les réidentifications ultérieures *via* des croisements de données. D'autres techniques ont alors été développées. Celle du *k*-anonymat est la plus employée aujourd'hui. Elle consiste à brouiller certains champs, dits quasi-identifiants parce que leur combinaison est parfois unique : l'âge, le sexe, le code postal… On remplace leurs valeurs précises par des intervalles de valeurs ou des catégories, de façon à ce que que la combinaison résultante soit partagée par au moins *k* personnes.

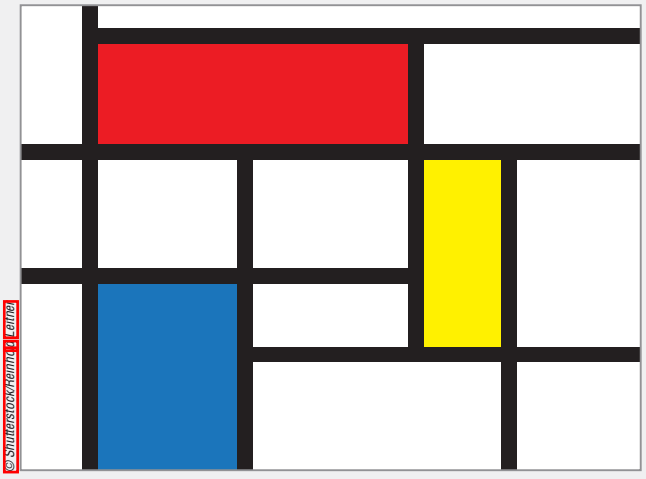
Pour appliquer cette technique, on utilise souvent l'algorithme de Mondrian, élaboré en 2006. Son nom est une référence à Piet Mondrian (1872-1944), peintre hollandais dont les œuvres partitionnent l'espace *(ci-contre, un dessin inspiré de Mondrian)* ; de même, l'algorithme partitionne l'espace des quasi-identifiants, où ces derniers sont indiqués sur des axes et où les individus sont repérés par des points.

L'algorithme commence par choisir un quasi-identifiant, selon lequel il divise les individus en deux groupes, par exemple ceux dont le code postal est inférieur ou égal à 75011, et les autres *(ci-contre en haut à droite)*. Si *k* est supérieur à 2, les groupes résultants sont de nouveau divisés en deux (en utilisant le code postal ou un autre quasi-identifiant, tel l'âge) pour former quatre groupes. Ce découpage continue jusqu'à ce qu'aucun groupe ne puisse plus être divisé sans englober moins de *k* individus. La dernière étape consiste à remplacer les quasi-identifiants de chacun par l'ensemble de valeurs couvert par son groupe.

Le problème est que les données personnelles à protéger, tel le diagnostic médical, peuvent être identiques au sein d'un groupe. Un attaquant qui sait que sa cible appartient à ce groupe a alors accès à sa valeur sensible (les valeurs sensibles étant les données à protéger). Plusieurs techniques visent à y remédier.

La *l*-diversité, par exemple, consiste à construire des groupes ayant au moins *l* valeurs sensibles. On choisit la valeur de *l* en fonction de ce que l'attaquant sait de sa cible, c'est-à-dire du nombre maximal de valeurs sensibles qu'on l'estime capable d'éliminer. Plus il sait de choses, plus *l* sera élevé, plus il faut inclure d'individus dans chaque groupe afin d'avoir une diversité suffisante, et moins les statistiques calculées à partir du jeu de données assaini sont fines. Le compromis entre utilité et protection est ici réalisé à travers le choix de *l*.

En pratique, on construit souvent les groupes grâce à l'algorithme de Mondrian. À chaque division d'un groupe en deux, on vérifie que les nouveaux groupes contiennent au moins *l* valeurs sensibles distinctes. Dans le cas contraire, on revient en arrière et on partitionne le groupe selon un autre quasi-identifiant *(en bas à droite)*.



© Shutterstock/Adonis ; Lainel

© Pour la Science - n° 433 - Novembre 2013

On sait alors qu'on a environ 250 valeurs fausses pour chacune de ces maladies. En comptant le nombre de grippés dans le jeu de données assaini et en lui soustrayant 250, on obtient le nombre réel de cas de grippe avec une bonne précision. Plus les vraies données sont nombreuses par rapport aux fausses, plus les informations extraites sont précises, mais moins les données sont protégées. Il existe de nombreuses façons de créer des fausses données, et les algorithmes qui s'en chargent font l'objet de multiples études.

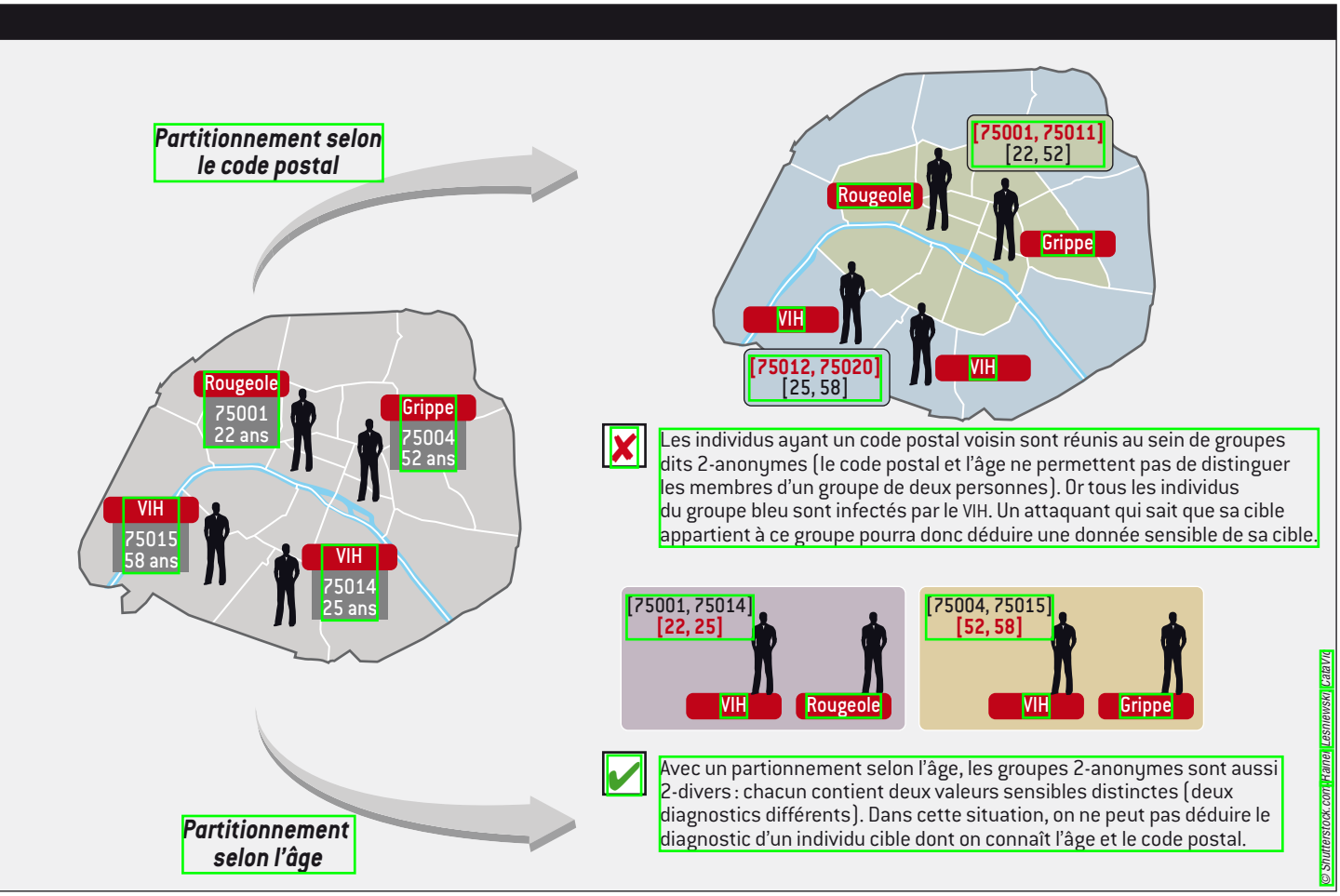
L'idée de la confidentialité différentielle est en plein essor. Elle est désormais applicable à des types variés de données, tels les graphes de réseaux sociaux *(voir la figure 3)*. Ces graphes représentent les individus par des nœuds, connectés par des liens. La perturbation peut alors consister à supprimer de vrais liens et à en introduire de faux. Des informations telles que le nombre de liens sont toujours extractibles du graphe, sans que l'on puisse déterminer les connexions précises d'un individu.

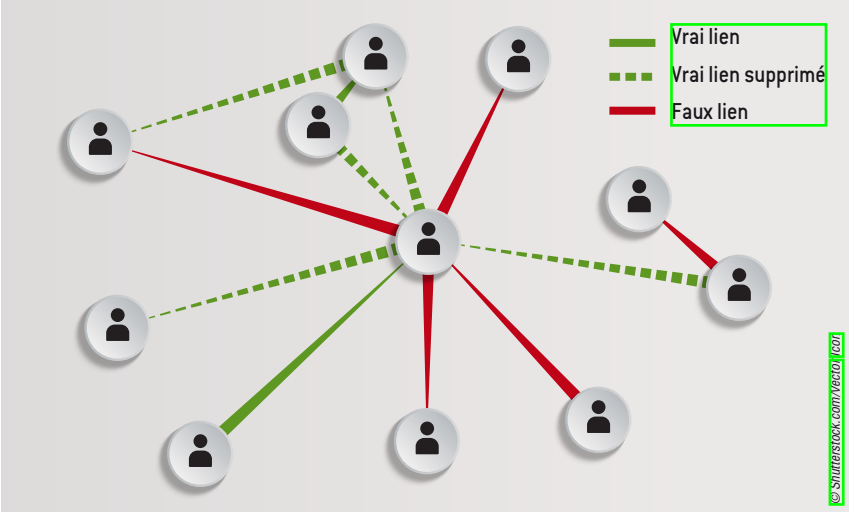
Le succès de la confidentialité différentielle s'explique en partie par sa simplicité : l'attaquant n'apparaissant pas dans les algorithmes, on évite les difficultés liées à l'estimation de ce qu'il sait de sa cible, et la distinction entre quasi-identifiants et données sensibles, parfois peu évidente, n'est pas nécessaire. Mais cette simplicité est à double tranchant : dans des travaux publiés en 2011, A. Machanavajjhala et ses collègues ont montré que la non-prise en compte des connaissances annexes de l'attaquant et des relations potentielles entre les individus d'un jeu de données ouvre des failles dans la protection.

UN MODÈLE D'ASSAINISSEMENT universel des données reste donc chimérique, chaque modèle étant plus ou moins efficace selon la situation.

Un modèle d'assainissement universel des données reste donc chimérique. Chaque modèle a ses défenseurs et ses détracteurs, et est plus ou moins efficace selon la situation.

Outre le modèle d'assainissement, l'architecture de gestion des données a une importance. Les données nominatives sont souvent extraites du système informatique assurant leur usage quotidien (tel le serveur d'un centre de soin), puis copiées sous une forme pseudonymisée dans un entrepôt de données. C'est à partir de cet entrepôt que seront produits à la demande des jeux de données assainis pour différents destinataires. En France et en Angleterre, le système de dossier médical personnel national fonctionne ainsi. Un tel processus facilite la production de multiples versions assainies d'un même jeu de données, chacune étant adaptée aux besoins du destinataire et limitée aux usages prévus par la législation. Ainsi,





3. LES GRAPHES DE RÉSEAU SOCIAL représentent les individus par des points et leurs liens par des traits. Pour qu'ils ne dévoilent pas la vie privée des individus, on peut leur appliquer un algorithme dit de confidentialité différentielle, consistant, par exemple, à supprimer de nombreux vrais liens et à les remplacer par autant de faux liens. Il est alors impossible de savoir avec précision à qui est connecté un individu, mais on peut toujours calculer des valeurs telles que la connectivité moyenne. On voit ici que les méthodes de perturbation des données doivent être choisies en fonction des informations que l'on souhaite pouvoir extraire.

les épidémiologistes peuvent recevoir des jeux de données plus précis que les industriels pharmaceutiques.

Toutefois, ce principe d'assainissement centralisé nécessite une fiabilité totale du gestionnaire de l'entrepôt de données. Or on constate régulièrement des fuites d'information sur de nombreux serveurs, dues à des négligences ou à des attaques internes ou externes (l'*Open Security Foundation*, organisation non gouvernementale américaine, recense les cas les plus significatifs sur son site Internet DataLossDB.org). Même si les données stockées dans les entrepôts sont pseudonymisées, nous avons vu que cela n'apporte pas une protection suffisante. Il est donc légitime de s'interroger sur les risques de la centralisation.

Les statisticiens ont proposé un mécanisme d'assainissement décentralisé dès les années 1960, pour parer aux réticences à leur confier des données personnelles sensibles. Le principe est de perturber les données de chacun au moment de leur collecte, ce qui les protége avant tout enregistrement. Illustrons ce principe dans le cas d'un sondage politique. On demande aux sondés de répondre par oui ou par non à une enquête d'opinion, en disant la vérité ou un mensonge avec une probabilité connue (par exemple, en suivant le résultat d'un tirage à pile ou face). Comme on connaît le pourcentage de réponses sincères, on peut faire des calculs statistiques sur ces données, mais on ne peut reconstituer l'opinion d'un individu précis.

Cependant, on sait aujourd'hui qu'une perturbation indépendante de chaque donnée ne permet pas d'atteindre le niveau de qualité d'un assainissement réalisé sur le jeu de données dans son ensemble : à protection équivalente, les données sont moins précises, et à précision égale, les données sont moins protégées. La centralisation des données personnelles est-elle alors nécessaire à un assainissement de qualité ? Non, car il est aussi possible d'élaborer des mécanismes décentralisés où les perturbations ne sont pas indépendantes – une piste étudiée par plusieurs laboratoires. En d'autres termes, la perturbation à appliquer n'est pas décidée localement (comme dans notre exemple précédent, où chaque sondé décide de la véracité de sa réponse en jouant à pile ou face), mais par une entité centrale. Celle-ci possède des informations sur toutes les réponses, sans connaître les réponses elles-mêmes. Dans le cas du sondage d'opinion, une telle entité saurait si un sondé a dit la vérité ou non, mais ne connaîtrait pas sa réponse. En pratique, l'entité centrale stocke tout de même les réponses, mais sous une forme chiffrée.

Aucune entité centrale ne doit tout savoir

Des algorithmes regroupés sous le nom de *Secure Multi-Party Computation* (littéralement, calcul multipartite sécurisé), qui font souvent intervenir des techniques de cryptographie, visent à assurer la confidentialité de l'assainissement distribué (où les calculs sont répartis entre plusieurs acteurs). Par exemple, si quatre personnes possèdent chacune un nombre qui doit rester privé, mais dont on souhaite que la somme soit calculable, on peut appliquer l'algorithme suivant : la première personne tire un chiffre aléatoire, auquel elle ajoute son nombre, avant de transmettre le résultat à la deuxième personne ; celle-ci y ajoute son nombre, transmet le résultat, et ainsi de suite jusqu'à ce que le résultat final revienne à la première personne ; cette dernière retranche le chiffre aléatoire qu'elle avait tiré et dispose alors de la somme des nombres, sans qu'aucun des participants n'ait découvert le nombre des autres. En effet, chacun a transmis sa donnée sous une forme perturbée, mais

BIBLIOGRAPHIE

T. Allard *et al.*, **MET$_A$P: Revisiting privacy-preserving data publishing using secure devices**, *Distributed and Parallel Databases*, pp. 1-54, 2013 (à paraître).

B.-C. Chen *et al.*, **Privacy-preserving data publishing**, *Found. and Trends in Databases*, vol. 2, n° 1-2, pp. 1-167, 2009.

A. Greenfield, **Everyware : La révolution de l'ubimédia**, Pearson, 2007.

C. Dwork, **Differential privacy**, *Proceedings of the 33rd international conference on Automata, Languages and Programming*, vol. 2, pp. 1-12, 2006.

L. Sweeney, ***k*-anonymity : a model for protecting privacy**, *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10(5), pp. 557-570, 2002.

© Pour la Science - n° 433 - Novembre 2013

avec un paramètre commun (le nombre aléatoire tiré au début).

Les mécanismes distribués constituent un pas majeur vers la sécurisation du processus d'assainissement, mais leur mise en œuvre pose des problèmes de passage à l'échelle, car ils nécessitent des calculs importants et une forte connectivité des participants. Ils sont pour l'instant relégués au traitement de jeux de données peu volumineux.

Avec notre équipe, nous nous sommes attaqués à ce problème. Nous avons développé une architecture de gestion de données personnelles fondée sur des dispositifs individuels sécurisés (telles des clefs USB dotées de processeurs sécurisés, insérables dans un smartphone, une tablette ou un ordinateur). Nous avons montré que les calculs nécessaires aux algorithmes d'assainissement peuvent être répartis entre ces dispositifs et une entité centrale sans que celle-ci ne dispose des données personnelles non chiffrées. La puissance de calcul de l'entité centrale assure l'applicabilité à grande échelle et, grâce au fait qu'elle coordonne les dispositifs personnels, ceux-ci n'ont pas besoin d'être interconnectés en permanence. Cette architecture est capable d'appliquer des algorithmes d'assainissement à des jeux de données massifs, correspondant à plusieurs millions d'individus.

Une telle architecture décentralisée pourrait assurer la gestion des dossiers médicaux, des factures ou de tout autre type de dossier personnel. En pratique, les données seraient stockées localement sur les appareils de l'utilisateur et ne seraient jamais exportées sous une forme non perturbée ou non chiffrée. Aucun serveur ne les regrouperait toutes, mais les échanges entre l'entité centrale et les appareils des utilisateurs permettraient tout de même de collecter certaines informations, d'une façon respectueuse de la vie privée.

Pendant la dernière décennie, une grande diversité de modèles et d'algorithmes d'assainissement de données ont été élaborés, afin de mieux prendre en compte les capacités de l'attaquant. Aujourd'hui, la pseudonymisation reste massivement utilisée, alors qu'elle ne garantit pas une protection suffisante ; dans le reste des cas, la *k*-anonymisation est le plus souvent appliquée, et les techniques plus complexes sont encore exotiques. Ces techniques doivent donc continuer à se répandre et à se perfectionner.

Pour autant, la science de l'assainissement reste la recherche du meilleur compromis entre utilité et confidentialité des données, dont la protection absolue est illusoire. Seule la destruction des données peut garantir qu'un attaquant n'en tirera aucune information sensible (cette destruction est d'ailleurs compliquée, en raison de leur présence sur de nombreux serveurs). L'analyse des nouveaux gisements de données personnelles pouvant apporter des bénéfices notables, l'assainissement de données est une question sociétale avant d'être un défi scientifique. Il faudra y apporter des réponses multidisciplinaires, impliquant informaticiens, juristes et analystes. ■



La protection des informations

Rémy Chrétien et Stéphanie Delaune

Un bon cryptage des données sensibles ne suffit pas à les protéger. Il faut aussi que les protocoles de communication utilisés pour coder ces données soient dépourvus de failles, dans leur logique comme dans leur mise en œuvre.

Le 6 septembre 2013, le monde apprenait avec stupéfaction l'existence de *Bullrun*. Ce programme de l'Agence de sécurité américaine, la NSA *(National Security Agency)*, vise à décoder un grand nombre d'informations cryptées circulant sur Internet. À partir des documents secrets que leur a livrés le lanceur d'alerte Edward Snowden, le *New York Times* et le *Guardian* révélaient qu'une percée technique avait multiplié la puissance de *Bullrun* en 2010, de sorte que les transactions et communications en ligne protégées des banques et de la plupart des grands fournisseurs de service sur Internet étaient désormais accessibles « en grands volumes » aux agences de renseignement *(voir la figure 1)* ; on apprenait aussi que ces agences disposent de superordinateurs pour décrypter par la « force brute » les données qui résistent…

© Pour la Science - n° 433 - Novembre 2013

sensibles



© Domaine public

Aujourd'hui, sans vraiment le savoir, nous réalisons tous quotidiennement des opérations de cryptage. Des chiffrements permettent, par exemple, de protéger notre numéro de carte bancaire lors d'un achat en ligne ; on les retrouve dans les protocoles (les procédures de communication) utilisés en téléphonie mobile ou pour lire les passeports électroniques. Pour autant, toutes ces données sensibles, nos données, sont-elles vraiment protégées ? Où sont les points faibles ? Nous allons analyser dans cet article ce qui conditionne la qualité des systèmes de sécurisation des données utilisés de façon routinière.

Sur un réseau, tout se passe comme si les informations protégées ne circulaient qu'à l'intérieur de coffres-forts réalisés dans un matériau solide : le chiffre (le cryptage). Toutefois, quelle que soit la solidité mathématique du chiffrement opéré, encore faut-il que l'assemblage des pièces qui constituent les coffres-forts en circulation ne laisse pas apparaître de fente exploitable par une personne mal intentionnée… Finalement, l'expédition de coffres-forts d'un côté et de leurs clefs de l'autre n'est pas une pratique sans risques, et il arrive qu'on y commette des erreurs, voire qu'on laisse traîner les clefs…

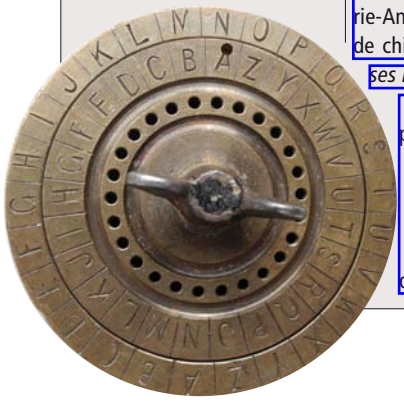
Chiffre et compagnie

Le chiffrement, nous l'avons dit, est l'analogue du matériau constituant les parois du coffre-fort. Il en existe de nombreux types, chacun caractérisé par des contraintes d'utilisation et des niveaux de sécurité différents. Quoi qu'il en soit, la qualité d'une primitive cryptographique, c'est-à-dire d'un algorithme de cryptage particulier, se mesure à sa robustesse et à sa facilité de mise en œuvre. Tout comme les parois d'un coffre-fort peuvent être plus ou moins épaisses, la robustesse d'une primitive cryptographique peut être plus ou moins importante. Une primitive sera considérée comme faible quand il est facile de découvrir comment inverser son chiffrement, inversion qui permet de révéler les données protégées *(voir l'encadré page suivante)*.

1. CES DÔMES RECOUVRENT LES ANTENNES utilisées (ici au Japon) par le système mondial d'interception des communications privées et publiques des agences de renseignement des États-Unis, du Royaume-Uni, du Canada, de l'Australie et de la Nouvelle-Zélande.

Chiffrement et clef de chiffrement

Chiffrer consiste à mélanger de façon plus ou moins inextricable les informations élémentaires d'un texte (lettres, chiffres, bits…). Le chiffre employé par Jules César dans ses correspondances secrètes fournit un exemple simpliste de chiffrement : César se contentait de décaler l'alphabet latin de trois lettres. Son propre nom, CESAR, devenait ainsi FHVDU. Le chiffre de César peut être étendu aux 26 décalages alphabétiques possibles, ce qui explique que des appareils à chiffrer ainsi aient été vendus aux commerçants et aux militaires jusqu'au XIXe siècle *(ci-dessous)*. Il peut aussi être généralisé aux chiffrements polyalphabétiques, qui consistent à substituer les lettres de l'alphabet par d'autres tirées du même alphabet, voire d'un autre. Pendant sa captivité, la reine Marie-Antoinette employa ce type de chiffre *(ci-dessus, l'une de ses lettres)*.



Tout chiffrement implique un algorithme de déchiffrement. Dans le cas du chiffre de César, il suffit de décaler l'alphabet de trois lettres dans le sens opposé pour déchiffrer.

Dans la pratique cryptologique, l'algorithme de (dé)chiffrement est public, de sorte que la sécurité réside dans la clef qui spécifie le chiffrement effectué. Cette clef est le nombre 3 dans le cas du chiffre de César, ce qui est bien trop simple pour résister à des essais de déchiffrement systématiques (la force brute). Pour produire une clef plus difficile à découvrir, on pourra par exemple décaler de 3 lettres la première phrase, de 1 lettre la deuxième, de 4 la troisième, puis de 1, de 5, de 9, de 2, de 6, de 5,… et nommer π la clef résultante. Mais cela serait encore bien trop simple !

Jacques Faizant et Valérie Nacha

Après la robustesse, le deuxième critère à prendre en compte dans le choix d'une primitive cryptographique est la facilité avec laquelle on pourra la mettre en œuvre dans la situation à traiter. Si placer dans un appartement ou une maison particulière un coffre-fort aux parois d'acier de dix centimètres d'épaisseur et pesant cinq tonnes semble une précaution exagérée, ce n'est pas le cas dans une banque. De même, les contraintes qui accompagnent l'emploi d'une primitive cryptographique restreignent ses applications.

Les systèmes de sécurisation des données embarqués dans un téléphone portable, par exemple, sont limités en mémoire et en puissance de calcul, ce qui y interdit les primitives trop complexes ; la situation est différente si les données à protéger ont vocation à être stockées sur un serveur.

Une fois la primitive choisie, il reste à fixer la taille de la clef de chiffrement à utiliser. Une telle clef consiste en une série de bits (de valeurs 0 ou 1) déterminant la façon dont s'opère le cryptage avec un algorithme donné. Dans le cas des algorithmes dits symétriques, la clef spécifie aussi le déchiffrement. Dans celui des algorithmes dits asymétriques, elle ne sert qu'au chiffrement, et une clef supplémentaire est nécessaire au déchiffrement. Quoi qu'il en soit, un principe général prévaut : plus une clef est longue, plus il sera difficile de casser le chiffrement correspondant et le cryptosystème dont elle est le cœur.

L'ESSENTIEL

- Les données stockées ou circulant en masse sur Internet sont protégées par des systèmes de sécurité.
- Les attaques peuvent porter soit sur leur couche cryptographique, soit sur leur couche logique.
- La solidité de la première dépend de la longueur de la clef de chiffrement ; celle de la seconde, des détails de la mise en œuvre du système.
- Les attaquants tentent aussi d'exploiter toute fuite involontaire d'information que présente le système.

Hubert Berberat

Attaques par force brute

Ce principe général s'explique facilement si l'on considère le cas d'un algorithme symétrique public – c'est-à-dire connu et employé par qui veut. Supposons ce mode de cryptage contrôlé par une clef comportant N bits. Pour le casser, il existe toujours, en dernier ressort, la méthode dite de la force brute : elle consiste tout simplement à essayer, avec l'aide d'un ordinateur, toutes les clefs possibles, soit 2^N clefs. Dans le cas d'une clef de 320 bits, on aura donc $2^{320} \simeq 2 \times 10^{96}$ clefs (plus que d'atomes dans l'Univers…) à tester. Face à de tels nombres, la méthode de la force brute paraît bien illusoire. Mais la prudence s'impose : avec l'augmentation incessante de la puissance de calcul des superordinateurs massivement parallèles (à très nombreux processeurs calculant en même temps), l'expérience a montré que l'on a pu casser des primitives d'usage courant, dont les clefs étaient considérées jusque-là comme assez longues pour être sûres…

L'algorithme de cryptage AES *(Advanced Encryption Standard)*, par exemple, s'emploie aujourd'hui avec des clefs de 128, 192 ou 256 bits. Cet algorithme fait subir aux blocs successifs de données un triturage numérique complexe comprenant des permutations, des rotations dans un espace mathématique abstrait, etc., le tout répété plusieurs fois selon la longueur de la clef. En 2011, Christian Rechberger, de l'Université technique de Copenhague, et deux collègues ont publié une méthode d'attaque d'AES-128 (avec clef de 128 bits) environ quatre fois plus rapide que la force brute. Son déroulement implique à peu près 2^{126} (environ 10^{38}) opérations pour trouver la clef. Un ordinateur effectuant dix milliards d'opérations par seconde mettrait de l'ordre de 3×10^{20} années pour faire le calcul, soit plusieurs milliards de fois l'âge de l'Univers : on comprend qu'AES-128 soit encore considéré comme sûr…

Ce cas illustre que plus la clef d'un algorithme est longue, plus il faut de temps

© Pour la Science - n° 433 - Novembre 2013

et de ressources de calcul pour le décrypter. La longueur de la clef détermine donc le niveau de sécurité du cryptage.

L'histoire des attaques contre l'algorithme RSA montre l'influence de la longueur de la clef sur la sécurité. Nommé d'après ses inventeurs Ron Rivest, Adi Shamir et Leonard Adleman, RSA est un algorithme asymétrique à clef publique (la clef de cryptage est connue de tous) dont la robustesse découle de la difficulté à décomposer les grands nombres en facteurs premiers. Le cryptage RSA a été utilisé dès les années 1980 avec des clefs de 320 bits par le groupement des banques émettrices de cartes bancaires. Or une version de RSA à clefs de 320 bits a été cassée à la fin des années 1990 par Serge Humpich. Cet ingénieur français a exploité une faille logique du système de sécurité de ce moyen de paiement tel que spécifié en 1983 (la puce contenue dans les cartes vérifie elle-même le code), puis cassé l'instance RSA utilisée dans les cartes bancaires, obtenant à partir de là des cartes fonctionnelles, mais non reliées à un compte bancaire…

Avant 1983, les cryptologues consultés par le groupement des banques émettrices de cartes avaient préconisé une clef d'au moins 640 bits, mais les banques ne les avaient pas suivis. Aujourd'hui, la taille minimale acceptable pour des clefs RSA est de 1 024 bits, tandis que la taille recommandée est de 2 048 bits…

Ce cas et d'autres illustrent que, dans la pratique, il faut non seulement que les clefs des algorithmes d'usage courant soient assez longues, mais aussi qu'on augmente au fil du temps leurs longueurs à mesure que les moyens de calcul deviennent plus puissants *(voir la figure 2)*.

La clef doit être secrète, pas l'algorithme

Comment mesurer la sécurité inhérente à un algorithme de chiffrement à un moment donné ? Pour ce faire, le mieux est de le soumettre librement aux attaques. L'expérience a en effet montré qu'il est préférable de rendre publics les algorithmes de cryptage (le fonctionnement du chiffre de César il y a 2 000 ans, ou celui de RSA aujourd'hui), car ces derniers finissent toujours par être connus des attaquants – lesquels ne peuvent parvenir à leurs fins que s'ils réussissent à obtenir la clef par un vol ou par un calcul. C'est pourquoi le cryptologue militaire néerlandais Auguste Kerckhoffs a énoncé en 1883 le principe qui porte son nom : la sécurité d'un système de cryptage ne doit reposer que sur le secret de sa clef. Ainsi, la confiance que les cryptologues accordent à une primitive découle soit de sa résilience aux méthodes de cryptanalyse développées pour en découvrir la clef raisonnablement vite, soit de la difficulté de résoudre les problèmes mathématiques qui lui sont associés ; cette confiance n'a donc rien d'absolu et change avec le temps (en général en diminuant...).

À ce propos, les algorithmes symétriques A5/1 et A5/2 de chiffrement à la volée, utilisés dans les communications hertziennes de téléphonie mobile selon les normes GSM et GPRS, fournissent un intéressant cas d'école. Les spécifications de ces algorithmes fondés sur un générateur de nombres pseudo-aléatoires ont été initialement maintenues secrètes. Mais David Wagner, de l'Université de Californie, et des collègues sont parvenus à les révéler en 1999. L'écoute en temps réel des communications sur téléphones mobiles est rapidement devenue possible, alors que les systèmes A5/1 et A5/2 avaient déjà été déployés sur la planète entière…

AES (chiffrement symétrique)		
Année	Longueur minimale (en bits)	
	Lenstra	ANSSI
2014	78	100
2015	79	100
2017	80	100
2020	82	100
2025	85	128

RSA (chiffrement asymétrique)		
Année	Longueur minimale (en bits)	
	Lenstra	ANSSI
2014	1309	2048
2015	1350	2048
2017	1435	2048
2020	1568	2048
2025	1805	4096

2. CE TABLEAU donne les évolutions recommandées de la longueur des clefs des chiffrements AES et RSA, telles que les ont estimées le cryptologue Arjen Lenstra, de l'École polytechnique fédérale de Lausanne (Suisse), et l'Agence nationale de sécurité des systèmes d'information (ANSSI, France).

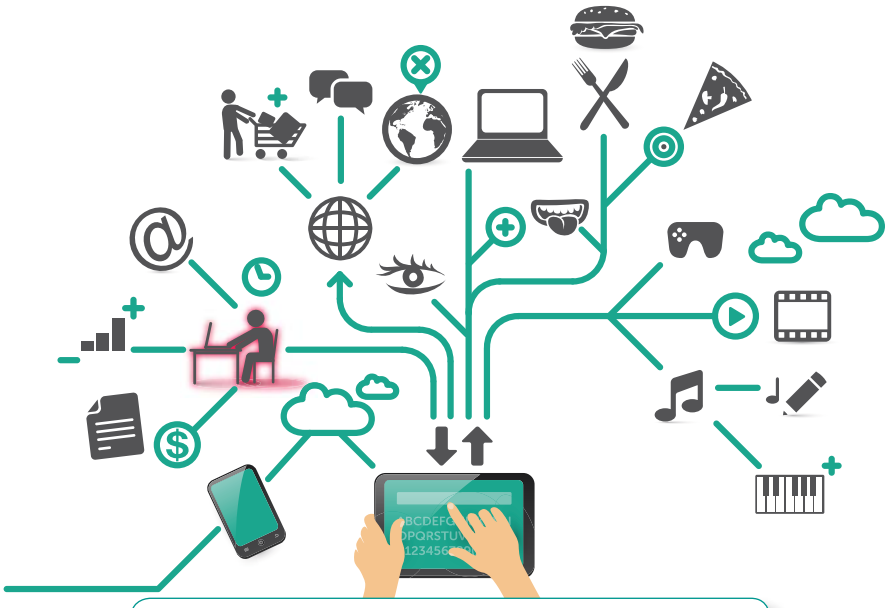


© Fedor Selivanov/shutterstock.com

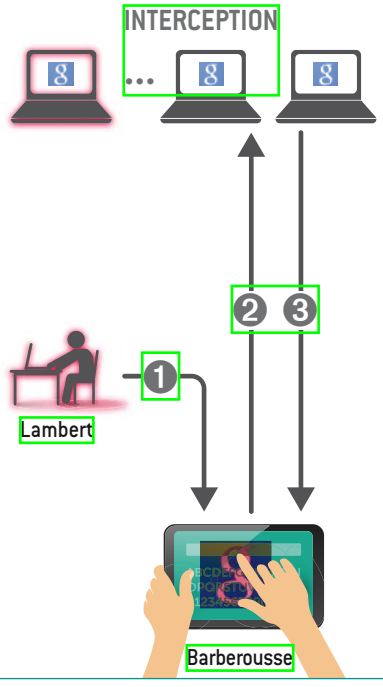
3. LES CRYPTAGES DES COMMUNICATIONS DE TÉLÉPHONIE MOBILE GSM ET GPRS déployés à l'origine *(ici une antenne-relais)* ont été cassés dès 1997. Puis, en 2000, des chercheurs ont prouvé qu'il était même possible de le faire à la volée, de sorte qu'après interception, l'écoute en temps réel des conversations GSM et GPRS sur la planète entière est devenue possible.

UN EXEMPLE DE VOL DE DONNÉES PERSONNELLES SUR INTERNET

L'usurpation d'identité sur Internet prend de nombreuses formes. Nous avons tous reçu des courriels dans lesquels des escrocs nous demandent des données sur notre identité ou nos moyens de paiement en se faisant passer pour l'administrateur de l'un des sites que nous fréquentons (banque, librairie en ligne, etc.). La ficelle grossière de cette escroquerie a son équivalent en temps réel, qui est décrit ci-dessous.



Transactions bancaires, achats de morceaux de musique, etc. : un individu a aujourd'hui de nombreuses activités en ligne et doit pour cela s'identifier souvent. Cette réalité rend la confidentialité sur Internet cruciale pour le fonctionnement de la société ; elle explique aussi que les escrocs font de nombreuses tentatives pour usurper notre identité ou voler nos mots de passe en temps réel.



Levé de bonne humeur, Lambert se connecte ❶ grâce à son compte *Google* à une nouvelle application affichant en ligne les résultats de l'équipe de France de basket. Mais il ignore que ce service a été créé par Barberousse, un escroc qui lui fait faire bien plus que seulement afficher des résultats sportifs... Barberousse utilise l'identité de Lambert pour demander ❷ une connexion à l'une des applications du compte *Google* de Lam-

LES AUTEURS




Stéphanie DELAUNE, chargée de recherche au CNRS, travaille au LSV (Laboratoire de spécification et vérification), le laboratoire d'informatique de l'École normale supérieure de Cachan.

Rémy CHRÉTIEN est doctorant au LSV et au Laboratoire lorrain de recherche en informatique et ses applications (LORIA).

De telles erreurs de conception montrent que, pour éviter ce genre de désastre, d'importants efforts de recherche sont nécessaires avant de déployer un système très utilisé. Il est fréquent que le même algorithme de chiffrement soit employé dans une myriade d'applications différentes, de sorte qu'elles deviennent toutes vulnérables si l'algorithme est attaqué avec succès. Et le nombre énorme de systèmes impliqués ainsi que la complexité inhérente aux primitives cryptographiques rendent impraticables, en général, les corrections à la volée (par exemple par un patch, c'est-à-dire une section de code correctrice).

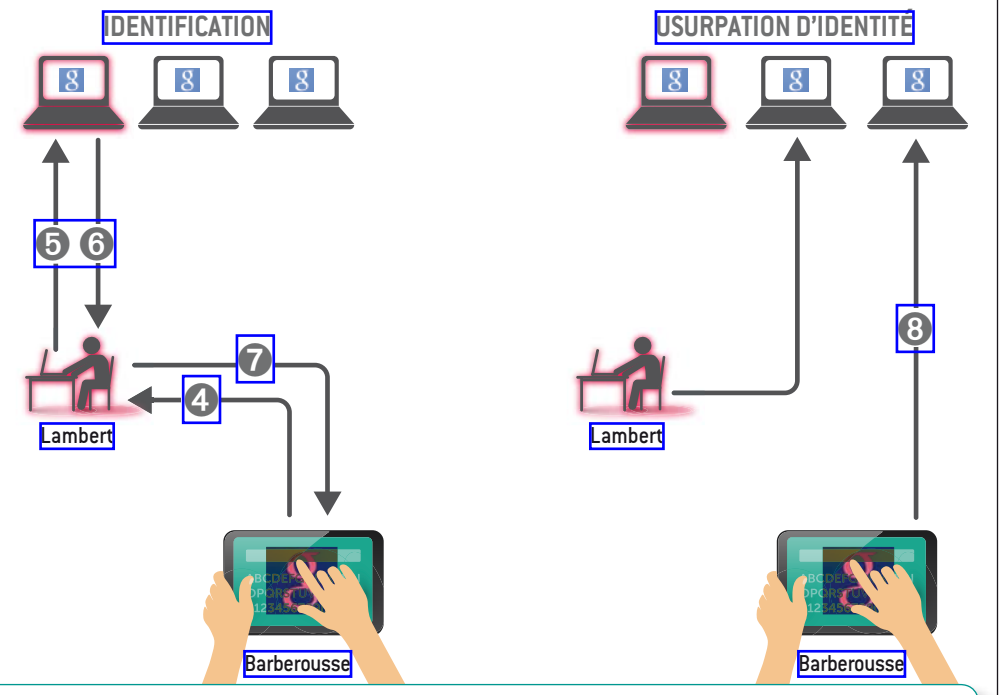
Une fois le coffre-fort choisi, il importe de maîtriser son expédition, ce qui n'a rien de facile. En effet, une fois effectué le choix des primitives cryptographiques adaptées à la situation, il reste à les mettre en œuvre sans commettre d'erreurs ou de maladresses fatales pour la sécurité des données. La construction d'un cryptosystème s'accompagne en effet de spécifications fixant la façon dont ses composantes échangent des informations, qui sont le plus souvent chiffrées, mais pas toujours. Comme une erreur dans le choix du coffre-fort peut être fatale, un mauvais choix de système de transport (de protocole de communication) annule vite la sécurité apportée par le blindage (par le chiffrement).

Des failles dans la couche logique...

Lorsqu'on traite les problèmes d'organisation et de mise en œuvre d'un système de sécurisation des données, on quitte la couche cryptographique pour entrer dans ce que l'on nomme la couche logique de sécurité. L'emploi d'un algorithme de cryptage dépourvu des faiblesses mentionnées plus haut ne garantit pas l'absence

© Pour la Science - n° 433 - Novembre 2013

ET D'ACCÈS AUX APPLICATIONS *GOOGLE*



bert. Celle-ci répond par une demande d'identification ❸, que Barberousse répercute à Lambert ❹. Ce dernier s'identifie ❺ auprès du serveur dédié à l'authentification des utilisateurs de *Google*. Ce serveur renvoie ❻ la preuve de l'identité de Lambert à son ordinateur, lequel, conditionné par l'application malveillante de Barberousse, lui communique ladite preuve ❼. Barberousse se sert de cette preuve pour se connecter ❽ par le compte de Lambert à une application *Google*, par exemple l'agenda ou la messagerie de Lambert ; dès lors qu'il s'est connecté avec l'identifiant et le mot de passe de Lambert, ce dernier est aussi simultanément connecté sur l'application sportive, qu'il consulte alors, toujours de bonne humeur, pendant que Barberousse lit son emploi du temps, vole ses secrets ou envoie des messages depuis son ordinateur...

© zaomaistudio/Shutterstock.com

d'erreurs dans la logique présidant à la conception du système.

La première erreur logique possible consiste à envoyer la clef sans la cacher, en d'autres termes à transmettre la clef en clair, ou d'une façon qui permet de la voler. Le vol d'identité résultant d'informations de connexion d'un utilisateur à un service en ligne (une banque en ligne, par exemple) non chiffrées sur Internet fournit un exemple trivial d'une telle erreur logique. Dans ce cas simple, un attaquant n'a qu'à intercepter la communication entre l'usager et son fournisseur de service pour dérober les informations confidentielles permettant de se connecter à la place de l'usager. De telles failles se sont révélées récemment sur des services bien connus du grand public. On peut citer une version ancienne du protocole de connexion des particuliers aux services de *Google* (dit *Google Single Sign-on*), qui a la particularité d'étendre à toutes les applications de *Google* la connexion d'un utilisateur à une seule application.

Ce genre d'attaque permet, par exemple, d'accéder à l'agenda d'une personne après qu'elle a entré son nom d'utilisateur et son mot de passe. Elle repose sur la ruse consistant à pousser l'utilisateur à se connecter à une application malveillante conçue pour l'intéresser, mais secrètement contrôlée par l'attaquant. Il pourra s'agir, par exemple, d'une application affichant les derniers résultats sportifs. L'attaquant intercepte le message envoyé par l'utilisateur pour accéder à cette application. Cette interception est intéressante pour l'attaquant, car ce message a le défaut de ne pas être spécifique à une application donnée, de sorte que l'attaquant peut réutiliser les informations de connexion pour se connecter à d'autres applications depuis le compte *Google* de l'utilisateur, par exemple à sa messagerie, à l'agenda, au service cartographique, à ses archives personnelles, etc. *(voir l'encadré ci-contre)*.

Ce cas illustre qu'en matière de sécurité informatique aussi, le diable est dans les détails, dans des détails de conception infimes, parfois presque invisibles, mais qui peuvent avoir d'importantes répercussions sur la sécurité. Ces failles dans la conception d'un système de protection des données sont d'autant plus graves que la faiblesse qu'elles engendrent affecte des services déjà déployés, et que le déploiement d'un patch ne suffit pas à les réparer. Dans le cas du protocole de connexion aux services de *Google*, changer un message de connexion utilisé par toutes les applications n'est guère réaliste, car cela a des effets à la fois sur la pratique de l'utilisateur et sur l'accessibilité des autres applications. Il est en effet souvent important de maintenir la compatibilité entre les divers protocoles donnant accès à un faisceau d'applications, même si l'un d'eux est défectueux. Une révision complète du système de connexion des services de *Google* a donc été nécessaire pour supprimer la faille.

Que faire pour éviter ce genre de difficulté ? Pour l'essentiel, il s'agit de tester de façon complète, avant de déployer le système de protection, la sûreté des procédures utilisées pour faire circuler les coffres-forts (les informations cryptées) qui protègent les données. Différents modèles théoriques de mise en œuvre existent. Dans le modèle dit symbolique, on procède en supposant d'abord que les primitives cryptographiques choisies sont incassables, puis en représentant par des symboles toutes les opérations effectuées par le système de sécurité. Sur la base d'une définition rigoureuse de la sécurité à obtenir, une technique mathématique rigoureuse permet ensuite de s'assurer que le protocole résiste à toutes les formes d'attaques décrites au sein du modèle utilisé. Une fois cette étape franchie, on peut passer à la mise en œuvre du système de sécurité.

... ou dans l'implémentation

Nous avons maintenant soigneusement choisi des coffres-forts adéquats (des algorithmes de cryptage), organisé scrupuleusement leur transport en nous méfiant de toutes les trahisons volontaires ou involontaires du transporteur (des procédures de communication) et des ruses imaginables

Le passeport biométrique attaqué

Les passeports biométriques français contiennent des puces dites RFID capables de communiquer sans fils. Les données personnelles du voyageur qu'elles contiennent ne deviennent accessibles qu'après que le lecteur de passeport a prouvé son droit de les recevoir en lisant, puis en transmettant à la puce, la clef chiffrée encodée en bas de la première page du passeport. Une fois cette formalité accomplie, le lecteur et le passeport créent une nouvelle clef temporaire pour chiffrer le numéro de passeport, le nom, la date de naissance, la nationalité et la photographie du propriétaire du document.

Toutefois, ce système de sécurité s'est révélé vulnérable. En 2005, il est apparu qu'il était facile de « casser » la clef de sécurité en employant la méthode dite de la force brute. Des failles logiques ont aussi été découvertes dans la conception du système. La première permettait de reconnaître le voyageur en exploitant un message d'erreur envoyé au cours de l'authentification.

En 2010, elle a été corrigée, mais on s'est alors aperçu que la durée exacte du temps de réponse livrait aussi l'identité du voyageur… Plus récemment encore, nous avons montré que la taille et le format des données chiffrées suffisent pour différencier de nombreux passeports, et donc pour les reconnaître dans un contexte particulier. Toutes ces attaques exigent toutefois de savantes mises en œuvre : le contenu des puces RFID de nos passeports est bien mieux protégé que celui de la plupart des puces RFID utilisées au quotidien…

Véronique Cortier,
CNRS-LORIA, Nancy



© cyrri/fotolia/shutterstock.com

BIBLIOGRAPHIE

T. Chothia et V. Smirnov, **A traceability attack against e-passports**, dans *Financial Cryptography*, Springer, vol. 6052, pp. 20-34, 2010.

A. Armando *et al.*, **Formal analysis of SAML 2.0 web browser single sign on : Breaking the SAML-based single sign-on for Google apps.**, 6th ACM Workshop on Formal Methods in Security Engineering (FMSE 2008).

A. K. Lenstra, **Key Lengths, Contribution to the Handbook of Information Security**, Université d'Eindhoven, 2004.

SUR LE WEB

Page Internet du LORIA sur le passeport électronique : www.loria.fr/~glondu/epassport/attaque-tailles.html

Cours et séminaires du Collège de France sur la sécurité informatique : www.college-de-france.fr/site/martin-abadi/

par les bandits (par les pirates informatiques) ; il nous reste à passer à la pratique (à implémenter le système). Dans cette étape aussi, des failles de sécurité peuvent passer inaperçues. Les spécifications des protocoles n'étant jamais assez exhaustives, il est fréquent que les interprétations qu'en font ceux qui les implémentent confèrent au système certaines particularités, des incompatibilités avec d'autres systèmes, voire des failles de sécurité…

En 2008, on a ainsi découvert une faille dans la version de la suite logicielle *Open*SSH utilisée sur les serveurs *Linux* tournant avec le système d'exploitation *Debian*. Cette suite a été développée par l'équipe du projet *Open*BSD (un projet d'environnement UNIX libre) pour faciliter aux programmeurs l'établissement de connexions sécurisées à partir du protocole SSH (*Secure Shell* en anglais, ou ligne de commande sécurisée). Or il s'est avéré que les paires de clefs RSA de 2048 bits engendrées à l'aide de la version *Debian* de la bibliothèque d'outils de chiffrement *Open*SSL (utilisée par la suite logicielle *Open*SSH) entre 2006 et 2008 avaient en réalité une robustesse équivalente à celle d'une clef de… 16 bits. De fait, 20 minutes suffisaient pour casser les cryptages SSH pour serveurs *Debian*, et prendre ainsi le contrôle de la machine… Compromises, toutes ces clefs ont dû être changées.

Un autre exemple est celui du passeport électronique français introduit en 2006. Ce passeport incorpore une puce de radio-identification (ou RFID, pour *radio frequency identification*) qui, dans la version biométrique en circulation depuis 2008, contient les photos du détenteur et de deux de ses empreintes digitales. Si les spécifications de l'un des protocoles de communication entre la puce RFID du passeport et la borne de lecture étaient initialement correctes, l'implémentation choisie pour la version française du passeport introduisait une faille. Ses concepteurs avaient prévu qu'en cas d'incident d'authentication du passeport auprès d'une borne, un message d'erreur serait émis pour en préciser la cause.

L'intention était bonne, mais cette circulation supplémentaire d'information s'est révélée suffisante pour qu'un attaquant interceptant les communications puisse identifier le passeport français en voie d'authentification, et ainsi le suivre à la trace à travers le monde *(voir l'encadré ci-contre)*.

Ce type d'attaque ne remet pas en cause la spécification du système et ne concerne que certaines implémentations. Il est donc généralement plus aisé de les bloquer. Bien entendu, le mieux est toujours de les éviter en amont de l'implémentation. Il s'agit, là encore, de tester le système de sécurité en le simulant par un modèle théorique le plus réaliste possible. Malheureusement, tout modèle reste… un modèle, c'est-à-dire un jeu permettant de découvrir des attaques possibles dans un cadre restreint, parce que théorique…

Les failles non encore rendues publiques se monnayent

Notons que les failles dont l'existence n'a pas encore été rendue publique se vendent à prix d'or. Il existe en effet un marché de ce que l'on nomme des vulnérabilités « jour zéro » (le jour 1 étant celui de la publication). Ces attaques restées secrètes sont vendues au plus offrant, le plus souvent à des entreprises ou à des États concernés.

Quand la sécurisation des données repose sur une primitive inviolable, qu'elle a été conçue sans aucune faute logique et implé-

© Pour la Science - n° 433 - Novembre 2013

mentée sans erreurs, devient-elle parfaitement sûre ? Pas encore. Il reste les attaques indirectes, qui sont comparables à la découverte de la combinaison numérique protégeant un coffre-fort grâce à l'écoute au stéthoscope des bruits de son mécanisme de verrouillage… C'est ce que l'on nomme les attaques par canaux auxiliaires ou canaux cachés.

Ce type d'attaque n'est malheureusement pas l'apanage des films d'espionnage : même un cryptosystème parfaitement réalisé peut être sujet à des fuites d'informations parasites qui, bien analysées et exploitées, suffisent parfois pour le casser. Il arrive ainsi qu'un processeur émette des rayonnemen ts exploitables, qu'il consomme de l'énergie suivant un profil traduisant les tâches de transmissions qu'il est en train d'accomplir, qu'un algorithme de chiffrement ait des temps d'exécution plus ou moins longs selon la valeur de la clef, etc.

De telles fuites ont déjà été utilisées pour casser certaines implémentations du cryptage RSA. Comment ? Les opérations d'élévation à la puissance nécessaires dans cet algorithme peuvent être réalisées soit par multiplication, soit par élévation au carré ; or ces deux opérations induisent des profils de courant électrique différents, lesquels, mesurés et rapportés à leur cause, ont *in fine* permis de déterminer la clef !

Attention aux fuites !

Les failles de ce type sont les plus difficiles à déceler. Les modèles de cryptosystèmes, qui ignorent toute la couche technique sous-jacente au fonctionnement des ordinateurs et des réseaux, sont inutiles pour les repérer. Une fois qu'elles sont connues cependant, il arrive que des contre-mesures simples suffisent à les faire disparaître, par exemple un brouillage supplémentaire de données, une égalisation des signaux ou de la consommation électrique, le blindage des puces pour les rendre opaques à l'espionnage électromagnétique, etc. Autant de défenses qui tiendront jusqu'à ce qu'un autre attaquant découvre encore une fuite insoupçonnée d'information…

La conception et la mise en œuvre d'un cryptosystème impliquent ainsi une science subtile. Cependant, toute la subtilité du monde ne peut rien contre l'utilisateur d'un coffre-fort qui laisse traîner les clefs ou contre la maladresse d'un utilisateur de communications chiffrées qui emploie le même mot de passe trivial (son prénom) pour se connecter au site de son groupe de fans de football, puis à sa banque et ensuite à sa messagerie d'entreprise…

De plus, les révélations du *Guardian* et du *New-York Times* suggèrent que le secret le mieux gardé de l'Internet aujourd'hui serait que les agences de renseignement anglo-saxonnes ont secrètement imposé aux « industries de l'Internet américaines et étrangères » d'insérer des « vulnérabilités dans leurs systèmes de cryptage ». Ces portes dérobées leur assureraient des entrées dans les protocoles de communication sécurisée largement utilisés en ligne, tels HTTPS (le protocole de transfert hypertexte sécurisé), la voix sur IP (téléphonie par Internet) ou encore SSL. Un coffre-fort doté d'une trappe indécelable, mais connue des espions, protégera bien mal nos secrets, c'est clair... ■



La quête d'un support numérique durable

Les disques durs, les CD, les DVD, les mémoires flash... se dégradent en quelques années. Cette situation est largement sous-estimée, mais diverses stratégies et de nouveaux dispositifs sont mis en place.



Franck Laloë et Erich Spitz

Les greniers des vieilles maisons abritent souvent des trésors. Quand on fouille dans des tiroirs ou dans des coffres, il n'est pas rare de retrouver de vieilles photos, des films huit millimètres ou des lettres manuscrites. Aujourd'hui, la plupart des documents, aussi bien les images que les vidéos et les textes, sont enregistrés sur des supports numériques, tels les disques durs des ordinateurs, les CD et les DVD.

À quoi ressembleront les chasses au trésor dans les greniers d'ici 50 ans ? Comment les supports numériques résistent-ils au passage du temps ? Les constructeurs ont annoncé que les supports numériques avaient une longue durée de vie. Les rares études réalisées montrent, au contraire, que ces supports se détériorent en quelques années. Dès lors, se pose la question de la pérennité des données numériques. Comment les conserver ? Ce problème est méconnu du grand public. Pourtant, il touche les documents personnels ayant une valeur sentimentale (les photos et les lettres), mais aussi les informations plus sensibles, telles les données administratives, financières, médicales ou scientifiques.

Aujourd'hui, le numérique est le véhicule dominant de l'information. Il a supplanté les supports analogiques parce qu'il présente de nombreux avantages : une grande densité d'information permettant un gain de place considérable, la facilité des recherches dans l'information stockée, un accès rapide, une circulation quasi instantanée et son universalité – la même clef USB peut contenir des textes, du son, des vidéos, des images et des tableaux de données. La question de la pérennité des données a été reléguée au second plan. Pour l'industrie, la priorité était de développer des dispositifs rapides et contenant des quantités d'informations toujours plus importantes.

La question de la pérennité reste essentielle. Nous aborderons les différents aspects

© Pour la Science - n° 433 - Novembre 2013

L'ESSENTIEL

- **La pérennité des données numériques est un problème peu étudié par l'industrie et les organismes publics.**
- **L'altération des supports est la difficulté principale pour l'archivage. Les dispositifs les plus courants ont des durées de vie courtes.**
- **La méthode d'archivage durable la moins onéreuse aujourd'hui consiste à graver des DVD en verre.**



© T. MACKAY PHOTOGRAPHY LLC/shutterstock.com

de la pérennité de l'information numérique et soulignerons que le problème le plus délicat est celui de l'altération des dispositifs de stockage. Des solutions existent, mais elles sont encore peu répandues.

Notons d'abord que sauvegarde et archivage à long terme des données ne sont pas synonymes. La première a pour but de prévenir les pertes soudaines de données (accidents divers, vols, etc.), la seconde, la disparition progressive des données utiles à long terme. En ce qui concerne la sauvegarde, comment garantir que la duplication d'un fichier est fidèle à l'original ? Cette difficulté trouve une solution efficace dans les codes correcteurs, des algorithmes qui identifient les erreurs de copie.

Dupliquer sans erreurs

Cet aspect des méthodes numériques passe souvent inaperçu des utilisateurs : la possibilité de copier un document numérique avec la garantie que la copie ne contient pas d'erreurs. En effet, quand vous enregistrez un texte, aucune altération accidentelle ne viendra s'y glisser et changer une lettre en une autre pendant la sauvegarde. Cette fiabilité ne se retrouve pas dans les supports analogiques. Il suffit de penser aux textes anciens qui nous sont parvenus par le travail séculaire des moines copistes qui, malgré leur soin extrême, ont introduit des erreurs ; il est humainement impossible de faire des copies identiques. C'est d'ailleurs valable pour toute technique analogique : quelle que soit sa qualité, chaque recopie introduit un peu de « bruit » qui s'accumule de façon exponentielle de copie en copie. Au bout d'un certain nombre de sauvegardes, l'information initiale devient inexploitable.

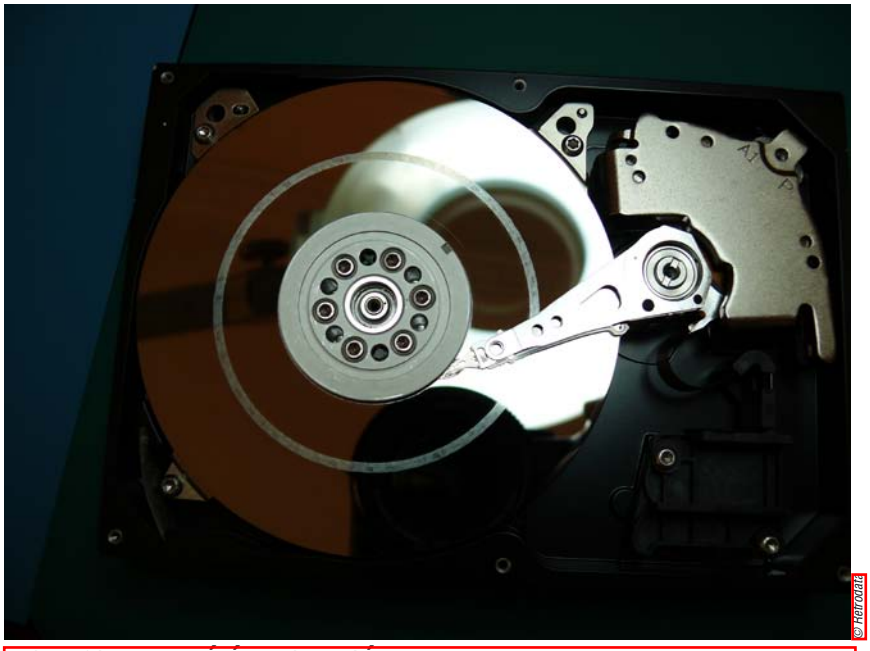
Derrière cette merveille de la copie sans erreurs se cache le travail de mathématiciens, tels les Américains Irving Reed et Gustave Solomon, qui, dans les années 1960, ont mis au point des codes de correction d'erreurs numériques permettant des millions de copies sans erreurs d'un fichier. L'algorithme est complexe, mais son principe est simple : il faut « suréchantillonner » l'information, c'est-à-dire donner plus d'éléments qu'il n'est nécessaire pour restituer l'information. La redondance permet de vérifier qu'il n'y a pas eu d'erreurs dans la copie. Les algorithmes sont si efficaces que si nous recopions un fichier légèrement endommagé, sur un CD de données un peu ancien par exemple, la copie éliminera les erreurs apparues entre-temps – tant que le taux d'erreurs apparues n'est pas trop grand. Le fichier sera restauré dans son état initial !

Le numérique semble avoir les mêmes vertus magiques que le Phénix renaissant de ses cendres. Et pourtant, il a un talon d'Achille, et pas des moindres : la préservation à long terme des données. Certes, il est de peu d'intérêt de conserver la plupart d'entre elles, dont la pertinence est éphémère ; mais il est essentiel d'en préserver la petite fraction qui garde un intérêt pendant des années, des décennies, parfois des siècles. L'intérêt de certaines données croît avec leur ancienneté, comme les données des satellites d'observation de la surface terrestre : en période de réchauffement climatique, il devient important de comparer les observations actuelles aux plus anciennes possibles, afin de discerner des évolutions à long terme. Les données des grands instruments scientifiques (telles celles du CERN, du satellite *Planck*, du radiotélescope SKA ou *Square Kilometer Array*) sont irremplaçables, rien ne garantissant qu'il sera possible un jour de répéter ces expériences très coûteuses. À plus petite échelle, les données du cadastre, de la comptabilité des entreprises, les données médicales, etc., doivent être conservées pendant longtemps. Enfin, à titre individuel, chacun d'entre nous souhaite conserver les souvenirs de sa vie personnelle et familiale.

La conservation de données numériques n'est donc pas une question anodine, puisqu'elle touche de nombreux domaines. Elle est confrontée à trois obstacles : documenter les fichiers pour savoir de quoi il s'agit, conserver le matériel et les logiciels nécessaires à la lecture du fichier et lutter contre le vieillissement des supports physiques.

Les deux premiers points sont ceux auxquels chacun songe avant tout. Toutes les formations de documentaliste enseignent l'importance des métadonnées. Celles-ci regroupent les informations qui accompagnent un document : sa nature, sa date, ses auteurs, des mots clefs associés, etc. Ces éléments facilitent une exploitation future.

Outre le problème des métadonnées, chacun a une anecdote à raconter de disquettes contenant des fichiers importants, mais illisibles faute de matériel adéquat.



1. CE DISQUE DUR A ÉTÉ ENDOMMAGÉ par la tête de lecture qui a touché le matériau ferromagnétique à la surface du disque et tracé le sillon circulaire. Les données enregistrées sur le disque et situées dans le sillon sont définitivement perdues.

Qui possède encore un lecteur de disquette souple ou un lecteur de disque Zip ou Jaz ? De nombreux supports sont devenus obsolètes parce qu'ils ont été remplacés par des modèles plus performants ou parce qu'ils ont été des échecs commerciaux.

Ces deux aspects de la pérennité de l'information peuvent être réglés avec un peu d'organisation et de bon sens. Transposer au numérique le dépôt de métadonnées qui était déjà essentiel pour l'analogique n'est pas hors de portée ; de plus, on peut parfaitement choisir des systèmes d'exploitation, des logiciels et des formats de fichiers ouverts – dont le code est accessible et que chacun peut utiliser – ou très répandus, qui offrent des perspectives raisonnables de pérennité *(voir l'article de Roberto Di Cosmo dans ce numéro)*.

Ces deux premiers problèmes sont importants, mais ils ne doivent pas masquer le troisième, qui est bien plus grave. On peut prendre toutes les précautions que l'on voudra en créant des métadonnées parfaites et en utilisant des formats robustes : si les données elles-mêmes disparaissent, cela ne sert évidemment à rien ! Ce problème a longtemps été sous-estimé et négligé. Hélas, nous allons voir que cet aspect de la pérennité des données pose des difficultés techniques et économiques qui sont loin d'être résolues.

La variété des dispositifs de stockage disponibles dans le commerce et leur évolution rapide peuvent donner l'impression qu'il existe de nombreuses méthodes différentes pour stocker et sauvegarder de l'information numérique. Certains supports sont-ils adéquats pour de l'archivage de longue durée ? En réalité, ces dispositifs fonctionnent sur un petit nombre de principes physiques et se rangent dans quelques catégories simples qui subissent toutes des détériorations en quelques années.

Tous les supports s'altèrent

La première catégorie fait appel à l'enregistrement magnétique : le système d'enregistrement est conçu pour orienter de façon contrôlée des petits domaines dans une couche magnétique déposée sur une surface plane. La lecture s'effectue ensuite par une tête électromagnétique sensible à cette orientation. C'est ainsi que fonctionnent les disques durs, les bandes magnétiques (ou LTO, pour *Linear Tape Open*) et même les disquettes de formats divers, dont la plupart ont maintenant disparu à cause du succès des mémoires flash. Les disques durs présentent de grandes capacités de stockage pour un coût faible, et avec un accès rapide aux données. Ils sont cependant fragiles et soumis à des pannes soudaines et imprévisibles (ou *crash*) pouvant entraîner une perte totale des données (par exemple, si la tête de lecture touche la couche magnétique, *voir la figure 1*). Pour ne pas perdre ses données, une possibilité est la sauvegarde sur plusieurs disques en parallèle (tels les systèmes redondants RAID, pour *Redundant Array of Independent Disks*). La longévité des disques durs est difficile à évaluer, car elle dépend de nombreux facteurs indépendants. On s'accorde à penser néanmoins que, hors *crash*, une durée de vie de cinq ans est un bon ordre de grandeur. Pour optimiser cette longévité, il faut laisser tourner les disques durs en permanence, ce qui minimise les incidents. Les grands centres de données appliquent cette méthode coûteuse en électricité. On estime que les centres de données contribuent ainsi à plus de deux pour cent de la production de dioxyde de carbone d'origine humaine.

Il y a quelques années, les bandes magnétiques étaient considérées comme obsolètes, car moins performantes que les disques durs. Mais, précisément du fait de la nécessité de l'archivage des données, les bandes magnétiques ont effectué un retour spectaculaire sur la scène du numérique : elles conservent les données de façon passive, sans consommation électrique. Cependant, leur durée de vie est aussi limitée, de l'ordre de cinq ans, temps au bout duquel il devient indispensable d'effectuer une « migration » – c'est-à-dire de recopier les données sur une bande neuve, si on ne veut pas les perdre. Il existe d'ailleurs des robots dont la fonction est de gérer un grand nombre de bandes, d'y suivre le nombre d'erreurs et d'effectuer une migration vers des bandes neuves dès que c'est nécessaire.

La seconde catégorie de supports numériques fait appel à un enregistrement électrique et non magnétique : dans les mémoires flash, de très nombreux petits transistors stockent une charge électrique qui définit la valeur du bit enregistré. Les mémoires flash sont devenues d'usage universel, dans les appareils photo, les caméscopes, les ordinateurs, les téléphones, etc. Elles autorisent des accès aux données encore plus rapides que les disques durs, mais le prix par mégaoctet de données est plus élevé. Quant à la longévité des données stockées sur des mémoires flash, on dispose de peu d'informations à ce sujet. On sait cependant que les rayons cosmiques

© Pour la Science - n° 433 - Novembre 2013

déchargent progressivement les transistors et effacent donc les données. On considère souvent que la durée de vie est de l'ordre de cinq ans.

La troisième catégorie est constituée par les disques optiques numériques (DON), tels que les CD, les DVD et les disques *Blu-ray*. Les temps d'accès aux données enregistrées sont bien plus longs que ceux des disques durs et des mémoires flash. De plus, leur capacité est nettement inférieure à celle des disques durs. On fabrique pourtant chaque année des milliards de disques optiques numériques, encore que leur nombre soit en constante diminution chaque année ; beaucoup pensent qu'ils sont en voie d'extinction progressive. En termes de longévité toutefois, ce sont les meilleurs supports disponibles actuellement : les disques « pressés » fabriqués en grande série ont une durée de vie qui serait comprise entre 20 et 30 ans. Cependant, ce n'est pas le cas des disques enregistrables, qui reposent sur des principes physiques plus délicats (inscription dans un colorant composé de molécules photosensibles ou, pour les disques réenregistrables, un changement de phase qui rend la surface plus ou moins réfléchissante). Les études du Laboratoire national de métrologie et d'essais en France et du Laboratoire lutherie acoustique musique de l'Université Paris 6 ont montré que certains disques enregistrés perdent leurs données en un an seulement ! D'autres, heureusement, ont des durées de vie dix fois plus longues.

Les disques optiques offrent des perspectives très intéressantes en termes de longévité. Il y a une vingtaine d'années une société française, *Digipress*, a commercialisé un CD ayant une excellente durée de vie, puisque les données étaient directement gravées dans du verre. Hélas, l'invention devait être trop en avance sur son temps, à une époque où beaucoup croyaient encore à la longévité des supports numériques. Ce produit n'a pas trouvé son marché. Mais, depuis deux ans, une société allemande, *Syylex*, a repris l'idée et propose des DVD gravés dans le verre dont la longévité est remarquable, comme l'ont montré des études du Laboratoire national de métrologie et d'essais *(voir l'encadré page 82)*. L'avantage d'une telle solution est que les données sont stockées dans le format DVD, très répandu dans le monde, et qu'elle ne demande aucun lecteur plus spécifique qu'un lecteur standard. Quelques autres produits sont proposés sur le marché et annoncés avec une grande longévité, mais à notre connaissance, aucun n'a subi des tests aussi soigneux que ces disques. Il faut espérer que les esprits aient maintenant suffisamment évolué pour que ce nouveau produit se répande rapidement, en particulier par des commandes institutionnelles, mais rien ne le garantit.

Graver dans le verre

Quelles stratégies peuvent être adoptées pour préserver l'information numérique à long terme avec les dispositifs existants ? Deux approches sont possibles : la stratégie dite passive, qui consiste à écrire l'information sur un support pérenne et à le stocker dans un endroit sûr ; la stratégie dite active, qui consiste à écrire l'information sur un support numérique standard et à en suivre constamment l'évolution, afin d'effectuer une migration vers un support neuf lorsque nécessaire. La stratégie passive est celle que l'humanité a pratiquée depuis toujours, mais elle utilisait des supports analogiques – que l'on songe à la bonne conservation des parchemins anciens, voire aux écritures cunéiformes sur des pierres ! Elle semble donc la plus naturelle, mais dans le cas du numérique, elle demande un investissement initial, par exemple pour la confection de DVD gravés dans le verre.

La stratégie active ne demande pas un tel investissement. Dans la plupart des cas,

L'altération, un processus mal compris

Les différents supports numériques ont des durées de vie relativement courtes. De nombreux facteurs entrent en jeu et il est difficile de dire précisément quels phénomènes sont en cause et lesquels sont les plus importants. Pour les disques durs, les pannes mécaniques sont une difficulté majeure. Mais d'autres problèmes peuvent surgir : certains éléments du disque stockant l'information peuvent devenir défectueux et illisibles.

Les rayons cosmiques peuvent décharger les transistors des mémoires flash et modifier l'information. Les supports optiques, les CD et les DVD, correspondent à un empilement de couches de protection en plastique, de couches contenant les données et d'une pellicule métallique qui sert de réflecteur pour le laser de lecture des données. La pellicule plastique est sensible à la surexposition aux ultraviolets de la lumière du Soleil et à la chaleur. Les tests de longévité de ces supports ont aussi montré qu'ils sont sensibles aux réactions d'oxydation par l'eau et par l'air. Les supports optiques enregistrables sont encore moins stables.

Les facteurs externes ne sont pas les seuls à influer sur la durée de vie des supports. La qualité des matériaux utilisés et les précautions prises lors de la fabrication sont aussi importantes. Le Laboratoire national de métrologie et d'essais a montré que sept pour cent des DVD enregistrables du commerce présentaient des microtaches. Elles seraient dues à des impuretés présentes dans la couche réflective et accéléreraient le processus d'oxydation, réduisant la durée de vie du support à un an.

LES AUTEURS

Franck LALOE est directeur de recherche au Laboratoire Kastler Brossel de l'École normale supérieure. Il est président du Groupement d'intérêt scientifique SPADON. Il a été rapporteur du groupe Pérennité des supports numériques.

Erich SPITZ est membre de l'Académie des sciences et de l'Académie des technologies. Il a présidé le groupe Pérennité des supports numériques.

BIBLIOGRAPHIE

J. Zhang *et al.*, **5D data storage by ultrafast laser nanostructuring in glass**, compte rendu de conférence CLEO : Science and Innovations, 2013.

J. Perdereau, **Tenue dans le temps des disques enregistrables DVDR et DVD en verre (Syylex) à température et hygrométrie élevées**, rapport du GIS DON, LNE, 2012.

J.-C. Hourcade *et al.*, **Longévité de l'information numérique ; les données que nous voulons garder vont-elles s'effacer ?, rapport à l'Académie des Sciences et à l'Académie des Technologies**, EDP Sciences, 2010.

l'archivage ne se fait pas localement, mais dans des centres de données spécialisés, ou « fermes », contenant de nombreuses armoires remplies de disques durs. Ces centres sont gérés par du personnel spécialisé, dont une des missions est de procéder aux migrations de supports quand cela est nécessaire. Cette solution fonctionne de façon satisfaisante, et elle est bien dans la ligne de l'engouement actuel pour le *cloud*. Elle pose néanmoins des difficultés. La première est que, sauf s'il dispose d'un centre informatique évolué, le propriétaire des données perd le contrôle de l'archivage ; il doit faire confiance à un prestataire de service, sachant que toute erreur de sa part peut engendrer une perte de données irréversible. Il faut aussi prévoir à l'avance le cas où ce prestataire viendrait à cesser ses activités, et les procédures de restitution des données. Le second problème est celui de la confidentialité : il est évident que des données sensibles sont mieux protégées si elles sont à l'abri dans un coffre-fort sur place que si elles sont transportées à des milliers de kilomètres.

La comparaison des coûts des deux stratégies est bien sûr déterminante. Celui de la stratégie passive est facile à évaluer : le coût de transfert des données sur des supports spéciaux et de très grande longévité peut être élevé ; par exemple, un DVD de cinq gigaoctets de données gravées dans le verre coûte actuellement à peu près 150 euros, soit plus de 100 fois le prix d'un DVD enregistrable standard ! Il est clair que la stratégie passive implique une dépense initiale importante. Ensuite, les coûts sont réduits : celui lié à l'immobilisation d'un espace d'archivage, par exemple dans une cave ou un coffre-fort.

La méthode passive, plus économique

Pour la stratégie active, il en va tout autrement. L'erreur classique est de ne prendre en compte que les coûts des supports nécessaires pour les migrations, coûts qui sont négligeables. Les coûts réels sont en fait induits par la nécessité du suivi technique (salaires des informaticiens, locaux avec environnement technique) et la consommation électrique des « fermes » de disques durs. Ces coûts croissent donc linéairement avec le temps, contrairement à ceux de la stratégie passive, qui restent quasi constants ; chacun comprend donc qu'il existe un certain délai à partir duquel c'est la stratégie passive qui devient la moins onéreuse.

Des comparaisons entre les coûts des deux méthodes ont été effectuées. Ainsi, pour l'archivage de 500 gigaoctets, la stratégie active sur bande magnétique est la moins chère pendant six à sept ans, puis c'est la stratégie passive sur DVD gravés en verre ; pour l'archivage de un téraoctet, le point de croisement est repoussé à 10 ans environ. Il est intéressant de noter que cette étude montre que la stratégie active sur disques durs RAID, pourtant la plus répandue aujourd'hui, n'est jamais la moins onéreuse, que ce soit à court ou à long terme.

À partir de ce simple constat, les organismes et entreprises qui ont besoin de conserver d'importants volumes de données pendant plus de dix ans devraient opter

Évaluation de la tenue dans le temps du DVD en verre *Syylex*

Des études menées par le Laboratoire national de métrologie et d'essais (LNE) sur le vieillissement naturel des disques optiques numériques enregistrables CD-R et DVD-R ont montré une grande disparité de comportement des disques dans le temps, avec des durées de vie comprises entre un an et plus de 20 ans selon les modèles. C'est en contradiction avec les durées de vie de 50 à 300 ans annoncées par les fabricants.

Les membres du Groupement d'intérêt scientifique SPADON ont souhaité étudier des solutions différentes des DVD-R classiques, tel le DVD *GlassMasterDisc* de la Société *Syylex*. Il n'utilise ni polycarbonate ni de couche d'enregistrement organique, points faibles potentiels des DVD-R classiques. L'information est directement gravée dans du verre.

L'estimation de la durée vie de ce disque aurait pu se faire en utilisant la méthode normalisée par vieillissement accéléré en température et humidité (ISO/CEI 16963). Cependant, d'une part, cette norme n'a pas été validée expérimentalement, d'autre part, elle ne semble pas applicable aux DVD, au vu des résultats de plusieurs études antérieures, probablement parce que les facteurs d'altération physico-chimiques sont multiples. Il n'est donc pas possible d'estimer la durée de vie réelle des DVD à partir d'un test de vieillissement accéléré. Pour étudier l'efficacité du *GlassMasterDisc* pour l'archivage, le GIS-SPADON et le LNE ont choisi d'évaluer la tenue dans le temps de ce support exposé à des températures et des taux d'humidité élevés et de comparer les résultats aux DVD enregistrables classiques. Les six modèles choisis comme éléments de comparaison sont réputés comme ayant une bonne tenue dans le temps (cas des DVD-R utilisant une couche enregistrable inorganique) ou ont montré dans les études précédentes du LNE une tenue au vieillissement accéléré (80 °C et 85 pour cent d'humidité) supérieure à celle des autres modèles.

L'étude a consisté à placer les DVD dans une enceinte climatique *(voir la photographie ci-contre de l'enceinte CLIMATS)* dans des conditions dites très sévères de température et d'humidité, pour une durée maximale de 1 000 heures, et cela par tranches de 250 heures. Le DVD *GlassMasterDisc* étant en verre, la température a été choisie égale à 90 °C, soit 10 °C de plus que pour les précédentes études de DVD, tout en maintenant l'humidité relative à 85 pour cent. La fin de vie des DVD correspond à un taux de données numériques erronées dépassant les limites normalisées ou à l'illisibilité du disque.

Il ressort de cette étude, qu'en vieillissement accéléré, le DVD *GlassMasterDisc* a une durée de vie nettement supérieure à celle des DVD-R existants sur le marché. Après 1 000 heures d'exposition dans ces conditions délétères, le DVD *GlassMasterDisc* ne présente pas d'altération significative de sa qualité, alors que le meilleur des autres modèles de DVD-R a une durée de vie maximale de 250 heures.


J. Perdereau, LNE

Le DVD *GlassMasterDisc* présente donc le meilleur potentiel, parmi les DVD enregistrables existants sur le marché, pour réaliser un archivage de très longue durée des données numériques.

Jacques Perdereau,
Ingénieur au LNE

© Pour la Science - n° 433 - Novembre 2013

Actuellement en kiosque

DOSSIER POUR LA SCIENCE

- Le bestiaire des ARN
- La mémoire des cellules
- La matière noire du génome
- Le cerveau modelé par l'expérience

Vers une médecine personnalisée

Le magazine thématique de l'actualité scientifique

N° 81 Octobre-Décembre 2013

L'hérédité sans gènes

Comment l'environnement rebat les cartes



DOSSIER HORS-SÉRIE

POUR LA SCIENCE • DOSSIER N° 81 • OCTOBRE-DÉCEMBRE 2013

n°81 - 120 pages - prix de vente: 6,95€

www.pourlascience.fr
le site de référence de l'actualité scientifique internationale

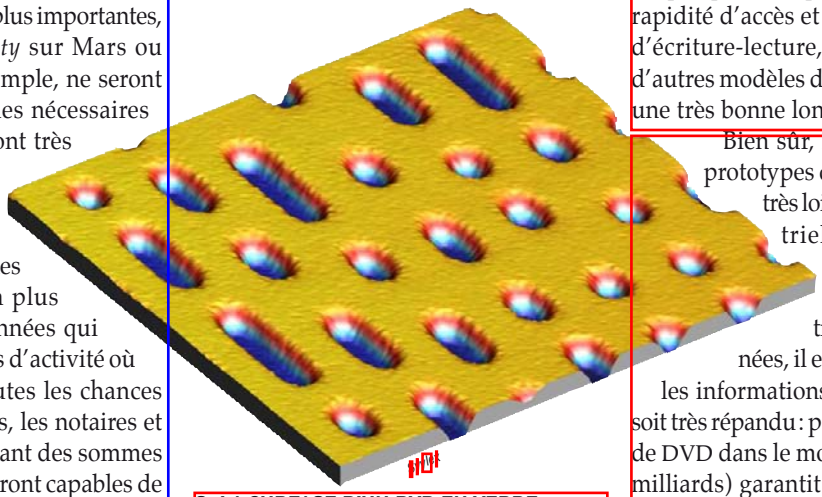
Disponible sur App Store
DISPONIBLE SUR Google play

pour la solution passive, la moins chère de toutes. Il est étonnant de constater qu'il n'en est rien, et de très loin, car presque personne ne s'intéresse encore aux supports numériques pérennes. La raison en est peut-être que beaucoup de décideurs, et même de particuliers, n'ont pas pris la mesure du problème et préfèrent conserver leurs illusions plutôt que d'envisager des catastrophes futures potentielles ; et cela nécessite d'engager des politiques à moyen et long termes. La stratégie active reste donc à peu près la seule utilisée dans le monde, à quelques exceptions près, comme à l'Institut national de l'audiovisuel, en France, pour les émissions radiophoniques.

Cependant, évitons un alarmisme excessif : les données les plus importantes, celles du robot *Curiosity* sur Mars ou celles du CERN par exemple, ne seront pas perdues ; les sommes nécessaires à leur conservation seront très probablement dépensées, car elles restent une faible fraction du coût de ces grands programmes de recherche. De façon plus générale, toutes les données qui touchent à des domaines d'activité où l'argent abonde ont toutes les chances de survivre : les banques, les notaires et d'autres professions gérant des sommes importantes sont et resteront capables de mettre en œuvre la stratégie active dans de bonnes conditions.

Assurer la pérennité des données est difficile dans d'autres secteurs. Pour les hôpitaux, les données médicales doivent être légalement conservées au minimum 20 ans, une gageure dans un contexte de restrictions budgétaires. Et dans bien d'autres cas, on peut être encore plus inquiet. Il est d'ores et déjà connu que les données issues de la numérisation de documents analogiques ont parfois été perdues (dans certaines archives départementales par exemple), et l'on est loin d'avoir une vue générale sur la question. Quant aux données personnelles et familiales, tout porte à croire qu'elles n'auront pas la durée de vie des photos ou des films huit millimètres que nous ont transmis nos parents.

Peut-on raisonnablement espérer voir émerger des supports numériques peu onéreux, permettant le stockage de grandes quantités d'informations et garantissant leur conservation pendant des siècles sans migration ? Il semble que les grands laboratoires industriels se sont longtemps concentrés sur la fiabilité, le volume de données stockées et la rapidité d'accès, mais une prise de conscience progressive et l'émergence d'un marché centré sur la longévité pourraient changer la donne. De plus, des structures publiques commencent à s'intéresser au sujet. Le Groupe d'intérêt scientifique SPADON *(Supports pérennes d'archivage pour les données numériques)* regroupe plusieurs laboratoires universitaires et du CNRS. Il suit et coordonne les progrès techniques dans ce domaine, et suscite de nouvelles recherches. De même, le groupe Pérennité des informations numériques, auquel participent, entre autres, la Bibliothèque nationale de France et le Centre informatique national de l'enseignement supérieur (CINES), effectue un suivi depuis plusieurs années.



2. LA SURFACE D'UN DVD EN VERRE vue au microscope à force atomique. L'information est gravée directement dans le verre. Ce support offre la meilleure durée de vie et la solution la plus économique à long terme.

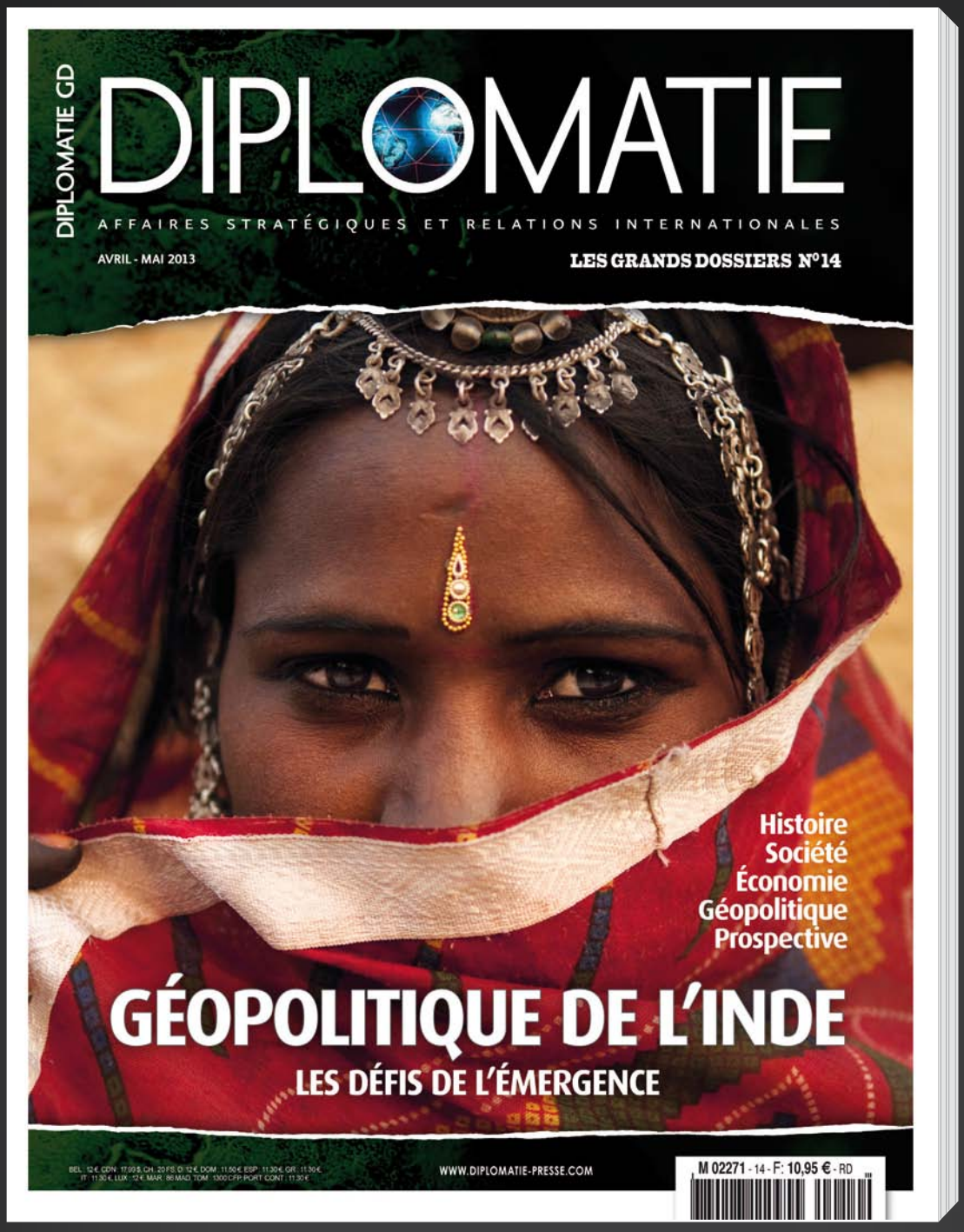
Ces efforts commencent à porter leurs fruits. Une voie prometteuse semble être fournie par des dispositifs optiques, avec l'écriture par des lasers à impulsions (nanosecondes ou femtosecondes) à l'intérieur même d'une structure vitreuse. Récemment, dans le cadre du projet européen *Femtoprint*, un groupe de l'Université de Southampton a annoncé avoir mis au point une technique d'enregistrement « à cinq dimensions » en créant, avec un laser ultrarapide, des nanostructures dans le verre – les cinq dimensions correspondent aux trois dimensions d'espace habituelles, puisque l'inscription se fait en volume, plus deux dimensions pour la taille et l'orientation des nanostructures. La capacité de stockage extrapolée est de plus de 300 téraoctets par disque de silice, et la durée de vie serait pratiquement infinie ! On a aussi beaucoup parlé récemment des possibilités offertes par les skyrmions, des configurations quantiques particulières dans les matériaux magnétiques, ou même de l'enregistrement de données dans l'ADN *(voir l'article de Christophe Dessimoz dans ce numéro)* ! Enfin, de façon moins futuriste, rien n'interdit d'imaginer que les mémoires flash, jusqu'à présent optimisées en termes de rapidité d'accès et de nombre de cycles d'écriture-lecture, puissent engendrer d'autres modèles de dispositifs assurant une très bonne longévité.

Bien sûr, il ne s'agit là que de prototypes de laboratoire, encore très loin de toute offre industrielle. De plus, il faut garder à l'esprit que, pour un archivage à très long terme des données, il est essentiel de stocker les informations dans un format qui soit très répandu : par exemple, le nombre de DVD dans le monde (des dizaines de milliards) garantit probablement qu'ils resteront lisibles dans un siècle ou plus. Un support d'information numérique pérenne doit donc atteindre un certain niveau de diffusion pour être un bon candidat à l'archivage.

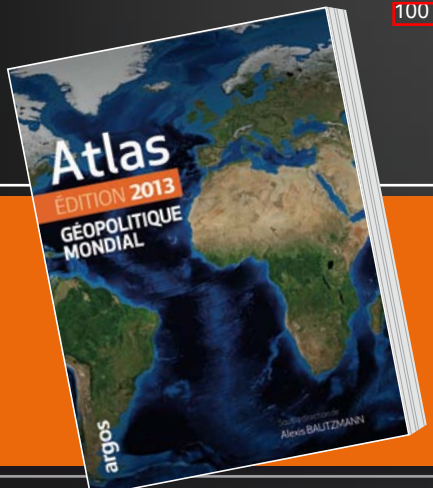
Aujourd'hui, le DVD gravé dans le verre est la meilleure piste. Il résiste bien à l'altération, mais sa capacité est limitée et son prix relativement élevé. C'est probablement pourquoi il est très peu utilisé. L'enjeu de la pérennité des données numériques est important, et il est crucial de trouver des solutions. Certes, cela nécessite une bonne organisation avec l'usage de métadonnées et de normes sur les formats pour les conserver sur le long terme. Mais l'aspect le plus important et le plus difficile à régler est celui de l'altération des supports numériques. La prise de conscience du problème va encourager la recherche et le développement de dispositifs à longue durée de vie. Il est difficile de dire si une solution pourrait apparaître bientôt et d'où elle viendrait. De l'optique ? De la physique du solide ? ■

© Pour la Science - n° 433 - Novembre 2013

En vente chez votre marchand de journaux



100 pages • Format 23x30 cm • 10,95 € (codification Presstalis : 02271)



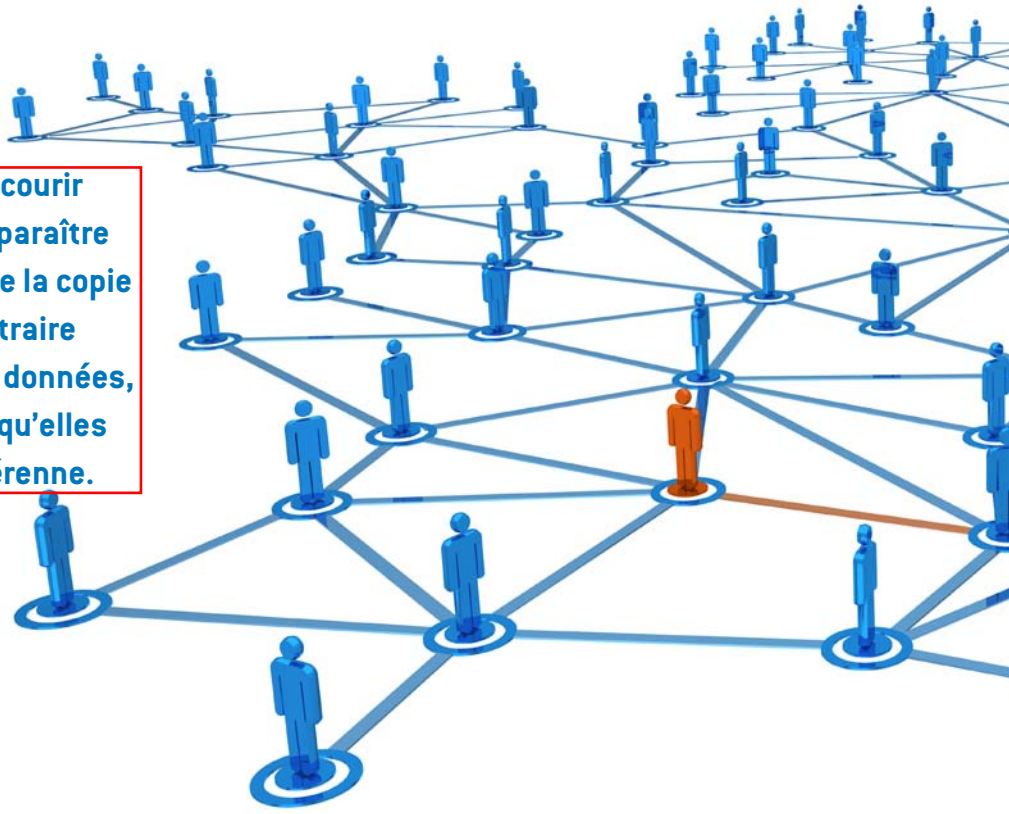
Commandez DIPLOMATIE sur Internet www.GEOSTRATEGIQUE.com

Et ne manquez pas en librairie

ATLAS GÉOPOLITIQUE MONDIAL 2013

270 cartes et graphiques pour décrypter toute l'actualité internationale
(Éditions ARGOS/AREION - Diffusion PUF - 23x30 cm - 212 pages - 22,50 €)

AREION GROUP • 91 rue Saint-Honoré • 75001 Paris (France) • Tél. +33(0) 175 435 271 • Fax : +33 (0) 811 622 931 • E-mail : diplomatie@areion.fr



Centraliser les données, c'est courir le risque de les voir toutes disparaître d'un seul coup. L'avènement de la copie numérique facile invite au contraire à partager le plus possible les données, afin d'augmenter les chances qu'elles soient conservées de façon pérenne.

Roberto Di Cosmo

Organiser le partage pour préserver les données

Qu'il s'agisse d'histoire, de philosophie, d'avancées scientifiques, d'œuvres littéraires ou artistiques, ou tout simplement d'histoires individuelles, l'humanité a toujours cherché à préserver ses savoirs et à les transmettre aux générations futures. Pour ce faire, elle a d'abord employé les récits oraux et les mythes. Puis l'invention de l'écriture a entraîné celle des livres, et tout a changé le jour où ces « recueils » ont pu être multipliés grâce à l'imprimerie, avant d'être partagés en masse dans les bibliothèques. Aujourd'hui, la copie numérique et Internet permettent une diffusion des connaissances plus efficace que jamais. Analysons les modalités, les risques et les promesses de la transmission numérique du savoir, une innovation, qui, une fois de plus, change notre vie culturelle.

Dupliquer des informations a longtemps été difficile. Pensons, par exemple, au coût prohibitif de la reproduction d'un livre par un copiste avant l'invention de l'imprimerie. La seule stratégie possible consistait à réunir les précieux recueils en un endroit à la fois sûr et accessible, où l'on pouvait les protéger tout en les rendant disponibles à la consultation à l'aide d'index centralisés. La grande bibliothèque d'Alexandrie, où, pendant l'époque hellénistique (entre –336 et –30), on tenta de rassembler « tout le savoir du monde », illustre que cette stratégie est ancienne. La destruction de cette bibliothèque, dans les premiers siècles de notre ère, montre les risques, bien connus des informaticiens, que courent les systèmes centralisés, puisqu'il suffit d'en détruire le centre pour tout faire disparaître…

Grâce aux avancées techniques du dernier demi-siècle, nous vivons aujourd'hui une nouvelle rupture dans l'histoire culturelle de l'humanité.

La copie parfaite…

Pour la première fois, nous disposons d'instruments pour répliquer parfaitement un nombre toujours plus grand de biens immatériels. Les moyens informatiques permettent en effet de numériser – c'est-à-dire d'encoder sous forme d'une suite de chiffres binaires – du texte, du son, des images et de la vidéo, puis de répliquer à volonté ces informations sans perte de qualité et pour un coût négligeable. Tout cela est devenu banal, et le domaine des biens duplicables à l'identique ne cesse

© Pour la Science - n° 433 - Novembre 2013



de s'élargir, comme l'illustre l'arrivée des scanners et des imprimantes 3D.

Dès lors, on est tenté de penser que le problème ancien de la préservation des données est résolu… Après tout, quand il est possible de dupliquer, sans perte de qualité, et à coût quasi nul, autant de fois qu'on le souhaite, comment pourrait-on encore perdre de l'information ?

… n'existe pas !

Mais à y regarder de plus près, on s'aperçoit vite qu'il n'en est rien. Non seulement les techniques actuelles de reproduction n'ont pas résolu le problème, mais elles l'ont en fait compliqué. En premier lieu, les mêmes outils qui rendent facile la duplication de l'information ont aussi pour conséquence d'engendrer dans la pratique une quantité gigantesque d'information : pour l'année 2011, on a calculé que 1,8 zettaoctet de données ont été produites *(voir la figure 2)*, et huit zettaoctets sont annoncés pour 2015 ! Précisons que la plus grande partie de cette information n'est pas créée directement par l'homme, mais par nos machines, qui gardent non seulement les traces numériques de nos opérations informatiques, telles la transmission ou la conversion de fichiers, mais aussi les données et les métadonnées (données sur les données) produites par nos échanges sur les réseaux sociaux…

Décider de ce qui doit être préservé dans cette gigantesque masse et de ce qui peut tomber dans l'oubli n'est pas aisé. Ainsi, la Bibliothèque du Congrès des États-Unis archive désormais les messages envoyés par les citoyens américains sur le réseau *Twitter*, alors que ces commentaires et autres micro-informations devaient à l'origine n'être qu'éphémères…

Même si l'on disposait d'une méthode pour décider exactement de ce qui doit être préservé, la conservation à long terme des objets numériques serait bien plus problématique qu'il n'y paraît. À la différence des livres et des tableaux, qui livrent immédiatement leur contenu, l'accès aux objets numériques suppose un environnement technique complexe. Ce dernier comporte, d'une part, des supports matériels de stockage de l'information susceptibles de se dégrader et, d'autre part, des logiciels complexes, utilisant des formats définis et fonctionnant seulement sur certains processeurs qui, eux, sont susceptibles de devenir rapidement obsolètes. À moyen et long termes, les risques de pertes de données numériques sont donc considérables.

Illustrons cette réalité peu connue par un exemple : si l'on retrouvait aujourd'hui un document électronique rédigé il y a seulement… 30 ans, sans doute aurait-il la forme d'une disquette magnétique de 5 pouces 1/4 contenant un fichier *WordStar*. À supposer que cette grosse disquette ne soit pas trop dégradée, ce qui est fort peu probable *(voir l'article de Franck Laloë et Erich Spitz dans ce numéro)*, nous nous heurterions à une autre difficulté : les lecteurs de tels supports ont presque tous disparu. De même, le logiciel *WordStar* n'est plus employé (car il n'est plus disponible), de sorte que le format dans lequel il enregistrait les données n'est plus reconnu par les traitements de texte actuels *(voir la figure 4)*.

L'ESSENTIEL

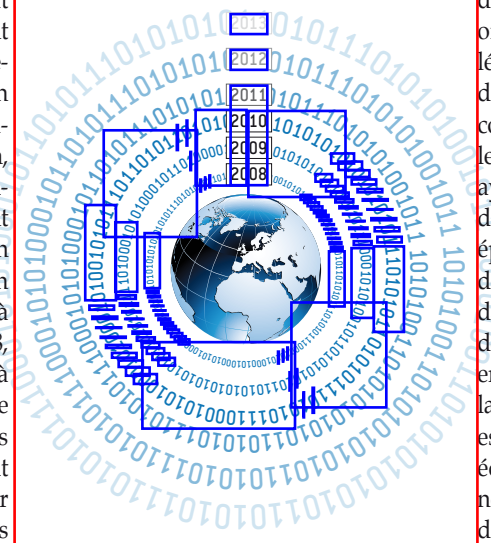
- La réplication numérique de données et leur diffusion par Internet semblent faciliter la conservation de la mémoire humaine.
- En réalité, elle la menace, car les moyens techniques de cette réplication se périment et rendent les données inaccessibles.
- Les offres d'hébergement gratuit des grands opérateurs d'Internet, pour utiles qu'elles soient, ne constituent pas non plus une solution sans risques.
- Il reste à refaire sur le réseau ce que l'humanité a toujours fait : répartir sa mémoire entre tous !

© Shutterstock.com/resta

1. LE RÉSEAU HUMAIN est depuis toujours le dépositaire de l'expérience de l'humanité. Pour en faire aussi le dépositaire de sa mémoire numérique, encore faut-il faciliter les échanges de données entre pairs…

Le cas du projet *Domesday* de la BBC (société britannique de radiotélévision) fournit un exemple de ces difficultés techniques. En 1986, soit quelque 900 ans après une entreprise similaire *(voir l'encadré page 90)* ordonnée par Guillaume le Conquérant (1027-1087), la BBC a lancé un recensement et un sondage sur l'ensemble du Royaume-Uni. Les données, recueillies par plus d'un million de volontaires dans les 23 000 cellules territoriales définies pour l'occasion, constituaient une sorte d'immense instantané de la société britannique ; elles furent mises à la disposition du public au moyen d'un vidéodisque de type *LaserDisc*, un support des données dont on considérait à l'époque qu'il serait durable. Or, dès 1998, il fut supplanté par les DVD, de sorte qu'à partir du début des années 2000, on ne trouvait déjà pratiquement plus de lecteurs de *LaserDisc*, ni d'ordinateurs où pouvait tourner le logiciel nécessaire pour présenter le contenu du disque. Résultat : plusieurs équipes durent fournir un important effort de recherche afin d'émuler (d'imiter) le fonctionnement des lecteurs de *LaserDisc* dans un environnement informatique actuel.

Ce cas montre que pour préserver des logiciels, il importe de maintenir les connaissances nécessaires pour les faire exécuter par une machine, et donc de disposer de leur code source et de leur environnement d'exécution. Dans le cas du *LaserDisc*, quelques années avaient suffi pour que tout cela disparaisse ! C'est bien à cause de telles situations qu'aujourd'hui, on reconnaît de plus en plus que l'usage de formats de données ouverts (qui peuvent être librement utilisés et modifiés, et dont les spécifications sont publiques et sans restriction d'accès) et de logiciels libres (dont le code source est disponible pour tous) est indispensable à la pérennité des informations.



2. EN 2011, 1,8 ZETTAOCTET de données a été produit, et diverses estimations en annoncent huit pour 2015. Un zettaoctet représente 1 000 milliards de gigaoctets (c'est-à-dire 10^{21} octets). Ainsi, en quelques jours, l'humanité accumule aujourd'hui plus d'information qu'elle n'en a produit en deux millions d'années avant l'invention d'Internet.

Mais une fois le problème de la lecture des *LaserDisc* du projet *Domesday* résolu, on s'aperçut que d'épineux problèmes légaux restaient à démêler. Selon le droit d'auteur en vigueur, la réalisation d'une copie exigeait que l'on recueille d'abord les autorisations des millions d'individus ayant contribué au projet, ainsi que celles des auteurs des logiciels utilisés… À une époque où reproduire l'information est devenu banal, l'imposant arsenal juridique des droits d'auteurs nous interdit nombre de réplications, et ce sur des durées de plus en plus longues… Si le droit protégeant la propriété intellectuelle sert aujourd'hui essentiellement à préserver les modèles économiques ayant précédé l'ère d'Internet, il a souvent pour effet de bloquer le développement des modèles nécessaires à l'économie numérique.

Copier est-il hors la loi ?

Si nous voulons préserver les données, nous sommes donc confrontés à un dilemme : nous devons d'un côté les recopier sur de nouveaux supports (avant leur inévitable dégradation) et sous de nouveaux formats (avant leur inévitable obsolescence) ; d'un autre côté, nous devons respecter des lois qui nous interdisent de réaliser ces copies… Notre civilisation est donc confrontée à la question suivante : qui doit se charger du renouvellement continu des données et dans quelles conditions ? Plusieurs


© Shutterstock.com / PoOPaz

3. LES SUPPORTS DE STOCKAGE se suivent et ne se ressemblent pas, tant en termes de capacité que de format d'enregistrement. Ci-dessus, quelques-uns des formats datant des 30 dernières années : à gauche, une disquette 5 pouces 1/4, au milieu une disquette 3 pouces 1/2 de haute densité, puis un CD-ROM et, à droite, une clef USB actuelle.



4. ***WORDSTAR*** est un logiciel de traitement de texte introduit en 1978, qui fonctionnait sous DOS, le système d'exploitation pour ordinateurs personnels de *Microsoft*. Il occupa une position dominante au début des années 1980, mais fut supplanté par *Word* dans les années 1990, puis abandonné et oublié, au point de devenir inutilisable aujourd'hui.

© Pour la Science - n° 433 - Novembre 2013

réponses sont possibles. Chacune a ses avantages et ses inconvénients.

Une première possibilité serait de sous-traiter la préservation des données à quelques grands opérateurs chargés de les recopier régulièrement sur les nouveaux supports avant que les anciens ne se dégradent et, éventuellement, de les convertir dans les nouveaux formats. À cet égard, les offres d'hébergement gratuit ne manquent pas, qu'il s'agisse d'abriter nos photographies, vidéos, documents bureautiques ou encore notre courrier. Le plus souvent, elles sont le fait de grands fournisseurs de services sur Internet, tels *Google*, *Amazon*, *Facebook*, *Apple*, *Microsoft*, etc. Toutefois, les récentes révélations sur le programme de surveillance PRISM de l'Agence américaine de sécurité soulignent un phénomène général : nous payons ces services en fournissant involontairement des informations très détaillées sur nos activités, préférences, relations personnelles, achats, voyages, etc.

Ces données sont employées à notre insu, ce qui se traduit non seulement par une surveillance occulte, mais aussi de plus en plus souvent par des violations à grande échelle de la vie privée, par exemple par des vendeurs de toutes sortes de produits.

D'autres grands acteurs ont lancé un gigantesque effort de préservation des données disponibles sur Internet ou sur certains réseaux sociaux : fondée en 1996, l'*Internet Archive* fait régulièrement des copies des sites disponibles sur la Toile et conserve une vaste collection d'images, de vidéos, de livres et de logiciels. S'agissant des œuvres littéraires, musicales ou cinématographiques, de grandes sociétés se réservent le droit exclusif de les archiver, et interdisent au public de les copier ou de les échanger ; en outre, elles mettent parfois en place des mesures de protection contre la copie qui entravent la conservation de ces données à long terme.

Aussi, tous ces acteurs, aussi puissants soient-ils, finissent-ils par constituer des points de fragilité comparables à la grande bibliothèque d'Alexandrie des temps hellénistiques. Malgré les investissements faits pour créer des centrales de serveurs plus sûres et mieux organisées, la concentration de grandes masses de données sous le contrôle d'une seule organisation est toujours source de fragilité. Et au-delà du problème technique de la sauvegarde des données contre les pertes accidentelles, il s'agit aussi de trier les informations qui méritent d'être préservées et celles qui peuvent être effacées. Or le fait qu'une seule organisation, ayant sa logique et ses intérêts particuliers, en décide, introduit manifestement un énorme risque de destruction. Aurions-nous aujourd'hui des éditions complètes initiales des *Fleurs du mal* de Charles Baudelaire (1821-1867) ou de *Madame Bovary* de Gustave Flaubert (1821-1880), si, il y a plus d'un siècle, alors que ces chefs-d'œuvre étaient censurés, leur conservation avait été confiée à une organisation étatique centralisée ?



5. L'AGENCE AMÉRICAINE DE SÉCURITÉ (NSA ou *National Security Agency*) exploite la concentration des données personnelles chez les grands fournisseurs de services sur Internet pour surveiller la population. Dans le projet PRISM, révélé par le lanceur d'alerte Edward Snowden, elle se fonde sur la loi américaine régulant les activités d'espionnage pour se fournir en données à décrypter chez les gros opérateurs, tels *Google*, *Amazon*, *Facebook*, etc.

L'AUTEUR



Roberto DI COSMO, professeur d'informatique à l'Université Paris-Diderot, est membre du Laboratoire Preuves, programmes et systèmes du CNRS.

L'échange entre pairs

Une solution alternative de conservation des données de l'humanité à long terme provient du très grand nombre des ordinateurs personnels et de la croissance exponentielle de la capacité de stockage des mémoires disponibles pour le grand public. Le déploiement d'un système d'échange pair à pair (de l'anglais *peer to peer*) rend envisageable la transformation du parc mondial d'ordinateurs personnels en une gigantesque base de données distribuées, puis de laisser chacun y répliquer et y partager ses données.

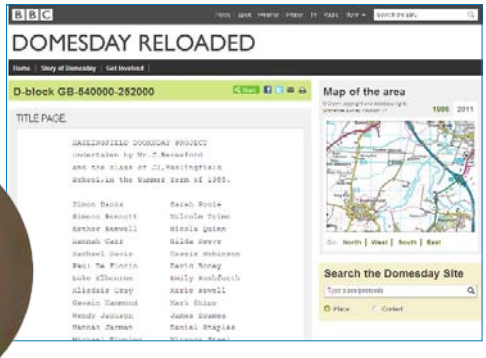
Sous réserve que les ordinateurs personnels continuent à exister (!), cette solution aurait l'avantage de faire l'impasse sur une autorité centrale fixant seule ce qui mérite d'être conservé. Cela permettrait à

LES DONNÉES DU JUGEMENT DERNIER

En 1085, Guillaume le conquérant ordonna un recensement général en Angleterre, qui devint ensuite la base indiscutable pour établir les taxes royales. Les Anglais nommèrent *Domesday Book* (le livre du jugement dernier) le volume où les résultats étaient consignés *(ci-dessous à gauche)*. En 1986, la BBC décida de stocker les données du nouveau *Domesday* – un nouveau recensement et une enquête générale qu'elle avait organisés – sur des vidéodisques de type *LaserDisc*, support que l'on pensait alors pérenne *(au centre)*. Toutefois, il se périma si vite que, 25 ans plus tard, il était déjà devenu impossible d'en lire le contenu ! Un long et coûteux effort technique fut alors nécessaire pour reproduire sur un ordinateur moderne le fonctionnement d'un lecteur de *LaserDisc*, ce qui conduisit au *Domesday reloaded* (*Domesday* refait), une nouvelle version consultable *(à droite)* qui, par prudence, a été versée aux Archives nationales britanniques…







l'humanité de décider spontanément des informations qu'elle souhaite conserver… Les années fastes du téléchargement de musique sur les anciens systèmes pair à pair nous ont donné un aperçu de ce qu'un tel réseau peut accomplir : à côté de tubes et autres succès commerciaux plus ou moins faciles, on y retrouvait souvent des morceaux de musique rares et, sinon, indisponibles. Il suffit en effet qu'existe un petit groupe de passionnés par tel ou tel type de données pour que celles-ci soient préservées quelque part dans le réseau social. Cela constitue un atout en faveur de la conservation de données à long terme, étant donné qu'il est difficile d'identifier quelle information pertinente aujourd'hui le sera demain ou, inversement, quelles données apparemment sans intérêt aujourd'hui pourraient se révéler précieuses dans l'avenir.

Comment résoudre les conflits d'intérêt ?

Encore faut-il que l'accès aux données dans la durée soit garanti par l'usage de formats ouverts et de logiciels libres. Dès lors, il serait utile de mettre en place une infrastructure de recherche répartie, comme les données, partout dans le monde, pour que les informations y soient stockées et indexées.

Le principal obstacle n'est pas technique, mais économique. Nombre d'acteurs des industries de la musique, du cinéma et du livre militent pour qu'on restreigne les échanges entre pairs et pénalise ceux qui y participent ; fondés sur la vente de copies individuelles, leurs modèles économiques s'écroulent si l'on donne la possibilité de réaliser à volonté un nombre illimité de copies numériques de toute œuvre. C'est pourquoi ils souhaitent l'application de lois contraignant les internautes à se détourner du modèle pair à pair pour passer à des échanges cryptés, lesquels ne peuvent que restreindre l'accès public aux données.

Nous ne pouvons détailler ici les propositions qui ont été faites pour concilier ces intérêts contradictoires, mais il nous paraît essentiel de souligner que le libre partage sans entrave est un instrument essentiel pour la préservation de nos données pour les futures générations. Il doit être encouragé et accompagné. De la même façon que les collections privées de livres ont préservé des titres dont les éditeurs ont disparu, les très nombreuses collections de copies numériques individuelles de données nouvelles ou anciennes seront pour nos descendants un moyen de conserver la culture actuelle, pourvu qu'elles soient enregistrées dans des formats ouverts, avec des logiciels libres. ■

■ BIBLIOGRAPHIE

J. Rothenberg, **Avoiding technological quicksand : Finding a viable technical foundation for digital preservation,** rapport au *Council on Library and Information Resources*, janvier 1999 (www.clir.org/pubs/reports/rothenberg).

P. et S. Aigrain, **Sharing : Culture and the Economy in the Internet Age**, Amsterdam University Press, 2012.

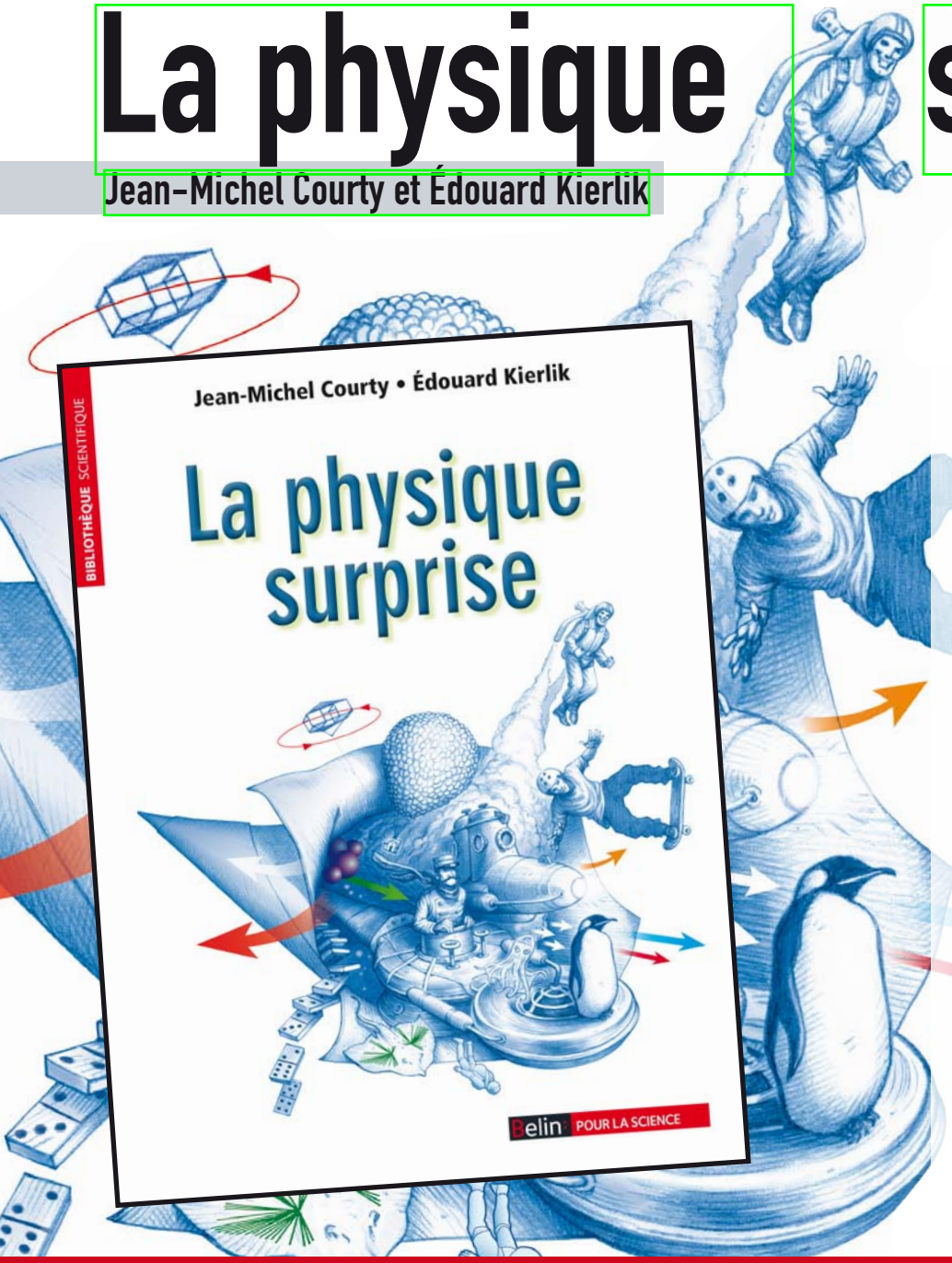
Th. Priol, **Vers le tout-en-réseau ?**, Dossier *Pour la Science*, n° 66, janvier-mars 2010.

Sur le parc de PC prévisible à l'échelle mondiale : www.gartner.com/id=1602818

© Pour la Science - n° 433 - Novembre 2013

La physique surprise

Jean-Michel Courty et Édouard Kierlik



Que nous réserve cette Physique surprise ? Trente-neuf pièces à déguster sans modération, mitonnées par les auteurs. Le style et les croquis, l'actualité et la variété des thèmes vous séduiront, vous amuseront, vous surprendront. Oui, on peut franchir le mur du son en chute libre. Non, les veines ne sont pas bleues. Oui, on peut séparer le chaud et le froid dans un courant d'air. Alors... Bon appétit !

Éditions Belin/Pour la Science 2013
184 pages – 24 euros
ISBN 978-2-8425-5121-9

BON DE COMMANDE

POUR LA SCIENCE

à découper ou à photocopier et à retourner accompagné de votre règlement à : Groupe Pour la Science • 628 avenue du Grain d'Or • 41350 Vineuil • Tél. : 0 805 655 255 • e-mail : pourlasciencevpc@daudin.fr

❑ Oui, je commande l'ouvrage *La physique surprise* de J.-M. Courty et É. Kierlik.

..... exemplaires × 24,00 € (Réf. 075121) = ... €
Frais de port (4,90 € France – 12 € étranger) + ... €
TOTAL À REGLER = ... €

Je choisis mon mode de règlement :
❑ par chèque à l'ordre de *Pour la Science*
❑ par carte bancaire
Numéro
Date d'expiration
Cryptogramme

Signature obligatoire

J'indique mes coordonnées :
Nom :
Prénom :
Adresse :
C.P. : Ville :
Pays : Tél. : (Pour le suivi client facultatif)

E-mail pour recevoir la newsletter *Pour la Science* (à remplir en majuscules) @

En application de l'article 27 de la loi du 6 janvier 1978, les informations ci-dessus sont indispensables au traitement de votre commande. Elles peuvent donner lieu à l'exercice du droit d'accès et de rectification auprès du groupe Pour la Science. **Par notre intermédiaire, vous pouvez être amené à recevoir des propositions d'organismes partenaires. En cas de refus de votre part, merci de cocher la case ci-contre ❑.**

PSPLS433 Offre valable jusqu'au 31/12/13 dans la limite des stocks disponibles.

Les mémoires du FUTUR

Comment stocker et traiter le déluge d'informations produites au quotidien ? Les mémoires numériques actuelles se rapprochant de leurs limites, les scientifiques développent de nouveaux dispositifs.

Luca Perniola

En 2012, l'industrie a fabriqué plus de puces électroniques de mémoire que l'agriculture n'a produit de grains de riz ! Une telle production est indispensable pour stocker et traiter les énormes quantités d'informations créées tous les jours. Dans le monde, $2{,}72 \times 10^{21}$ octets de données ont ainsi été produits en 2012. Les dispositifs actuels sont-ils capables de soutenir la croissance rapide du volume de données ? Cela implique des capacités de stockage importantes et des dispositifs qui permettent d'accéder rapidement aux données. Or les techniques utilisées aujourd'hui semblent proches de leurs limites. Aussi les laboratoires développent-ils une nouvelle génération de mémoires numériques, plus performantes et moins coûteuses.

La nécessité de conserver des informations a débuté très tôt dans l'histoire de l'humanité. Les premiers supports étaient analogiques – parmi les nombreux dispositifs, on trouve les tablettes d'argile, les parchemins ou les livres. Ils avaient pour principal inconvénient d'être encombrants. L'émergence du numérique a bouleversé cet état de fait. Toute donnée – musique, texte, image – est convertie en une séquence de bits (des « 0 » et des « 1 ») qui permet, entre autres avantages, un gain de place important, un traitement automatisé, la consultation à distance, etc.

Le premier exemple de mémoire numérique est celui des cartes perforées utilisées sur les métiers à tisser Jacquard à partir du début du XIX^{e} siècle. Ces cartes présentaient une suite de trous (un trou correspondant, par exemple, à 0 et son absence à 1). Elles contrôlaient le mouvement des crochets du métier à tisser et ainsi automatisaient la fabrication des textiles. Ces cartes perforées se limitaient à indiquer une succession d'instructions.

En 1834, le mathématicien anglais Charles Babbage a eu l'idée d'utiliser les cartes perforées avec sa machine à calculer. Les cartes contenaient les instructions et les données pour les opérations à effectuer. Cette innovation a conduit à l'invention des ordinateurs et au développement des supports de mémoire numériques. Ceux-

L'ESSENTIEL

- Les appareils numériques utilisent des mémoires rapides d'accès pour interagir avec le microprocesseur et des mémoires à grande capacité de stockage, mais plus lentes. Chaque type a ses avantages et ses limites.
- Les mémoires universelles, rapides et capables de stocker de grands volumes de données, n'existent pas.
- Les scientifiques conçoivent de nouveaux dispositifs, fondés sur des effets physiques différents, qui seront rapides d'accès, fiables, économes en énergie...

© Pour la Science - n° 433 - Novembre 2013



© Vasja Kobalev/shutterstock

ci ont connu un essor important au cours des dernières décennies. Ils représentent aujourd'hui l'essentiel des supports de mémoire. Durant ces 20 dernières années, les performances de ces dispositifs ont aussi encouragé la production massive de données *(voir la figure 1)*. La mise en mémoire étant devenue simple et peu chère, il n'y avait plus besoin de limiter le volume de données à stocker.

Les mémoires universelles : utopiques ?

Si l'accroissement du volume d'informations à stocker se poursuit au même rythme, les mémoires d'aujourd'hui seront-elles à la hauteur, en capacité de stockage comme en rapidité d'accès ? Il est probable que non. Les ingénieurs rêvent de mémoires dites universelles qui seraient notamment rapides d'accès, capables de stocker de grands volumes de données, fiables et économes en énergie. Malheureusement, un tel dispositif relève plus de l'idéal vers lequel tendre que d'une réalité concrète.

Les systèmes numériques d'aujourd'hui associent toujours plusieurs dispositifs : des mémoires dites de travail (ou « mémoires vives »), qui conservent de petits volumes de données mais qui ont des temps d'accès très courts, et des mémoires de stockage, qui peuvent enregistrer d'énormes quantités d'information, mais au prix d'un temps d'accès plus long. En associant ces deux types de mémoires, il est possible de construire un système qui peut exécuter des opérations en peu de temps et conserver un grand volume de données.

De nombreux concepts sont à l'étude et certains pourraient être opérationnels dans les années à venir. Pour comprendre l'intérêt de ces dispositifs, il faut commencer par s'intéresser aux mémoires utilisées aujourd'hui, avec leurs qualités et leurs limites. Je présenterai ensuite quelques solutions techniques qui pourraient succéder aux dispositifs actuels d'ici cinq à dix ans.

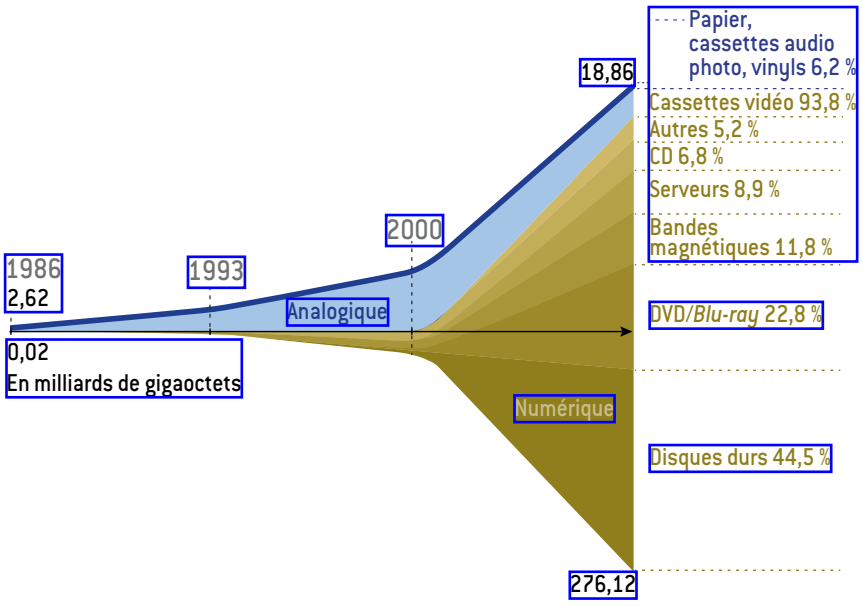
Comme je l'ai évoqué, nous ne savons pas fabriquer de mémoires universelles. La nature ne semble pas faire mieux dans ce domaine. Il existe, par exemple, différents modèles théoriques de la mémoire du cerveau humain. Dans tous les cas, la mémoire est subdivisée en sous-systèmes spécialisés en différentes tâches. On observe ainsi de nombreuses analogies entre la mémoire biologique et les mémoires artificielles. Dans un modèle classique des années 1960, on distingue la mémoire sensorielle qui retient une grande quantité d'informations provenant des différents sens pendant un temps très bref et sans faire d'interprétation. Certaines informations sont ensuite transmises à la mémoire à court terme, dont la capacité est limitée et où les informations sont disponibles pendant plusieurs dizaines de secondes. Cette mémoire permet, entre autres, à un lecteur de ne pas oublier le début de la phrase pour pouvoir en comprendre le sens.

Enfin, une mémoire à long terme permet de conserver des informations longtemps et en quantités importantes. L'évolution a probablement conduit à cette structure qui permet par exemple de traiter rapidement une information perçue, comme quand nous sommes confrontés à un prédateur ou à un autre danger.

Depuis les années 1960, ce modèle de la mémoire humaine a subi des modifications et s'est enrichi de sous-catégories *(voir l'article de Franck Chaillan dans ce numéro)*. Cependant, la conclusion est que le cerveau n'est pas une mémoire universelle et que la combinaison de mémoires spécialisées semble présenter des avantages.

L'industrie suit un schéma similaire, puisqu'elle associe différents types de mémoires. Si, à plusieurs reprises, de nouveaux dispositifs ont été présentés comme universels, des études approfondies les ont toujours mis en défaut. Quoi qu'il en soit, les fabricants se concentrent sur l'amélioration de leurs caractéristiques : la durée de vie, la fiabilité, etc. Une propriété importante est le temps d'accès, la durée nécessaire pour la lecture ou l'écriture d'une information sur une mémoire ; elle est comprise entre la nanoseconde et la milliseconde, suivant les dispositifs.

Le processeur d'un ordinateur, qui exécute des calculs, a besoin d'obtenir très rapidement les instructions et les données nécessaires à l'opération. Il doit donc être associé à des mémoires de travail à temps d'accès très court, pour ne pas ralentir les performances du processeur. De telles mémoires coûtent cher, et leur densité d'information stockée est faible : elles ne peuvent pas emmagasiner de grands



1. LES DONNÉES NUMÉRIQUES ont connu un essor spectaculaire. Le développement de supports numériques bon marché et simples d'utilisation marque un tournant au début des années 2000. Le numérique a supplanté les supports analogiques et le volume total de données enregistrées a augmenté rapidement. Cette tendance ne semble pas faiblir.

volumes de données. Pour archiver les informations, il faut donc ajouter au dispositif des mémoires de stockage.

Les deux types de mémoires sont complémentaires et cohabitent dans tous les dispositifs numériques *(voir la figure 2)*: ordinateurs, tablettes, téléphones portables, clefs USB... On peut définir une hiérarchie pyramidale des mémoires. Au sommet se situent les mémoires de travail, aux temps d'accès très courts et aux capacités de stockage faibles. En descendant dans la pyramide, les volumes d'information qu'il est possible de stocker augmentent au détriment des temps d'accès. La base de la pyramide correspond aux mémoires de stockage.

Les mémoires de travail et de stockage se distinguent aussi par leur capacité à conserver l'information sans alimentation électrique. Les mémoires de travail sont dites volatiles: en l'absence d'une source d'énergie électrique, l'information qui y est enregistrée disparaît en quelques dizaines de millisecondes. Ainsi, quand un ordinateur effectue une opération et qu'une panne de courant survient, toutes les informations stockées dans la mémoire de travail sont perdues. Les mémoires de stockage sont, au contraire, non volatiles: l'information reste en mémoire. Si vous sauvegardez des fichiers sur une clef USB ou sur un disque dur et que vous éteignez l'ordinateur, l'information ne sera pas perdue et pourra être réutilisée ultérieurement.

Examinons le fonctionnement des différentes mémoires. Au plus près du processeur, on distingue deux types de mémoires de travail: les mémoires statiques (ou SRAM pour *Static Random Access Memory*) et les mémoires dynamiques (ou DRAM pour *Dynamic Random Access Memory*). Les premières occupent le sommet de la pyramide. Elles sont directement construites sur la puce de silicium du microprocesseur, car elles sont constituées de transistors, comme ce dernier. Leur temps d'accès est de quelques nanosecondes, mais le volume de stockage est très faible.

Les mémoires dynamiques, juste en dessous dans la pyramide des mémoires, ont une capacité de stockage plus importante; elles sont en général placées sur des puces indépendantes, dont le temps d'accès est compris entre 10 et 100 nanosecondes. Ces mémoires de travail sont utilisées pour effectuer les opérations de calcul des microprocesseurs.

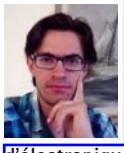
Des mémoires volatiles qui perdent les données

Dans une mémoire dynamique, le bit d'information est enregistré par un système composé d'un transistor et d'un condensateur. Ces composants sont gravés dans le substrat semi-conducteur en silicium de la puce de mémoire. Le transistor agit comme un interrupteur qui autorise, s'il est ouvert, la lecture du bit par un courant électrique. Physiquement, l'état du bit est défini par la charge électrique (l'accumulation d'électrons) du condensateur. Si le condensateur est chargé, le bit vaut 1, s'il est déchargé, le bit vaut 0.

Le problème des condensateurs est qu'ils retiennent mal les électrons, qui fuient dans le silicium environnant. Pour cette raison, la mémoire dynamique est volatile. Il est donc nécessaire de réécrire le bit d'information régulièrement, toutes les quelques dizaines de millisecondes, pour éviter que le condensateur ne se décharge trop et que le bit change d'état. Ainsi, sans alimentation électrique, l'information disparaît de la mémoire.

Dans les premières mémoires dynamiques des années 1960, les transistors et les condensateurs étaient accolés. La miniaturisation a conduit à réduire la taille des composants, mais aussi à changer leur placement relatif. Les condensateurs sont maintenant construits au-dessus des

L'AUTEUR



Luca PERNIOLA dirige le groupe des technologies de mémoires avancées au Laboratoire d'électronique et de technologie de l'information (LETI) au Commissariat à l'énergie atomique et aux énergies alternatives, à Grenoble.

© Pour la Science - n° 433 - Novembre 2013

transistors. De plus, la forme des condensateurs est aujourd'hui cylindrique et très allongée, ce qui réduit la consommation électrique. Avec un rapport de la longueur du condensateur sur la largeur de sa base qui peut atteindre 50, et une densité élevée de condensateurs, les fabricants se trouvent confrontés à plusieurs défis : la production de cavités aussi longues et étroites, et la gestion des problèmes de fuites. La limite de miniaturisation semble atteinte pour ces dispositifs, dont la densité de stockage (quatre gigaoctets) augmente difficilement.

Ainsi, les ingénieurs travaillent en priorité sur deux aspects relatifs aux mémoires dynamiques : l'augmentation de la densité d'information et l'économie d'énergie, cette dernière étant obtenue en supprimant la volatilité.

Les « mémoires flash » sont un type de mémoire non volatile. Mais elles ne peuvent pas remplacer les mémoires dynamiques. Situées plus bas dans la pyramide des mémoires, les mémoires flash ont un temps d'accès d'une dizaine de microsecondes – soit deux ordres de grandeur de plus que les mémoires dynamiques. Dans les mémoires flash, chaque bit est représenté par l'état ouvert ou fermé d'un transistor. Un réservoir au sein du transistor peut accumuler une certaine quantité d'électrons qui créent un champ électrique. Celui-ci laisse passer ou bloque le courant qui traverse le transistor.

Les mémoires flash sont non volatiles, car les électrons restent bloqués dans le réservoir même en l'absence d'une source extérieure (qui, dans les mémoires dynamiques, rafraîchit les informations). Le blocage est dû au fait que le réservoir d'électrons est intercalé entre deux couches d'oxydes qui l'isolent du substrat environnant. Ainsi, une information archivée n'est pas altérée.

Depuis leur développement dans les années 1980, les mémoires flash ont été miniaturisées. Elles atteignent actuellement des capacités de stockage de 128 gigaoctets. Cependant, pour conserver les propriétés de non-volatilité, les couches d'oxydes ne peuvent pas être soumises à la même miniaturisation, qui affaiblirait le blocage des électrons. En outre, comme pour les mémoires dynamiques, la miniaturisation augmente le risque d'erreurs d'écriture et de lecture.

Aujourd'hui, pour augmenter la quantité d'information qu'il est possible d'enregistrer sur une plaque de silicium, les fabricants créent plusieurs couches de mémoire empilées les unes sur les autres. La capacité de mémoire est multipliée approximativement par le nombre de couches.

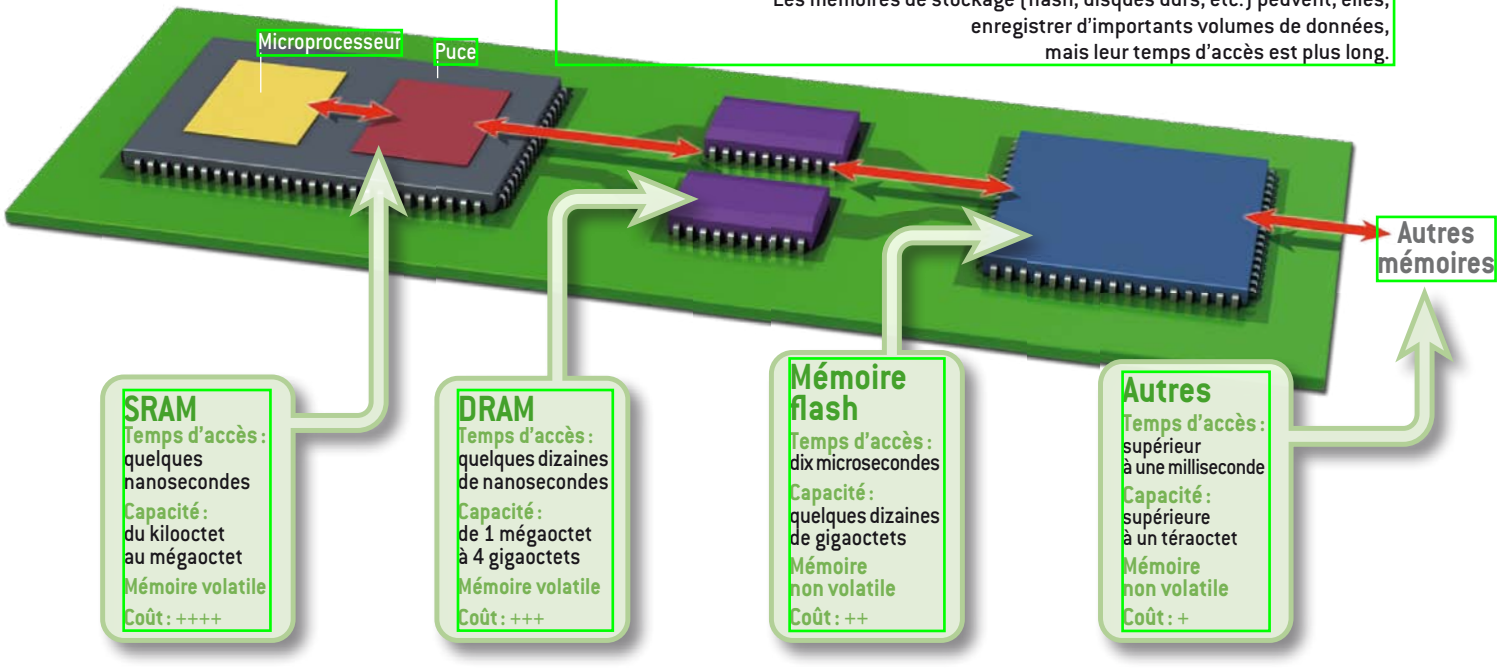
Même si les mémoires flash semblent aussi atteindre des limites de miniaturisation, elles pourraient remplacer une partie des « disques durs ». Ces supports numériques ont été développés dans les années 1950 par IBM. L'information est enregistrée sur des couches de matériau ferromagnétique déposé sur des disques rotatifs. Un élément, la tête de lecture, monté sur un bras mécanique, lit l'information sur le disque à partir de l'état magnétique local du bit. Il peut inscrire un bit grâce à des effets magnétiques réversibles.

Disques durs : volumineux, mais lents

Les disques durs occupent le bas de la pyramide des mémoires. Leur densité d'information est supérieure à celle des mémoires flash et leur prix par unité de mémoire stockée est plus faible. En revanche, le temps d'accès est de l'ordre de la milliseconde, car la tête de lecture doit d'abord être positionnée au bon endroit par un déplacement du bras mécanique et par une rotation du disque.

Dans les années à venir, les disques durs de certains appareils pourraient être remplacés par des dispositifs contenant des mémoires flash, les SSD, ou *Solid State*

2. UN SYSTÈME NUMÉRIQUE associe toujours des mémoires de travail et des mémoires de stockage. Les premières (mémoires statiques et mémoires dynamiques) sont placées au plus près du microprocesseur, chargé de faire les calculs. Elles échangent rapidement les informations avec le processeur, mais ne présentent que de petits volumes de stockage. Les mémoires de stockage (flash, disques durs, etc.) peuvent, elles, enregistrer d'importants volumes de données, mais leur temps d'accès est plus long.



Sauf mention contraire, les illustrations sont de Bruno Bourgeois

Drive. Ces derniers ont plusieurs avantages. Les ordinateurs équipés de ce type de mémoires sont beaucoup plus rapides. Et l'absence d'éléments mécaniques les rend plus robustes et moins sujets aux pannes.

Du flash pour les nomades

Les habitudes nomades des utilisateurs de tablettes, d'ordinateurs ou de téléphones portables favorisent l'utilisation de systèmes résistants aux chocs. Toutefois, les disques durs ne disparaîtront pas dans un futur proche, leur coût restant dix fois inférieur à celui des mémoires flash.

Les mémoires dynamiques, les mémoires flash et les disques durs semblent atteindre leurs limites. Il est donc crucial de développer de nouveaux dispositifs pour les remplacer. Mais les mémoires à venir ne pourront peut-être pas prétendre au statut de dispositifs universels. Les scientifiques continuent de développer des mémoires qui auront les caractéristiques soit de mémoire de travail, soit de mémoire de stockage, en améliorant leurs capacités de stockage, la vitesse de leur lecture ou de leur écriture, ou leur consommation d'énergie.

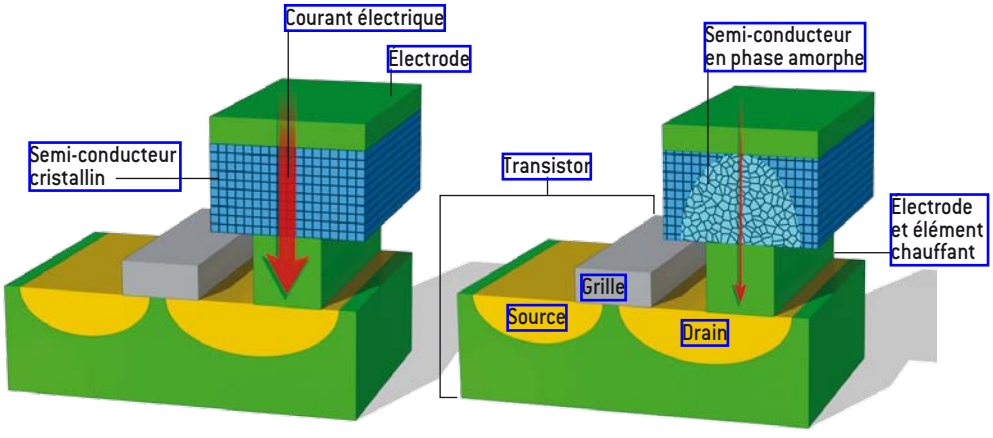
Ce dernier point est particulièrement étudié. Les mémoires non volatiles ont ici un réel avantage et devraient prendre une part de marché plus importante. Il est cependant difficile d'envisager un système uniquement composé de mémoires non volatiles. La mémoire la plus proche du microprocesseur est sollicitée en permanence et requiert des accès ultrarapides. La mémoire statique reste, dans ce cas, la plus efficace et la plus économe en énergie. Il n'y a donc pas besoin de remplacer les mémoires statiques par des mémoires non volatiles (pour être exact, on distingue trois types de mémoires statiques, dont l'un fait l'objet de recherches pour le substituer par un système plus performant).

De nouveaux dispositifs sont pressentis pour remplacer les mémoires dynamiques et pour remplir l'espace de performance entre les mémoires dynamiques et les mémoires flash. Dans ces concepts, que nous allons présenter, le condensateur des mémoires dynamiques est remplacé par une résistance électrique qui peut varier, le bit correspondant valant 0 ou 1 selon que la résistance est faible ou élevée. Il n'y a pas de problème de fuite de charges électriques dans ces dispositifs ; dès lors, ces mémoires sont non volatiles.

Pour gérer l'énorme flux d'informations à traiter et à archiver, de nouveaux dispositifs sont en cours de développement. Ils pourraient être bientôt sur le marché et remplacer les systèmes actuels. Les mémoires dynamiques sont des mémoires volatiles, qui nécessitent d'être réécrites en permanence. De ce fait, elles sont peu économes en énergie. Les mémoires à changement de phase et les mémoires magnétiques pourraient leur succéder en remplaçant les condensateurs, à l'origine de la volatilité, par des résistances.

Mémoires à changement de phase

Les mémoires à changement de phase (PCRAM ou *Phase Change Random Access Memory*) sont constituées d'un transistor et d'une résistance. Celle-ci se compose de deux électrodes séparées par un semi-conducteur, tel l'alliage germanium-antimoine-tellure, aux propriétés particulières. La structure de ce dernier peut être cristalline ou amorphe. Dans le premier cas, la circulation des électrons du courant électrique est aisée : la résistance est faible et le bit vaut 1 *(à gauche)*. Avec une phase amorphe, la circulation des électrons est plus difficile, la résistance est supérieure *(à droite)*. Cela correspond à un bit égal à 0.



Pour changer la valeur du bit, il faut changer la structure du semi-conducteur par chauffage. Un alliage résistif sert de source de chaleur. Il convertit un courant électrique en chaleur par effet Joule. Avec un chauffage bref, on obtient une structure amorphe. Si le chauffage est soutenu, une structure cristalline se met en place. Pour effectuer ces transitions réversibles, des températures de 500 °C à 800 °C sont nécessaires, ce qui implique des courants assez importants dans les dispositifs. La différence de résistance des deux états est bien marquée (rapport supérieur à 100), ce qui permet d'avoir un temps de lecture assez court, car il n'y a pas d'ambiguïté entre les deux états.

Une telle amplitude de résistance permettrait d'utiliser un tel dispositif pour plusieurs bits. Dans ce cas, la plage de résistance n'est pas divisée en deux intervalles correspondant aux bits 0 et 1, mais en davantage d'intervalles, ce qui permet de coder deux bits (00, 01, 10 ou 11). Ce fonctionnement est délicat à appliquer à cause de phénomènes de relaxation. Le matériau subit des chocs thermiques durant les étapes d'écriture et d'effacement. Entre ces étapes, une relaxation structurelle conduit à une variation stable de la résistance dans le temps, qui empêche une bonne utilisation des mémoires à changement de phase en mode multiniveaux.

En outre, ces mémoires ont une grande stabilité quand elles sont utilisées à haute température. Elles peuvent être utilisées dans certaines applications – automobiles, cartes à puces… – où les contraintes de température sont fortes.

© Pour la Science - n° 433 - Novembre 2013

Il existe toute une variété de mémoires à résistance variable, selon les principes physiques qui créent la résistance. Présentons ici deux successeurs possibles des mémoires dynamiques : les mémoires dites magnétiques et les mémoires dites à changement de phase. Ces mémoires consomment moins d'énergie et présentent, en principe, une densité d'information équivalant à celle des mémoires dynamiques.

Les mémoires magnétiques (MRAM pour *Magnetic Random Access Memory*) existent depuis les années 1950. Leur utilisation était cependant limitée à des niches, tel le domaine aérospatial. Ces mémoires résistent mieux aux effets des rayons cosmiques, qui peuvent induire des erreurs d'écriture des bits dans les autres dispositifs de stockage. Depuis dix ans, les mémoires magnétiques ont connu de nombreuses innovations qui ont relancé leur intérêt.

MRAM et PCRAM, des pistes sérieuses

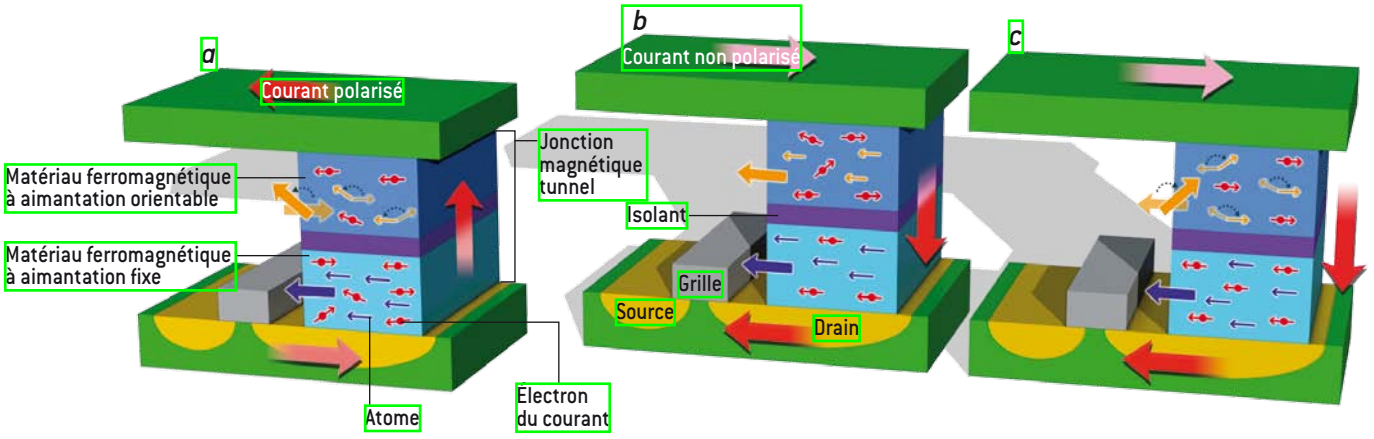
Il existe aujourd'hui des mémoires magnétiques ayant une capacité de stockage de 64 mégaoctets. La résistance qui code le bit est contrôlée par l'aimantation des éléments dont est formé le composant *(voir l'encadré ci-dessous)*. Le temps d'écriture est de l'ordre de la nanoseconde, mais le temps nécessaire pour lire une information est plus long. En effet, la différence de résistance entre les états « 0 » et « 1 » n'étant pas très marquée (le rapport des résistances est inférieur à dix), il faut plus de temps pour lire correctement le bit d'information.

Le deuxième dispositif pour remplacer la mémoire dynamique est la mémoire à changement de phase (PCRAM ou *Phase Change Random Access Memory*). Il a été développé à la fin des années 1960 par l'inventeur américain Stanford Ovshinsky (un autodidacte qui a déposé plus de 400 brevets en 50 ans et dont un grand nombre sont exploités industriellement). Dans cette mémoire, la résistance est déterminée par la structure atomique d'un semi-conducteur. Ce dernier peut être cristallin ou amorphe, et devient ainsi conducteur ou isolant *(voir l'encadré ci-dessous)*. Un élément chauffant permet de changer de façon réversible la structure du semi-conducteur. Ce type de mémoire demande donc plus d'énergie pour écrire ou effacer un bit que dans le cas de la mémoire magnétique.

Mémoires magnétiques

Pour remplacer les mémoires dynamiques et dépasser leurs limites, une des pistes les plus intéressantes est celle des mémoires magnétiques (MRAM pour *Magnetic Random Access Memory*). Dans l'exemple de la mémoire magnétique à transfert de spin, le bit de mémoire est constitué d'un transistor et d'une résistance. Cette dernière est constituée d'une superposition de deux films minces de matériau ferromagnétique séparés par une barrière isolante qui ne peut être franchie par des électrons que par effet tunnel – on parle de jonction magnétique tunnel. L'information est portée par la résistance de la jonction : on a « 0 » si la résistance est importante, « 1 » si la résistance est faible. L'aimantation des deux couches ferromagnétiques contrôle cette résistance. L'aimantation de la couche inférieure est fixée. Dans l'autre, il est possible de l'orienter grâce à un courant dont le spin des électrons – leur moment cinétique intrinsèque – est polarisé : tous les spins sont dirigés dans le même sens. Ce spin permet d'orienter l'aimantation du matériau. Le schéma *(a)* indique le processus d'écriture et le passage d'un bit 0 à un bit 1. Un flux d'électrons *(en rouge)* non polarisé arrive par le bas. Lors du passage de la barrière de potentiel, le transfert des électrons de spin dirigé vers la gauche est facilité. Ainsi, dans le matériau ferromagnétique, le flux d'électrons est polarisé. Le spin des électrons interagit avec les spins des atomes du matériau *(en jaune)* et les réoriente de la droite vers la gauche. L'aimantation des deux matériaux ferromagnétiques *(violet et jaune)* est alors dans le même sens (parallèle). La résistance est alors faible. Un courant plus faible peut être injecté dans le dispositif pour mesurer cette résistance et signaler que l'état correspond à un bit 1.

Pour faire passer l'état d'un bit de 1 à 0, on inverse le sens du courant de polarisation. En arrivant par le haut *(b)*, il n'est pas polarisé, mais seuls les électrons dirigés vers la gauche traversent la barrière. Il ne reste que des électrons de spin dirigé vers la droite dans le milieu ferromagnétique du haut *(c)*, ce qui entraîne un basculement de l'aimantation de la gauche vers la droite. Les aimantations sont antiparallèles. La résistance est grande, ce qui correspond à un bit 0.

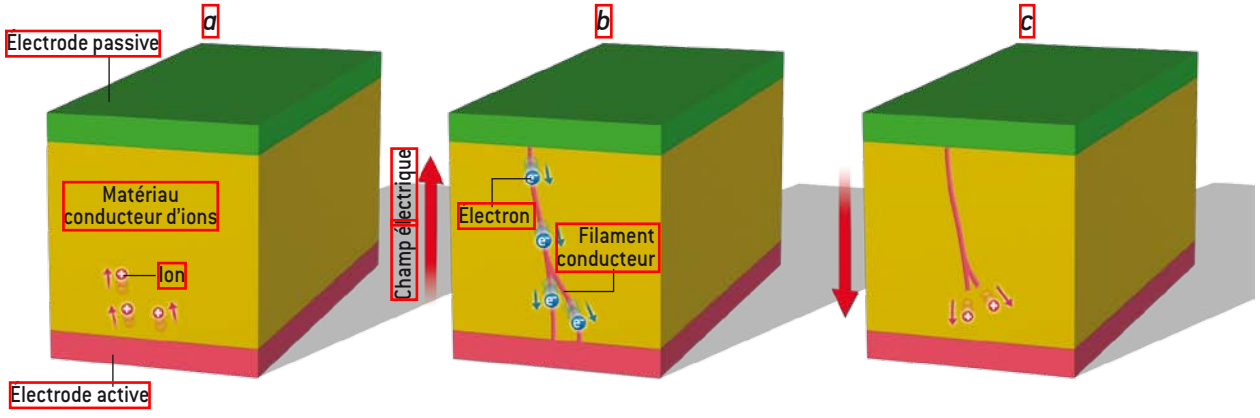


LES MÉMOIRES À OXYDES ET À PONT CONDUCTEUR

Les mémoires flash pourraient être remplacées par des dispositifs fondés sur les mémoires à oxydes (OxRAM) et les mémoires à pont conducteur (CBRAM). Dans les deux dispositifs, le bit d'information est porté par un transistor couplé à une résistance. Cette dernière est composée de deux électrodes séparées par un oxyde isolant ou un solide perméable à la conduction des ions. L'information est codée par la résistance du dispositif, qui dépend de la présence ou non d'un filament conducteur entre les électrodes. Ce filament est créé ou détruit par un stimulus électrique. Le phénomène électrique de claquage de l'oxyde (correspondant à la formation du filament conducteur) a été étudié depuis les années 1960, mais l'industrie n'a commencé à s'y intéresser qu'à partir de 2004. En général, pour les mémoires à oxydes, le substrat est « vierge » lors de la première utilisation : il faut appliquer une tension importante pour amorcer la formation du filament. La tension d'utilisation par la suite sera beaucoup plus faible, car seul un petit élément du filament est construit ou détruit.

La physique des mémoires à pont conducteur est mieux établie que celle des mémoires à oxydes. La nature du pont dépend de la composition de l'électrode active (cuivre, argent…). Le filament se forme à partir de la dissolution de l'électrode active, qui émet des cations se propageant vers l'électrode inerte grâce au champ électrique appliqué *(a)*. Les cations sont réduits et forment un filament d'atomes neutres qui relie les électrodes *(b)*. La résistance du dispositif chute (et fait passer le bit de l'état 0 à 1). Le processus est réversible : il faut inverser la polarité du champ électrique pour que les cations retournent à leur électrode initiale *(c)*.

De nombreuses études cherchent à optimiser les performances de ce dispositif en associant divers matériaux pour les électrodes, aussi bien dans les mémoires à oxydes que dans celles à pont conducteur. De plus, la compréhension et la description microscopique des effets impliqués pourraient nécessiter plusieurs années.



SI LES DISQUES DURS COMMENCENT déjà à disparaître dans les appareils nomades, les mémoires dynamiques pourraient être remplacées d'ici trois à cinq ans.

Les mémoires à changement de phase atteignent aujourd'hui une capacité de un gigaoctet et les temps d'écriture sont de l'ordre de plusieurs dizaines de nanosecondes. De ce point de vue, ces dispositifs sont moins performants que les mémoires magnétiques. Mais la différence de résistance des deux états étant plus marquée (d'un facteur supérieur à 100), le temps de lecture est plus court. Par ailleurs, ces mémoires sont robustes et fonctionnent à haute température.

Le paysage des mémoires de stockage pourrait aussi changer d'ici cinq à dix ans. Comme nous l'avons vu, les disques durs sont les mémoires les moins chères, mais ils sont trop fragiles pour des utilisations nomades. Bien qu'ils soient déjà en partie remplacés par des dispositifs contenant des mémoires flash, de nouvelles mémoires de stockage se profilent. Si leurs caractéristiques deviennent intéressantes, elles pourront remplacer les mémoires flash dans quelques années. Deux pistes sont prometteuses parmi les nombreuses solutions étudiées par les industriels : les mémoires à oxydes (OxRAM ou *Oxide-based Random Access Memory*) et les mémoires à pont conducteur (CBRAM ou *Conductive Bridge Random Access Memory, voir l'encadré ci-dessus*). Ces dispositifs sont récents, mais les recherches sont très actives et le nombre de prototypes augmente rapidement. Il n'est pas impossible que ces concepts de mémoires remplacent aussi les mémoires dynamiques. En 2013, une mémoire à oxydes de 32 gigaoctets a été annoncée par les Sociétés *Sandisk* et *Toshiba*.

Comme pour les mémoires magnétiques et à changement de phase, ces dispositifs sont constitués d'un transistor et d'une résistance variable. La valeur de la résistance détermine si le bit vaut 0 ou 1. Elle est contrôlée par un filament conducteur qui s'établit entre deux électrodes. Celui-ci peut être créé ou détruit par un champ et un courant électriques. Par exemple, dans une mémoire à pont conducteur, le filament métallique se

© Pour la Science - n° 433 - Novembre 2013

forme à partir d'ions positifs qui sont extraits d'une des électrodes. Quand la chaîne métallique relie les deux électrodes, le courant d'électrons peut facilement circuler et la résistance est faible. Le bit d'information vaut alors 1. En inversant le champ électrique, les atomes redeviennent des ions et migrent vers l'électrode : la chaîne se brise, la résistance augmente et le bit d'information passe de 1 à 0. Le processus est réversible.

Les progrès sont si rapides dans ce domaine qu'il est difficile de donner les caractéristiques de ces mémoires. Le temps d'écriture varie de la nanoseconde à la milliseconde suivant les dispositifs. Par ailleurs, les techniques de fabrication, surtout pour les mémoires à oxydes, sont relativement courantes, et leur coût de production est très compétitif. En revanche, les phénomènes microscopiques mis en jeu restent à mieux comprendre.

Tous les dispositifs présentés n'en sont pas au même point de maturité et leur présence sur le marché peut prendre encore quelques années. Si les disques durs commencent déjà à disparaître dans les appareils nomades, les mémoires dynamiques pourraient n'être remplacées que d'ici trois à cinq ans.

Un horizon encore flou

La demande de mémoires toujours plus efficaces va continuer de croître. Cela encouragera les universités, les centres de recherche et les constructeurs à explorer les pistes évoquées, ainsi que de nombreuses autres. La concurrence étant forte et favorable à l'inventivité, l'évolution peut être rapide, et il est difficile de voir clairement quels composants s'imposeront dans cinq, dix ou vingt ans. La tendance qui se dégage est l'abandon de l'idée d'une technologie prédominante et universelle qui remplacera toutes les mémoires existantes. La solution optimale est de faire cohabiter plusieurs techniques et d'améliorer les performances de chacune. Jusqu'où ? On l'ignore. ■

BIBLIOGRAPHIE

H.-S. Wong *et al.*, **Metal oxide RRAM**, *Proceeedings of the IEEE*, vol. 100, n° 6, pp. 1951-1979, 2012.

T. Kawahara *et al.*, **Spin-transfer torque RAM technology : Review and prospects**, *Microelectronics Reliability*, vol. 52, pp. 613-627, 2012.

M. Hilbert et P. Lopez, **The wolrd's technological capacity to store, communicate and compute information**, *Science*, vol. 332, pp. 60-65, 2011.

G. W. Burr *et al.*, **Phase-change memory technology**, *J. Vac. Sci. Technol. B*, vol. 28, n° 2, pp. 223-262, 2010.



Stocker de l'information dans l'ADN

Christophe Dessimoz

Grâce à l'avancée des techniques de la génomique, il devient possible d'utiliser l'ADN comme support numérique pour archiver des données sur le long terme.

Deux piscines olympiques : tel est le volume qu'il faudrait pour contenir l'ensemble des données produites dans le monde en 2012 (environ 3×10^{21} octets), si on les stockait dans les toutes récentes clefs USB de un téraoctet (10^{12} octets). Au fil des ans, ce volume croît de façon exponentielle et, avec lui, le besoin de supports numériques fiables pour archiver l'ensemble des informations sur le long terme. De plus, la durée de vie des supports actuels étant limitée (pas plus de quelques dizaines d'années), la pérennisation des données nécessite une maintenance constante et lourde, fondée sur des copies multiples et régulières sur divers supports *(voir l'article de Francis Daumas et Marion Massol dans ce numéro)*.

Pourtant, il existe depuis plus de trois milliards d'années un support numérique universel, compact et stable : l'ADN, la molécule porteuse de l'information génétique des organismes vivants. Codé sous la forme d'une succession de quatre molécules – les nucléotides A, T, G, C –, le programme de la vie est numérique : les nucléotides codent de l'information, au même titre que les bits (0 ou 1) de l'informatique. L'ADN est de plus très stable – on a retrouvé, dans des carottes de glace, des gènes datant de 450 000 à 800 000 ans – et dense : si l'on codait sur de l'ADN toute l'information produite en 2012, elle tiendrait dans le coffre d'une voiture...

En outre, le fait que l'ADN soit universel limite les risques d'obsolescence. Tant qu'il y aura de la vie sur Terre et un intérêt pour la comprendre, la technique pour lire l'ADN existera. Aussi les progrès récents de la génomique nous ont-ils conduits, à l'Institut européen de bio-informatique, à étudier la possibilité d'utiliser l'ADN comme support d'archives numériques.

Tous les sonnets de Shakespeare sur ADN

Peu après la caractérisation de l'ADN, publiée en 1953 par James Watson et Francis Crick, l'idée d'utiliser cette molécule comme support d'information a été envisagée. En 1964, un physicien et officier radio russe, Mikhail Neiman, voit dans les systèmes et processus d'information biophysiques un moyen de miniaturiser les dispositifs de stockage et de traitement de l'information. L'idée a été mise en œuvre à la fin des années 1990, avec le développement du génie génétique. Plusieurs équipes ont su encoder de l'information sur l'ADN, mais les informations étaient courtes (un message secret annonçant le débarquement du 6 juin 1945, une comptine anglaise avec paroles, musique et petite image, la Déclaration des droits de l'homme…). Surtout, les manipulations génétiques étaient laborieuses et *ad hoc*.

Ces dernières années, cependant, les progrès des techniques de séquençage ont entraîné un tournant important. En 2010, le biologiste américain Craig Venter, de

L'AUTEUR



Christophe DESSIMOZ est professeur de bio-informatique au Collège universitaire de Londres. Il est aussi affilié à l'Institut européen de bio-informatique (EMBL-EBI). Article écrit en collaboration avec Marie-Neige Cordonnier.

BIBLIOGRAPHIE

N. Goldman *et al.*, **Towards practical, high-capacity, low-maintenance information storage in synthesized DNA**, *Nature*, vol. 494, pp. 77-80, 2013.

G. M. Church *et al.*, **Next-generation digital information storage in DNA**, *Science*, vol. 337, pp. 1628-1629, 2012.

Entretien avec Ch. Dessimoz à écouter sur : bit.ly/pls 433 dessimoz

© Pour la Science - n° 433 - Novembre 2013

l'Institut du même nom, et son équipe ont synthétisé un génome complet d'un million de paires de bases (ou nucléotides) à partir du code du génome d'une bactérie, qu'ils ont introduit dans une autre espèce de bactérie dont ils avaient ôté le génome. Si l'opération était à nouveau *ad hoc* et fort coûteuse (estimée à plusieurs dizaines de millions d'euros), elle montrait qu'il était désormais possible de coder une grande quantité d'information sur de l'ADN en un temps raisonnable.

Depuis, le coût du séquençage a baissé de façon drastique, ce qui ouvre encore de nouvelles perspectives. Encoder de l'information sur l'ADN n'est pas compliqué. Il suffit de passer d'un système binaire (0, 1) à un système codé en base 4 en convertissant chaque couple binaire – 00, 01, 10, 11 – en un des quatre nucléotides A, T, G, C. Toutefois, deux aspects fondamentaux sont à prendre en compte. D'une part, coder une longue séquence coûte cher. Nous avons donc pris le parti de synthétiser des fragments courts (150 paires de bases), meilleur marché.

D'autre part, il s'agit de limiter au maximum les risques d'erreurs. Les erreurs de lecture des séquences se concentrent là où se suivent plusieurs nucléotides identiques. Pour éviter ces répétitions, nous avons converti les données binaires en base 3 et non en base 4, et utilisé le quatrième nucléotide comme une sorte d'opérateur de permutation. Le code en base 3 est composé de 0, 1, 2, auxquels on fait correspondre un nucléotide A, T, G ou C. Toute répétition est évitée en modifiant la correspondance chiffres-nucléotides après chaque chiffre codé. Par exemple, si l'on veut coder 22 et que le premier 2 est codé par A, le second 2 sera codé par une nouvelle correspondance ne faisant intervenir que les trois autres nucléotides (T, G ou C).

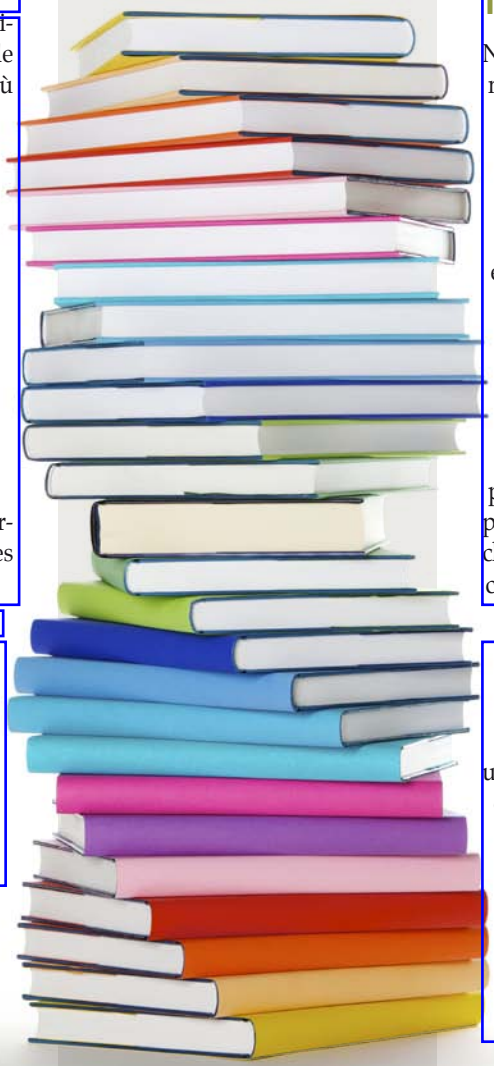
Enfin, pour limiter encore les erreurs, nous avons synthétisé des fragments d'ADN dont les séquences se chevauchent, de façon que chaque nucléotide soit couvert par quatre fragments différents, et nous avons ajouté, au début et à la fin de chaque fragment, des informations indiquant sa position dans la séquence complète.

Nous avons testé notre méthode sur cinq fichiers différents : un fichier texte en code ASCII comportant tous les sonnets de Shakespeare, le PDF d'un article scientifique, un extrait sonore du discours de Martin Luther King (format mp3), une photo (format JPEG 2000) et le texte (ASCII) du code qui nous sert à convertir les données binaires en base 3, pour montrer que l'on peut encoder le codage lui-même…

En deux jours, les données ont été encodées sur 150 000 brins d'ADN. L'ADN a ensuite été lyophilisé et envoyé par avion en Allemagne, où il a été réhydraté et séquencé en deux semaines. Nous avons retrouvé l'intégralité des données, sauf celles contenues dans deux fragments de 25 nucléotides. Ces deux fragments appartenaient à une région répétitive, ce qui nous a permis de les reconstituer manuellement et d'atteindre 100 pour cent de réussite (contre 99,999 pour cent sans ces deux fragments). Les deux fragments avaient chacun la particularité d'avoir des séquences leur permettant de se replier sur eux-même, ce qui a empêché leur séquençage.

Où stocker l'ADN synthétisé ?

Un endroit frais, sombre et sec suffirait à conserver de l'ADN lyophilisé pendant plusieurs milliers d'années. Une telle banque existe déjà : inaugurée en 2008, la Réserve mondiale de semences du Svalbard, une immense enceinte réfrigérée creusée dans l'île norvégienne du Spitzberg pour préserver la diversité des semences, n'a pas de personnel permanent sur place. Une cave à vin conviendrait aussi…



Un intérêt pour le stockage à long terme

Nous avons ainsi réussi à stocker de l'information sur de l'ADN et à la restituer selon un procédé déclinable pour toute donnée numérique. Bien sûr, notre technique a ses limites. Avant tout, même s'ils seront sans doute réduits à quelques heures à l'avenir, les temps d'accès, d'écriture et de lecture sont moins bons que ceux d'une clef USB. L'ADN ne peut constituer un support compétitif que pour l'archivage de données à long terme. De plus, le volume de données testé était assez petit – 739 kilooctets, l'équivalent d'une disquette –, mais en théorie, il est possible de passer à une échelle beaucoup plus grande. De même, le procédé reste cher – 7 000 euros par mégaoctet –, mais ce coût diminuera vite si les tendances actuelles se poursuivent.

La technique d'encodage peut encore être améliorée et optimisée. Notamment, le passage à des données plus importantes nécessitera une automatisation et une miniaturisation plus poussées. D'ici quelques années, pour des utilisations sur le long terme, il pourrait devenir plus économique d'investir dans un support certes plus cher à l'achat, mais avec très peu de frais de maintenance, l'ADN, que d'opter pour un support bon marché tel que les bandes magnétiques, mais qui doit être relu tous les cinq ans. ■

© Shutterstock / Lonmadeta

L'ADN

mémoire numérique du vivant

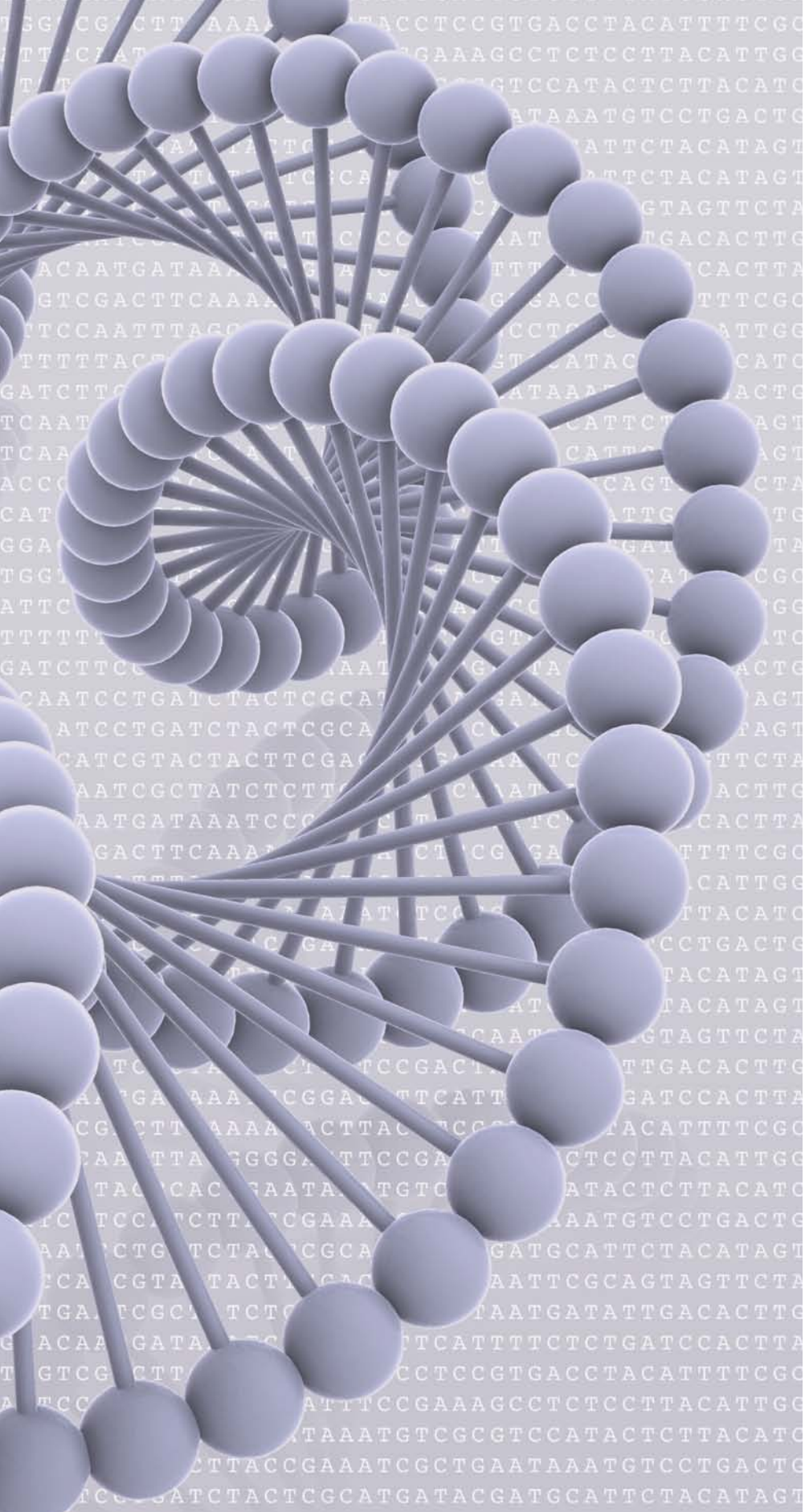
Vincent Daubin, Simon Penel et Éric Tannier

L'ADN, un code à quatre lettres, est une source d'information sur l'histoire du monde vivant. En comparant les génomes actuels, on reconstitue les gènes d'espèces disparues il y a plusieurs centaines de millions d'années.

L'ESSENTIEL

- **Par les mécanismes** de l'hérédité, les génomes actuels recèlent des informations sur l'évolution du vivant depuis plusieurs milliards d'années.
- **L'information** est brouillée par le jeu des mutations, de la sélection naturelle et du transfert de gènes entre espèces.
- **Néanmoins,** en modélisant l'histoire des gènes, les phylogénéticiens ont accès à des informations sur des espèces dont on n'a plus trace aujourd'hui.

© Pour la Science - n° 433 - Novembre 2013



Depuis plus de trois milliards d'années, des informations numériques écrites en langage ADN sont copiées, plus ou moins fidèlement, puis interprétées, conduisant à la formation de nouvelles générations d'organismes. À ce titre, la théorie de l'évolution est une théorie de l'information, analogue à celle qu'on utilise en mathématiques depuis sa formalisation en 1948 par le mathématicien américain Claude Shannon : un ensemble de lois décrivant comment un objet – un texte dans le cas de la théorie mathématique, un organisme dans le cas de la reproduction – peut être reconstitué à partir d'une information sur cet objet, en général codée de façon plus compacte que l'objet lui-même. La biodiversité actuelle résulte de ce processus de duplication – ou réplication – avec erreurs des données génétiques héritées d'une multitude d'ancêtres.

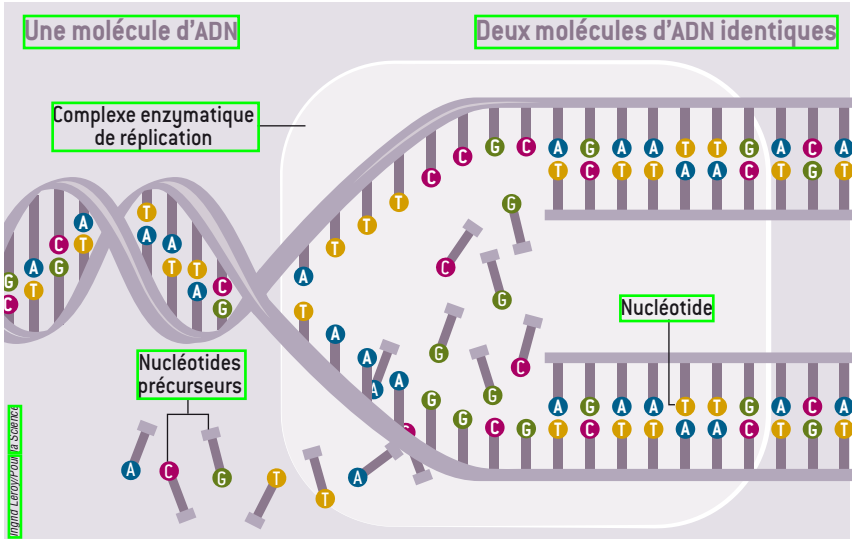
Dans les génomes des organismes actuels figurent les traces de ces milliards d'années de transmission d'informations codées génétiquement. Aussi, chaque lignée ayant enregistré des événements différents depuis sa séparation des autres, il est possible, en comparant ces génomes, de reconstruire l'histoire du vivant. Depuis 50 ans, les phylogénéticiens, qui étudient les relations évolutives entre les organismes, comparent les séquences ADN des espèces actuelles. Avec les progrès du séquençage génomique, ces recherches ont changé d'échelle. Elles permettent aujourd'hui de comprendre à la fois les mécanismes de l'évolution et l'origine des espèces. Quelles informations anciennes peut-on espérer décrypter dans les génomes actuels ? Quelles sont les méthodes employées ? Quelles sont les limites de cette approche ? L'analogie avec la théorie de l'information apporte plusieurs réponses.

L'hérédité : une théorie de l'information

Si, en semant des graines de tomate, on obtient des tomates et non des courgettes, c'est parce que dans les graines figure une information qui dicte la forme de la plante. La nature de cette information, celle qui est transmise des parents à leur progéniture dans tout le monde vivant lors de la reproduction, a été longuement débattue. Selon la théorie de la préformation, par exemple, dont on trouve des traces dès l'Antiquité, chaque graine contenait

© Shutterstock / ateena3

© Pour la Science - n° 433 - Novembre 2013



1. LE MÉCANISME qui confère sa pérennité à l'information codée dans le génome est la réplication de l'ADN, orchestrée par un complexe protéique, le complexe enzymatique de réplication. Les deux brins de la molécule d'ADN sont séparés et, sur chaque brin, sont recrutés les nucléotides complémentaires de la séquence lue par le complexe (A et C sont respectivement les complémentaires de T et G). On obtient ainsi deux nouvelles molécules d'ADN identiques.

l'intégralité du futur organisme, déjà formé, mais en plus petit – une sorte de modèle réduit de l'organisme. Mais pour la théorie de l'épigenèse, sa concurrente depuis l'Antiquité jusqu'au XVIII[e] siècle, les graines constituaient un plan de l'organisme, que le développement allait réaliser selon des lois d'organisation qui restaient à découvrir. L'épigenèse était ainsi déjà presque une théorie de l'information au sens de Shannon.

Au début du XX[e] siècle, la génétique a donné à l'hérédité son caractère numérique : le gène – qui n'était pas encore identifié comme un segment d'ADN – était un objet abstrait que l'on pouvait observer dans un nombre fini d'états disjoints, les allèles, codant les caractéristiques héréditaires des organismes. La découverte, en 1953, de la structure moléculaire du support de l'information héréditaire a clarifié le concept de gène : le plan des organismes contenu dans les graines est un code numérique porté par une molécule, l'ADN.

Cette molécule est un assemblage linéaire de petites molécules – des nucléotides – de quatre types : adénine (A), guanine (G), cytosine (C) et thymine (T). Cet assemblage porte une information numérique : comme les mots dans une écriture alphabétique ou les séquences de 0 et de 1 qui codent les données et les programmes dans les ordinateurs, l'ADN forme un texte dont les mots sont les gènes, écrits dans un alphabet de quatre lettres A, G, C et T, correspondant aux quatre nucléotides. Ce texte code le plan de l'organisme tel que le prévoyait l'épigenèse.

Lexique

Aufhebung : Ce terme allemand cher à Hegel désigne, en biologie, la dualité de la mémoire génétique du monde vivant, à la fois préservée et peu à peu détruite par les mutations.

Dérive génétique : Évolution d'une population au hasard de mutations sans effet sur la reproduction, et donc non soumises à une pression de sélection.

Épigenèse : Théorie selon laquelle les graines constituent un plan de l'organisme et non une copie complète miniature de l'organisme parent.

Épigénétique : Ensemble des modifications de l'expression des gènes qui ne sont pas codées dans le génome.

Phylogénie : Généalogie du vivant fondée sur l'étude des caractères anatomiques et moléculaires des organismes vivants.

La réplication, un excellent moyen de conservation

Portée par l'ADN, une information numérique est donc transmise, de génération en génération, lors de la reproduction. Cette information, que l'on nomme le génome, est contenue dans toutes les cellules de tous les êtres vivants. Elle se propage par la réplication de l'ADN, qui le rend pérenne *(voir la figure 1)*. En effet, le support lui-même, comme les multiples supports de l'information numérique en informatique (carte perforée, cassette, cartouche, disquette, disque, mémoire flash), n'est pas conservé indéfiniment. L'ADN d'une cellule qui meurt se dégrade vite, notamment parce qu'il est assimilé par d'autres organismes.

Dans des conditions exceptionnelles de conservation, on peut retrouver une partie de l'information qu'il contient après quelques milliers d'années, difficilement après quelques dizaines de milliers d'années. Ce temps de conservation peut paraître long à l'échelle de nos vies, mais la réplication, elle, permet une préservation de l'information à une tout autre échelle : depuis plus de trois milliards d'années elle permet la copie des génomes et, en quelque sorte, la conservation de cette information dans les organismes vivants.

Grâce à l'ADN et à sa réplication, l'information contenue dans les génomes se propage, se transmet, se copie, se pérennise. Les organismes vivants, en se reproduisant, permettent cette transmission. En informatique aussi, la copie et la transmission de l'information sont un moyen de conservation des données pour faire face aux défaillances possibles des supports. Néanmoins, toute copie, toute transmission *via* un support, provoque des erreurs. Aucun support ni aucun canal de transmission n'est totalement sûr.

Le codage et la transmission de l'information numérique afin de minimiser ces erreurs sont l'objet de la théorie des codes. Le principe est d'introduire de façon méticuleuse et parcimonieuse une « redondance » dans l'information, afin de détecter

© Pour la Science - n° 433 - Novembre 2013

et de corriger les erreurs. De même, la réplication de l'ADN ne produit jamais des copies parfaites de l'information. La structure de l'ADN et des mécanismes de réparation permettent de corriger une partie des erreurs, sur le principe de la redondance. L'ADN est constitué de deux brins enroulés en une double hélice. Indispensable à la réplication, cette structure en double brin est aussi une façon de conserver l'information en deux copies identiques. Chaque nucléotide a un complémentaire, auquel il est apparié dans la double hélice : adénine est complémentaire de thymine, et guanine de cytosine. En conséquence, l'information est présente deux fois – sur une séquence de nucléotides et sur son complémentaire.

Lors de la réplication, il arrive que les mécanismes dérapent et que dans le brin copié, un nucléotide soit perdu ou inséré, ou pris pour un autre. Il en résulte un « mésappariement », c'est-à-dire la mise en correspondance d'un nucléotide avec un autre qui n'est plus son complémentaire. La plupart du temps, les erreurs ainsi détectées sont réparées en rétablissant le nucléotide complémentaire.

Mais le mécanisme de correction a ses limites, et les génomes répliqués contiennent toujours des différences par rapport à leurs ancêtres. D'abord parce que la fidélité a un coût pour l'organisme, en énergie et en temps passé à fabriquer la génération suivante. Ces erreurs constituent ce que l'on nomme les mutations. Il en existe de multiples possibles, de toutes les formes et de toutes les tailles : remplacement d'un nucléotide par un autre, insertion ou perte d'un nucléotide ou d'un segment de nucléotides, duplication ou déplacement d'un segment, duplication du génome tout entier. Au moins une centaine de mutations séparent chez l'homme les génomes des parents et des enfants, qui contiennent environ trois milliards de nucléotides. La seconde raison pour laquelle les systèmes vivants ne sont pas parfaitement fidèles, c'est parce que dans ces erreurs figurent parfois les clés de leur survie.

Une information qui évolue est-elle préservée ?

Aufhebung, en allemand, signifie à la fois préservation et destruction. C'est un de ces mots à double sens que le philosophe Georg Wilhelm Friedrich Hegel (1770-1831) affectionnait pour évoquer la dialectique, et c'est le mot qu'ont repris, en 1965, le biologiste austro-américain Emile Zuckerkandl et le chimiste américain Linus Pauling pour décrire la conservation de la mémoire du monde vivant par les molécules : préservation infinie en théorie grâce à la réplication, mais aussi destruction progressive par les mutations.

La plupart des mutations dans les génomes ne se propagent pas aux générations suivantes. D'une part, la sélection naturelle fait un tri. Si une mutation modifie l'efficacité de la reproduction dans un

LES AUTEURS



Vincent DAUBIN est directeur de recherche CNRS au Laboratoire Biométrie et biologie évolutive (UMR CNRS 5558) de l'Université Claude Bernard Lyon I.



Simon PENEL est ingénieur d'étude CNRS dans le même laboratoire.



Eric TANNIER est chargé de recherche de l'Inria dans le même laboratoire.

Ils sont tous trois membres du projet ANCESTROME qui vise à reconstruire les caractéristiques des organismes disparus en analysant leurs génomes avec des méthodes phylogénétiques (ancestrome.univ-lyon1.fr).

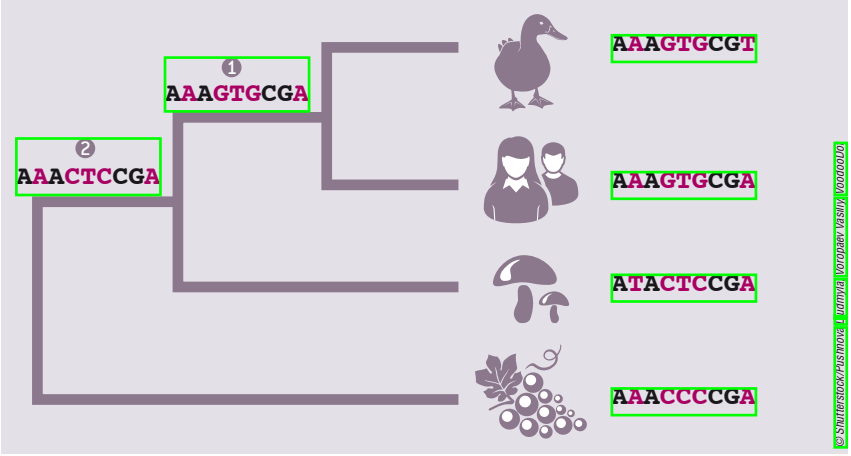
L'ADN et la métaphore informatique, comparaisons et raisons

Voir l'ADN comme une mémoire numérique pousse à l'analogie entre le monde vivant et les ordinateurs. L'ADN est le support d'une information numérique – le génome –, tout comme un disque dur est le support du logiciel ou des données. La métaphore peut être poussée loin, puisque les génomes sont des chaînes de caractères, à l'instar des données informatisées, suites de 0 et de 1. Le génome peut être vu comme un programme contenant les instructions nécessaires à la fabrication et à la reproduction d'un organisme, et le code génétique, qui permet la traduction des séquences génomiques en protéines, est une sorte de moyen d'interpréter ces instructions.

On peut aller encore plus loin. L'évolution a été qualifiée de bricolage par le biologiste François Jacob. Or la conception d'un programme informatique pourrait souvent, sans mettre en cause le professionnalisme des programmeurs, supporter cette qualification : un logiciel évolue toujours au fil du temps, chaque programmeur reprenant des morceaux de code de ses prédécesseurs en les amendant ou en intégrant des modules extérieurs, tels des transferts horizontaux. L'évolution des données numériques est ainsi comparable à celle des génomes.

Le génome a toutefois un statut particulier. Il constitue une sorte de programme autonome, puisqu'il code aussi la fabrication de la machine qui sera capable d'exécuter le programme et de se dupliquer. Là, on dépasse les possibilités des ordinateurs et des programmes informatiques (ce qui sera peut-être démenti sous peu par les imprimantes 3D autoréplicatives *RepRap*). De plus, le génome ne porte pas tout le matériel héréditaire : une part importante d'information est transmise lors de la reproduction, en dehors de la séquence, par « épigénétique ». Faute de bon modèle d'évolution intégrant ces caractères, la métaphore informatique s'arrête à l'information numérique, c'est-à-dire au génome.

On a les métaphores de son temps. Au XVIII^{e} siècle, les machines les plus perfectionnées étaient les automates, et la métaphore a beaucoup été utilisée pour décrire de façon mécaniste le fonctionnement du corps humain. On a aussi beaucoup comparé l'ADN à un langage, à une carte ou à un atlas. Aujourd'hui, la métaphore informatique est facilitée, d'une part, par l'utilisation constante des ordinateurs pour stocker les génomes et, d'autre part, par l'idée qu'il serait possible d'utiliser l'ADN pour le stockage de données informatiques.



2. SUR CET ARBRE PHYLOGÉNÉTIQUE SIMPLIFIÉ, on peut reconstruire, à partir des séquences d'un même gène repérées chez le canard, l'humain, le cèpe et le raisin, les séquences du même gène de l'ancêtre commun de l'humain et du canard d'une part *(1)*, et de l'ancêtre des animaux et des champignons d'autre part *(2)*. On suppose que les ressemblances sont dues à un héritage commun plutôt qu'à des événements indépendants. Les séquences les plus proches sont celles du canard et de l'humain (une seule mutation). La séquence du même gène chez leur ancêtre commun sera donc AAAGTGCGA ou AAAGTGCGT. Les séquences du cèpe et du raisin portant un A à la fin, la première option est sans doute la bonne. La séquence de l'ancêtre commun du cèpe, du canard et de l'humain est celle qui minimise le nombre de mutations entre cet ancêtre et le dernier parent dans l'arbre, le raisin, soit AAACTCCGA.

partie infime de l'histoire du monde vivant. En revanche, la sélection permet une meilleure conservation de l'information : on retrouve dans les génomes actuels des traces de l'origine du vivant, il y a presque quatre milliards d'années.

Pour reconstituer l'information ancienne, on compare plusieurs génomes actuels dans un cadre phylogénétique. Par exemple, prenons les génomes d'une variété de vigne, d'un cèpe, d'un canard et d'un humain *(voir la figure 2)*. Les ressemblances entre séquences permettent de reconstituer les relations de parenté entre ces espèces, ce qui se traduit par un arbre phylogénétique, c'est-à-dire une généalogie des espèces. Cette généalogie implique que le canard et l'humain ont un ancêtre commun qui leur est exclusif et qui descend lui-même de l'ancêtre qu'ils partagent avec le cèpe. On identifie alors des morceaux (des gènes) suffisamment similaires pour que l'on suppose qu'ils descendent d'une même séquence ancestrale : on trouve ATACTCCGA dans la vigne, AAACCCCGA chez le cèpe, AAAGTGCGA chez le canard et AAAGTGCGT chez l'humain.

À partir de l'analyse de ces séquences actuelles d'un même gène, on reconstitue les séquences du gène chez les ancêtres communs des espèces considérées en supposant que les ressemblances entre espèces sont plus fréquemment dues à leur héritage commun qu'à des événements indépendants qui auraient donné le même résultat : dans notre exemple, cette hypothèse se traduit par un scénario unique où chacune des positions variables *(en rouge)* n'a changé qu'une seule fois au cours de l'évolution des espèces.

En multipliant le nombre de gènes et de génomes comparés (aujourd'hui on peut avoir accès aux séquences de plusieurs milliers de génomes, chacun contenant des milliers de gènes, *voir l'article de G. Perrière dans ce numéro*) et avec l'aide de modèles plus réalistes prenant en compte nos connaissances des processus biologiques, on peut ainsi estimer des séquences de génomes ancestraux et décrire l'histoire des mutations.

En pratique, on sait aujourd'hui reconstruire les séquences quasi complètes d'ancêtres récents, par exemple le génome de la bactérie qui fut probablement responsable des grandes épidémies de peste du Moyen Âge, à l'aide du séquençage de ses descendantes, des

sens ou dans l'autre, alors sa fréquence augmente ou diminue dans une population au fil des générations. D'autre part, il n'est pas systématique qu'une mutation induise un avantage ou un inconvénient pour la reproduction. Dans certains génomes, par exemple ceux des mammifères ou de certaines plantes qui sont de longs textes verbeux, la plupart des mutations sont sans effet immédiat. Dans ce cas, la fréquence des génomes qui portent une mutation suit une marche aléatoire sans pression de l'environnement. C'est le principe de la dérive génétique.

Que cette variation soit influencée ou non par la sélection naturelle, la fréquence finit toujours par atteindre 0 ou 1. Si elle atteint 0, la mutation a disparu de la population ; il n'en reste pas de trace dans les génomes. Si elle atteint 1, tous les génomes de la population sont porteurs de la mutation. Dès lors, il ne reste pas de trace dans les génomes de leur état ancestral, avant la mutation.

Avec un nombre réaliste de mutations par génération (une centaine), la dérive remplace la plupart des lettres du texte génomique en quelques millions de générations. Ainsi, on ne peut retrouver des traces que de génomes ayant au plus quelques millions d'années – ce qui représente une

BIBLIOGRAPHIE

G. J. Szöllösi *et al.*, **Lateral gene transfer from the dead**, *Systematic Biology*, vol. 62, pp. 386-397, 2013.

J.-G. Dumas *et al.*, **Théorie des codes**, Dunod, 2013.

G. J. Szöllösi *et al.*, **Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations**, *PNAS*, vol. 109, pp. 17513-17518, 2012.

S. Abby *et al.*, **Lateral gene transfer as a support for the tree of life**, *PNAS*, vol. 109, pp. 4962-4967, 2012.

B. Boussau *et al.*, **Parallel adaptations to high temperatures in the Archaean eon**, *Nature*, vol. 456, pp. 942-945, 2008.

E. Zuckerkandl et L. Pauling, **Molecules as documents of evolutionary history**, *Journal of Theoretical Biology*, vol. 8, pp. 357-366, 1965.

© Pour la Science - n° 433 - Novembre 2013

souches de *Yersinia pestis* contemporaines. Pour des ancêtres plus éloignés, tel celui des mammifères qui vivait il y a une centaine de millions d'années, une partie de l'information a été perdue par dérive. Cependant, pour les parties du génome contraintes par la sélection, les gènes en particulier, la reconstruction reste tout à fait possible. On peut même connaître le nombre de chromosomes et la façon dont les gènes étaient organisés dans ces génomes. À partir de ces données, on commence à comprendre les modes de vie des espèces disparues (température optimale de croissance, profondeur préférée d'anciens mammifères marins, utilisation de l'oxygène par des bactéries anciennes, alimentation, taille ou longévité de mammifères anciens, histoire des migrations et métissages humains…), un peu comme si l'on retrouvait le code d'un vieux programme et que l'on reconstituait son fonctionnement à partir de ses versions nouvelles *(voir la figure 3)*.

Cependant, l'application de ces méthodes à l'ensemble du monde vivant se heurte à une difficulté : nombre d'organismes ont la capacité d'intégrer dans leur génome des gènes de leurs contemporains, proches apparentés ou non. Ces transferts horizontaux violent les hypothèses de reconstruction des génomes ancestraux, qui supposent que l'ancêtre commun des séquences appartient à l'ancêtre commun des espèces.

Le transfert de gène : obstacle ou indice ?

Il est apparu que chaque gène, quelle que soit sa fonction, pouvait, à un moment ou un autre de son histoire, avoir été transféré. À long terme, aucun gène n'a donc la même histoire que les organismes qui le portent. Ce constat semblait remettre en cause la possibilité théorique de reconstruire l'évolution du vivant à partir des gènes.

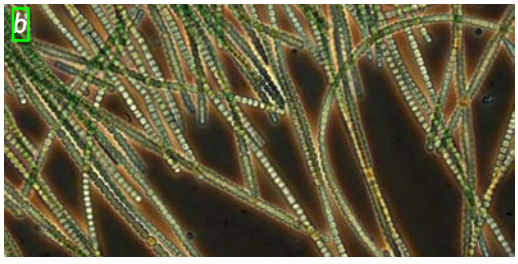
Ainsi, le gène était trop « individualiste » pour refléter l'histoire des organismes. N'avait-il pour autant rien à en raconter ? De fait, lorsqu'on observe l'ensemble des gènes et, pour peu que l'on sache interpréter les singularités que revèle chaque histoire de gènes, on s'aperçoit que les transferts horizontaux contiennent une information inédite sur l'histoire du vivant. D'une part, ils peuvent venir à l'appui d'une hypothèse phylogénétique : si un ancêtre a reçu un gène par transfert, ce gène doit être observé chez ses descendants. Ainsi, la présence du gène peut être un caractère partagé permettant de reconnaître les organismes qui sont proches.

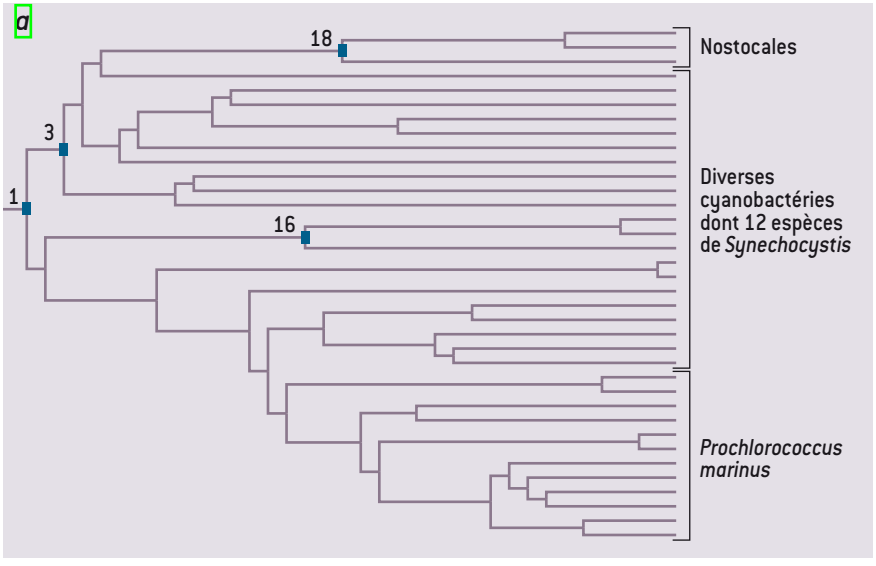
D'autre part, un échange de gènes entre deux organismes ne peut avoir lieu que s'ils sont contemporains. Le transfert horizontal contient donc une information sur les âges relatifs des lignées qui échangent des gènes : on ignore à quelle époque exactement vivaient les organismes anciens, mais, avec des relations de descendance et des échanges entre espèces contemporaines, on peut espérer établir une chronologie de l'histoire de la vie. Or les transferts ont lieu principalement chez les organismes microscopiques (bactéries, virus, champignons unicellulaires, autres

TRANSFERT DE GÈNES ET DATATION

Par définition, le transfert de gènes entre deux espèces n'est possible que si les deux espèces sont contemporaines. Ce processus, qui transparaît dans le génome des espèces actuelles, donne aux phylogénéticiens de précieuses informations sur les âges relatifs d'espèces anciennes dont les traces fossiles sont rares. En 2012, à partir de 36 génomes d'espèces différentes, les auteurs ont ainsi reconstitué la phylogénie chronologique des cyanobactéries, les premiers organismes à avoir fait de la photosynthèse et produit de l'oxygène en grande quantité *(a)*. Les plantes ont récupéré cette faculté en intégrant des cyanobactéries au fonctionnement de leurs cellules.

Les nœuds de l'arbre sont positionnés de gauche à droite selon leur chronologie. Le nœud 1, le plus ancien, est parfois daté à 3,8 milliards d'années. Les plantes ne figurent pas sur cet arbre, mais on voit la naissance au nœud 18 du clade des cyanobactéries multicellulaires, les nostocales, qui forment des filaments verts de cellules différenciées *(b)*. Comme ce clade a reçu, par transfert, des gènes d'autres morceaux de la phylogénie, on estime que cette forme de multicellularité est relativement récente. De même, la présence de nombreux tranferts du clade du nœud 3 vers le clade du nœud 16 lui donne une origine relativement récente, alors que les hypothèses habituelles plaçaient cette diversification à la racine des cyanobactéries.





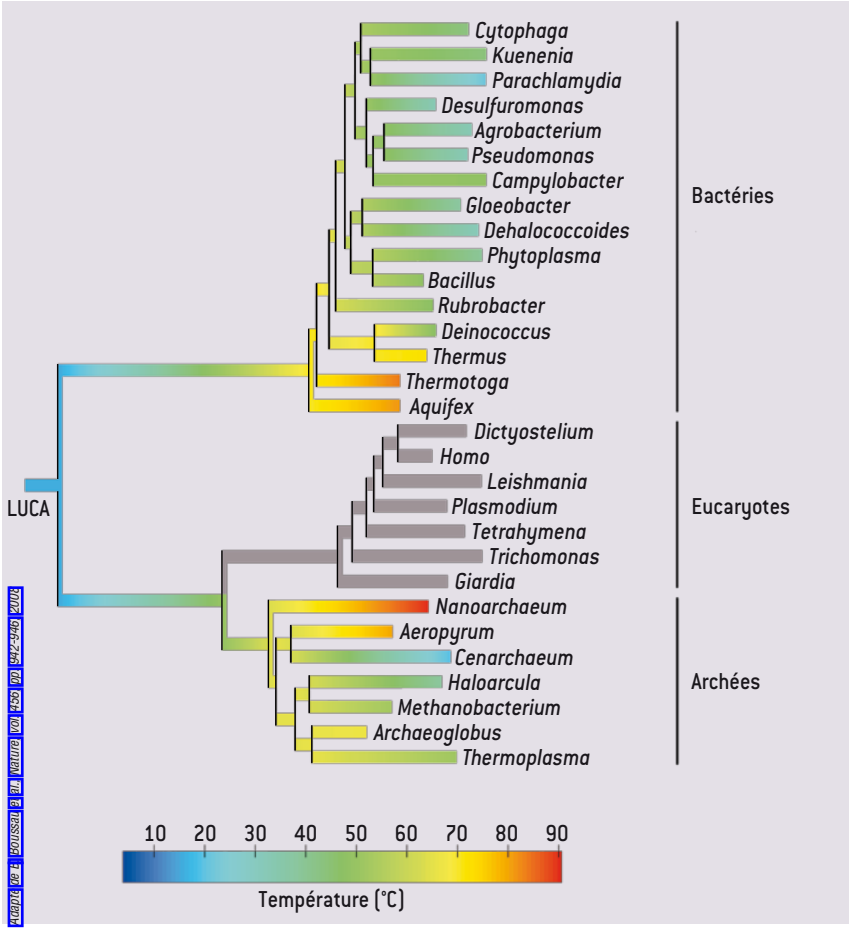
Publié par V. Szöllosi et al., PNAS, Volume 109, pp. 7513-7518, 2012

organismes unicellulaires tels que l'amibe ou la paramécie). Ces organismes ne laissent presque aucune trace fossile, ce qui rend la datation des trois quarts de l'histoire de la vie inaccessible aux méthodes classiques. Le transfert fournit ainsi l'outil de datation dont la petite taille de ces organismes nous privait.

Le transfert recèle encore d'autres traces inexploitées de l'histoire de la vie. Parmi les êtres vivants anciens, certains ont laissé des descendants, d'autres non. Il est raisonnable de penser que ceux qui n'ont laissé aucun descendant sont bien plus nombreux. Pourtant, on reconstitue l'histoire à partir des génomes actuels – les descendants de ceux qui, par définition, ont laissé des descendants ! On n'a ainsi accès qu'à une faible partie de la biodiversité ancienne.

Et si l'on détecte un échange de gènes entre deux organismes anciens, c'est à partir des descendants de celui qui a reçu le gène. En revanche, il n'y a aucune raison pour que celui qui l'a donné, lui, fasse partie de la minorité qui a eu des descendants. Statistiquement, il est probable qu'on ait ainsi accès à un gène d'une espèce ancienne sans descendant, qui n'aura donc pas transmis ses gènes en dehors de ce transfert. On est donc face à une biodiversité beaucoup plus importante que celle que nous offrirait la seule transmission d'ascendant à descendant.

Les gènes ne sont pas les marqueurs dont on avait rêvé pour reconstruire l'histoire de la vie. Ils racontent parfois une autre histoire, la leur. Mais si on sait modéliser la complexité de ces histoires, ils permettent de trouver, dans les génomes, des informations dont on pensait avoir perdu la trace. Il reste sans doute encore beaucoup à découvrir, dans nos cellules, sur notre passé.



3. RECONSTITUTION DE LA TEMPÉRATURE OPTIMALE DE CROISSANCE de l'ancêtre commun à tout le vivant (LUCA) à partir de l'analyse phylogénétique d'organismes actuels (65 eucaryotes, 60 archées et 331 bactéries). En se fondant sur l'observation selon laquelle chez les procaryotes (archées et bactéries), la température optimale de croissance est corrélée à la proportion de nucléotides C et G dans le génome, des chercheurs du laboratoire des auteurs ont décrit une évolution de la température en deux temps : LUCA vivait à environ 20 °C, puis la tolérance thermique de ses descendants a augmenté, avant de diminuer pour les espèces les plus récentes.

D'autres supports que l'ADN ?

En informatique, les supports de l'information numérique ont beaucoup évolué depuis les cartes perforées. Ils sont encore très divers aujourd'hui. Dans le monde vivant, il existe également plusieurs types de supports, mais c'est la double hélice d'ADN qui s'est établie comme le support universel chez les organismes cellulaires. Pourtant, ce support n'aurait pas été le premier.

L'ARN, toujours utilisé aujourd'hui par certains virus et conservé comme intermédiaire pour la lecture de l'information par les cellules, pourrait bien l'avoir précédé. Il avait l'avantage d'être capable de catalyser des réactions enzymatiques, offrant la possibilité d'organismes primitifs assez simples, dont le support d'information génétique est en même temps actif biochimiquement. Mais cette activité est également un facteur d'instabilité, et les organismes de ce monde à ARN auraient finalement adopté un média dédié pour préserver leur précieuse information génétique.

Cette transition est comparable à l'invention de l'imprimerie qui, plus rapide pour copier l'information écrite et moins sujette à l'erreur, a remplacé le scribe. Ou, en informatique, à l'abandon des cartes perforées qui stockaient l'information numérique lisible par les ordinateurs, pour des supports électromagnétiques.

Aujourd'hui, l'ensemble du vivant a adopté l'ADN comme média de stockage et de transmission de l'information numérique. Même l'informatique envisage d'effectuer cette transition : l'ADN pourrait remplacer les supports magnétiques pour stocker l'information numérique et être utilisé par les ordinateurs à la place des disques durs *(voir l'article de Ch. Dessimoz dans ce numéro)*. Cette innovation pourrait avoir des conséquences sur le monde vivant lui-même : pour des organismes habitués à récupérer l'ADN dans l'environnement par transfert horizontal, les grandes quantités d'ADN présentes dans des machines constitueraient un substrat pour inventer de nouveaux gènes. L'informatique elle-même changerait alors le cours de l'évolution ! ■

© Pour la Science - n° 433 - Novembre 2013

COLLECTION POUR LA SCIENCE LES PETITS THEMATIQUES

Spécial T.I.P.E. 2013-2014

Retrouvez tous les articles de référence parus dans *Pour la Science* sur le thème Transfert-Échanges

Déclinés en biologie, physique, informatique, archéologie et environnement, ces cinq feuilletables numériques vous aideront pour préparer votre T.I.P.E. !



3,99 €
35 p.
environ

Des dizaines d'autres sujets à découvrir sur
www.pourlascience.fr/petitsthematiques

La mémoire biologique est-elle codée ?

Franck Chaillan

La mémoire biologique repose sur des mécanismes beaucoup plus élaborés que le codage binaire des mémoires numériques. Elle est efficace, pertinente, adaptée au contexte et capable d'oubli. Mais, malgré la plasticité cérébrale, sa capacité de stockage est probablement limitée.

Aujourd'hui, c'est mon anniversaire. Ce matin, je suis venu au laboratoire en moto, en passant par le bord de mer, puis en empruntant l'autoroute. Quand je conduis, dès que j'aperçois la moindre petite lumière rouge à l'arrière d'une voiture, je ralentis pour éviter tout freinage d'urgence. En revanche, je ne prête plus attention à la haute tour qui domine le port. Pour dépasser rapidement la voiture qui précède, je regarde dans mes rétroviseurs, je coupe les gaz, j'embraye, je passe de la 6e à la 5e vitesse, je relâche l'embrayage, j'accélère et je me penche sur la gauche pour déboîter, tout cela automatiquement. Enfin, arrivé au laboratoire, je fais une recherche bibliographique, je rassemble mes connaissances et réfléchis à la façon de les organiser pour rédiger cet article sur la mémoire.

Pourquoi vous ai-je raconté cela ? Si ce court récit ne vous a guère intéressé (et c'est tout à fait normal !), il illustre la place que prend la mémoire dans notre quotidien et ses innombrables formes. En quelques lignes, j'ai fait appel à la mémoire épisodique, celle qui me concerne, en évoquant mon anniversaire et pour rédiger cette introduction ; la mémoire spatiale en décrivant mon trajet ; la mémoire non déclarative qui m'aide à ne pas faire attention aux événements ou objets habituels (la tour) ; la mémoire non déclarative procédurale, celle que j'utilise pour conduire, dépasser, embrayer, passer les vitesses sans que j'aie à y réfléchir ; la mémoire déclarative sémantique lors de ma recherche bibliographique et ma mémoire de travail pour organiser mes idées.

Les prémices de la plasticité

La mémoire biologique, qui permet le stockage des informations, est multiple. Elle est étroitement liée à l'apprentissage, c'est-à-dire à l'acquisition d'informations qui permettent à un individu d'adapter son comportement à l'environnement. Dès 1890, le psychophysiologiste américain William James proposa l'existence d'une mémoire primaire qui maintient temporairement une information dans le cours de la pensée, et d'une mémoire secondaire qui stocke cette information à long terme.

En 1911, le psychologue américain Edward Thorndike émit l'idée que la plasticité du système nerveux permet la création des traces mnésiques (ou souvenirs), lesquelles sont associées à des comportements. Aujourd'hui, le rôle de la plasticité cérébrale dans la formation, le stockage, le rappel et la reconstruction des souvenirs, leur éventuelle évolution et l'oubli est avéré. Nous examinerons ici en quoi consistent la plasticité cérébrale et son rôle dans la mémoire. Puis nous comparerons les performances de la mémoire biologique et des mémoires numériques, et nous nous interrogerons sur la façon dont les secondes peuvent perturber la première.

Commençons par rappeler que la mémoire est à court terme ou à long terme. La mémoire de travail est dite à court terme, car elle maintient actives durant quelques secondes ou minutes des informations nécessaires à un raisonnement ou à une opération mentale. Dès que l'opération est terminée, les données présentes dans la mémoire de travail sont oubliées ou stockées dans une mémoire à long terme. Cette dernière correspond au stockage des souvenirs. Elle se subdivise en une mémoire déclarative et une mémoire non déclarative. La mémoire sémantique, celle des connaissances que l'on a sur le monde, et la mémoire épisodique, celle des événements qui nous sont propres, sont des mémoires déclaratives. La mémoire procédurale, celle

© Shutterstock / VLADGRIN



L'ESSENTIEL

- La mémoire biologique repose sur une chaîne d'acteurs, des plus élémentaires (diverses molécules) aux plus complexes, de vastes réseaux de neurones activés simultanément.
- La plasticité cérébrale est la clé de la mémoire biologique.
- La mémoire numérique n'est que la somme de toutes les connaissances humaines. Elle ne crée pas – pour l'instant – de savoir nouveau, contrairement à la mémoire biologique.

des apprentissages moteurs, des automatismes, la mémoire des apprentissages non associatifs (ne plus réagir à un stimulus, par exemple la sonnerie de l'horloge), ou encore le conditionnement classique (de Pavlov), sont des mémoires non déclaratives. Et cela n'est qu'un panorama superficiel de « l'anatomie de la mémoire ». Cette rapide description macroscopique posée, que s'est-il passé quand j'ai appris à conduire une moto et quand, aujourd'hui, je conduis ? Pour le comprendre, nous devons nous placer au niveau des neurones, les supports de nos pensées et de nos souvenirs. Rappelons d'abord comment fonctionne un neurone et comment il communique avec ses voisins.

Tout est une question de communication entre neurones

Un neurone se compose de dendrites qui captent les informations, d'un corps cellulaire qui les intègre et d'un axone qui les transmet vers d'autres neurones. Les informations qu'il véhicule circulent des dendrites vers les terminaisons des axones. Ces informations sont des signaux électriques résultant des différences de charges électriques régnant de part et d'autre de la membrane neuronale. En effet, la répartition des ions négatifs (chlore, Cl^-, et protéines chargées négativement) et celle des ions positifs (sodium, Na^+, potassium, K^+ et calcium, Ca^{++}) diffèrent entre l'intérieur et l'extérieur des neurones.

Le milieu intracellulaire est riche en ions potassium et en protéines, les ions sodium et chlore étant plutôt concentrés à l'extérieur. L'intérieur d'un neurone contient davantage de charges négatives que l'extérieur. Il en résulte une différence de potentiel de l'ordre de –65 millivolts, valeur qui varie en fonction du type de neurones et de l'espèce animale. Cette différence de potentiel est nommée potentiel de repos. Quand une telle différence de potentiel règne de part et d'autre de la membrane neuronale, le neurone est dit polarisé. Il est inactif et ne véhicule aucune information, aucun potentiel d'action.

Mais cette différence de potentiel varie quand des ions traversent la membrane du neurone (y entrent ou en sortent) par différents types de canaux *(voir l'encadré pages 112 et 113)*. Prenons l'exemple de la propagation d'une information entre deux

neurones qui fonctionnent avec le glutamate, le principal neurotransmetteur excitateur (ces neurones glutamatergiques sont impliqués dans la mémoire). Au niveau de la terminaison présynaptique, les potentiels d'action véhiculés par l'axone entraînent l'ouverture de canaux ioniques calcium dépendant du voltage (aussi nommés canaux voltage-dépendants), c'est-à-dire qui s'ouvrent ou non selon la différence de potentiel qui règne de part et d'autre de la membrane neuronale. Rappelons qu'une synapse est composée des boutons pré- et postsynaptiques et de l'espace qui les sépare.

L'entrée massive d'ions calcium dans le bouton présynaptique favorise la fusion des vésicules synaptiques (petits réservoirs sphériques) qui contiennent le glutamate avec la membrane du neurone présynaptique. Après la fusion, la vésicule s'ouvre vers l'extérieur, et le neurotransmetteur est libéré dans la fente synaptique. Il s'y propage et se fixe sur des récepteurs portés par les dendrites ou le corps cellulaire du neurone postsynaptique. Ces récepteurs – AMPA et NMDA – sont des canaux qui laissent (ou non) passer les ions. La plupart du temps, les récepteurs NMDA ne s'ouvrent pas, car le canal est bloqué par un ion magnésium (Mg^{++}) volumineux.

Mémoire et plasticité synaptique

En revanche, les récepteurs AMPA s'ouvrent, et des ions sodium entrent dans le neurone : la différence de potentiel entre l'intérieur et l'extérieur du neurone diminue. Le neurone est activé : on dit qu'il est dépolarisé, c'est-à-dire qu'un potentiel postsynaptique excitateur est produit. Mais l'influence d'une synapse d'un neurone présynaptique sur le potentiel du neurone postsynaptique est limitée spatialement et, s'il n'est pas entretenu, le signal électrique émis au début de la dépolarisation s'affaiblit, puis disparaît.

Comment le signal est-il entretenu ? Au début d'un axone se trouve une zone – le segment initial ou la zone gâchette –, qui contient beaucoup de canaux ioniques voltage-dépendants. L'ensemble des potentiels postsynaptiques excitateurs et inhibiteurs provenant des dendrites et du corps cellulaire sont intégrés. Cela signifie qu'au niveau du segment initial, si la somme des excitations et des inhibitions donne une différence de potentiel égale à –40 millivolts, alors les canaux voltage-dépendants (sodium et potassium) sont automatiquement activés, et un potentiel d'action (ou signal électrique) est produit.

Supposons qu'un premier canal sodium du segment initial s'ouvre : il en résulte une dépolarisation et un potentiel d'action qui se termine par la fermeture de ce canal et l'ouverture d'un canal potassium qui se referme rapidement à son tour. Mais la zone de dépolarisation s'étend jusqu'au deuxième canal sodium, c'est-à-dire qu'à son niveau, la différence de potentiel passe de –65 millivolts à –40 : un potentiel d'action est émis. Il se termine avec la fermeture du canal sodium et l'ouverture temporaire d'un canal potassium. Et ainsi de suite. De proche en proche, les canaux sodium et potassium voltage-dépendants s'ouvrent, puis se ferment, permettant au potentiel d'action de se propager sans diminuer d'intensité tout le long de l'axone, jusqu'aux boutons synaptiques.

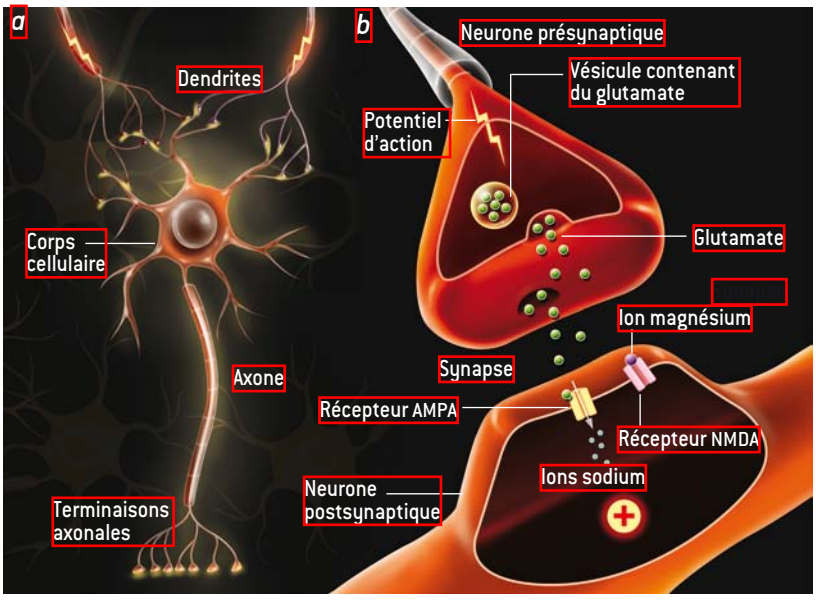
Pourquoi la plasticité cérébrale a-t-elle une importance dans le sujet qui nous occupe, à savoir la formation des souvenirs, leur stockage et leur rappel ? Dès 1949, le neuropsychologue canadien Donald Hebb posait les fondements des mécanismes de la mémoire. Selon lui, des « assemblées cellulaires », ou réseaux de neurones, sous-tendent la perception, mais aussi l'attention, la pensée ou la mémoire. Certains neurones sont parfois communs à plusieurs assemblées neuronales. Leur activation répétée et simultanée par une même stimulation modifie le fonctionnement ou la forme des contacts synaptiques entre ces neurones.

LES MÉCANISMES MOLÉCULAIRES DE LA MÉMOIRE

La mémoire s'organise en plusieurs niveaux : moléculaire, cellulaire, réseaux de neurones localisés, aires cérébrales, réseaux de neurones communiquant à grande distance et activés de façon intense et simultanée. Si l'on ne connaît pas encore de façon précise comment fonctionnent ces réseaux à grande échelle, certains mécanismes moléculaires et cellulaires ont été précisément décrits.

Chaque neurone est constitué d'un corps cellulaire, de dendrites et d'un axone. Les informations circulent des dendrites vers le corps cellulaire, puis dans l'axone jusqu'aux terminaisons axonales qui se connectent avec les dendrites ou le corps cellulaire du neurone suivant *(a)*.

Supposons qu'un signal, ou potentiel d'action, arrive à une terminaison axonale. Que se passe-t-il au niveau du bouton synaptique ? Les neurones impliqués dans la mémoire libèrent du glutamate, le principal neurotransmetteur excitateur. À l'arrivée d'un signal dans le neurone présynaptique, des vésicules contenant le glutamate fusionnent avec la membrane du neurone et déversent leur contenu dans la synapse *(b)*. Dans cet exemple, le neurone postsynaptique porte deux des quatre sous-types de récepteurs du glutamate : NMDA et AMPA. Ces récepteurs sont des canaux laissant passer respectivement des ions calcium et des ions sodium. Généralement, le canal NMDA est bloqué par un volu-



Sauf mention contraire, les illustrations sont de Sophie Nadot

© Pour la Science - n° 433 - Novembre 2013

C'est ainsi que certains contacts synaptiques sont renforcés : ce sont les traces mnésiques.

Prenons un exemple. Sur la route que j'emprunte pour me rendre au laboratoire, il y a cette tour que j'ai mentionnée dans l'introduction. La première fois que je l'ai vue (et les fois suivantes), certains de mes neurones se sont activés « quasi simultanément ». Il s'agissait des neurones codant la couleur de la tour, sa forme, sa hauteur, son environnement. Les neurones activés simultanément par la tour font désormais partie d'une même assemblée cellulaire qui code et stocke l'information « tour du port ».

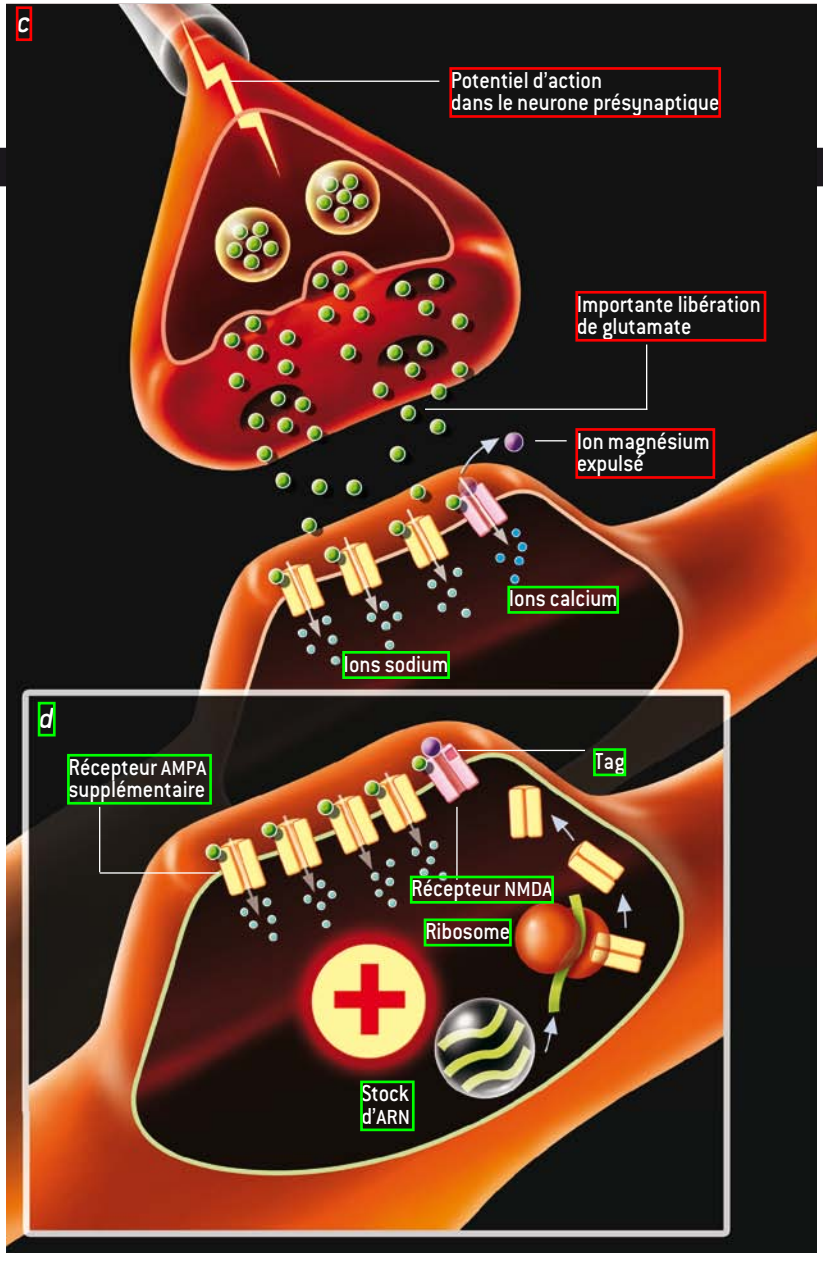
En 1973, les Britanniques Timothy Bliss et Tony Gardner-Medwin et le Norvégien Terje Lømo ont montré, chez le lapin, qu'un phénomène dit de potentialisation à long terme renforce l'efficacité de la transmission synaptique dans le système nerveux central, confortant la théorie de Hebb. Une dizaine d'années plus tard, on découvrait l'existence du phénomène inverse, une diminution de l'efficacité de la transmission synaptique dans le cervelet : c'est la dépression à long terme.

En quoi consistent les phénomènes de potentialisation et de dépression à long terme ? Les plus étudiés reposent sur l'activation des récepteurs AMPA et NMDA. Reprenons l'exemple des deux neurones glutamatergiques. Une potentialisation à long terme est déclenchée par l'arrivée de plusieurs potentiels d'action

L'AUTEUR



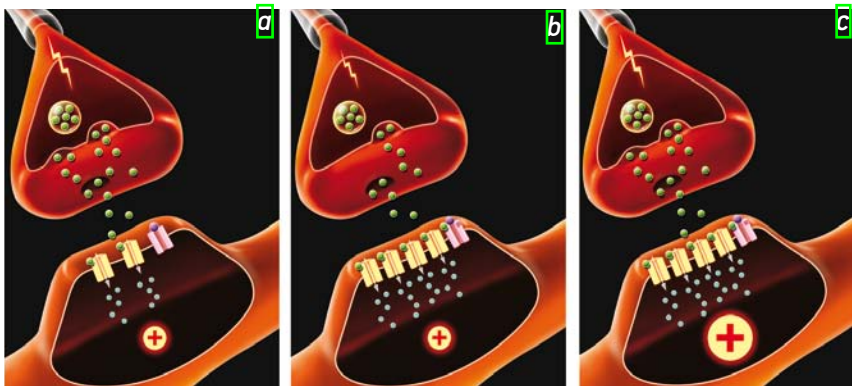
Franck CHAILLAN est maître de conférences à l'Université Aix-Marseille, UMR-CNRS 7291, LNC, à Marseille.



mineux ion magnésium qui empêche le passage des ions. Quand le glutamate se fixe sur ses récepteurs AMPA, des ions positifs entrent dans le neurone postsynaptique, et ce dernier est activé.

La mémoire repose sur le renforcement de certaines synapses. C'est le phénomène de potentialisation à long terme : quand un signal intense véhiculé par le neurone présynaptique parvient à une synapse, de nombreuses vésicules libèrent leur contenu dans la synapse. Les récepteurs AMPA sont activés « plus fortement », ce qui permet une entrée massive d'ions sodium. Qui plus est, les ions magnésium qui bloquaient les canaux NMDA sont expulsés par répulsion électrostatique en raison de la présence de nombreux ions positifs dans le neurone postsynaptique, de sorte que des ions calcium entrent en masse dans le neurone, ce qui déclenche une potentialisation *(c)*.

Diverses modifications permettent l'entretien de la potentialisation à long terme. Ainsi, des canaux AMPA sont produits sur place et exposés en nombre sur la membrane. En effet, il existe des réserves d'ARN messagers codant ces canaux dans le bouton synaptique. La machinerie de traduction de ces ARN en protéines (des ribosomes) est également disponible localement *(d)*. Ces récepteurs sont produits et dirigés vers la membrane. Des messages sont émis vers le noyau cellulaire, indiquant que le stock s'amenuise ; de nouveaux ARN messagers et des protéines issus du noyau sont acheminés pour restaurer le stock. La synapse où ces ARN doivent être délivrés est marquée (« tag » porté par un complexe impliquant probablement le canal NMDA). Tous ces événements moléculaires simultanés renforcent cette synapse.



1. LORS D'UNE POTENTIALISATION à long terme, le neurone postsynaptique est davantage activé. Supposons qu'un signal présynaptique active un neurone postsynaptique *(a)*. Après potentialisation, l'activation est plus efficace : le neurone postsynaptique est activé par moins de potentiels d'action *(b)* ; ou alors, pour un même potentiel d'action présynaptique, le neurone postsynaptique est davantage activé *(c)*. Quant à la dépression à long terme, elle réduit l'activation. Pour un signal présynaptique donné *(d)*, l'activation est plus faible *(e)*, voire absente : la synapse est dite silencieuse *(f)*.

transmis simultanément dans le neurone présynaptique. Il en résulte une augmentation brève, mais importante, de glutamate dans la synapse. Comme la stimulation est importante, non seulement les récepteurs AMPA sont activés, mais c'est aussi le cas des récepteurs NMDA : les ions positifs dans le neurone postsynaptique sont si nombreux que l'ion magnésium qui bloquait le canal NMDA est expulsé (par répulsion électrostatique). Cette ouverture renforce notamment la concentration des ions calcium dans le neurone postsynaptique.

Une potentialisation à long terme entre deux neurones se traduit par une augmentation de l'efficacité de la transmission synaptique. Cela signifie que le neurone présynaptique active plus facilement le neurone postsynaptique (l'activation se produit avec moins de potentiels d'action) ou plus fortement (pour le même nombre de potentiels d'action, le signal véhiculé par le neurone postsynaptique est plus intense).

Inversement, une dépression à long terme affaiblit la transmission synaptique : le neurone présynaptique active moins efficacement le neurone postsynaptique et le signal véhiculé est moins intense *(voir la figure 1)*. Elle résulte d'une augmentation faible, mais durable, de la concentration postsynaptique en ions calcium grâce, par exemple, à une activation des récepteurs NMDA qui laissent entrer le calcium. On admet généralement que la potentialisation à long terme est le phénomène moléculaire qui sous-tend la mémorisation, alors que la dépression à long terme serait associée à l'affaiblissement des connexions, donc à l'oubli. Toutefois, la dépression à long terme participe aussi à l'apprentissage, ce qui peut sembler paradoxal *a priori*.

Comme l'ont montré Graham Collingridge, de l'Université de Bristol, et T. Bliss, la mise en place de la potentialisation et de la dépression à long terme dépend notamment de la simultanéité de mécanismes pré- et postsynaptiques. Différents mécanismes facilitent (ou inhibent) la dépolarisation du neurone postsynaptique, de sorte que la production de potentiels d'action augmente (ou diminue). La régulation des flux ioniques à travers les canaux est l'une des façons de contrôler la production des potentiels d'action, c'est-à-dire l'efficacité de la transmission synaptique. L'augmentation ou la diminution du nombre de ces récepteurs présents sur la membrane postsynaptique en est une autre.

Plasticité de courte ou de longue durée

De plus, certaines synapses sont silencieuses, car elles portent des récepteurs NMDA et des récepteurs AMPA inactifs sur le versant postsynaptique. Mais si les récepteurs AMPA deviennent actifs, ou si de nouveaux récepteurs AMPA sont intégrés, elles deviennent actives et fonctionnelles. Enfin, en fonction des sous-unités qui les composent, les récepteurs AMPA laissent aussi passer les ions calcium, favorisant la plasticité synaptique.

Les événements moléculaires qui ont lieu sur le versant présynaptique ont également des répercussions sur la transmission des signaux électriques. Quand la quantité de glutamate libérée dans la fente synaptique augmente, la potentialisation à long terme aussi ; quand elle diminue, c'est la dépression à long terme qui se met en place.

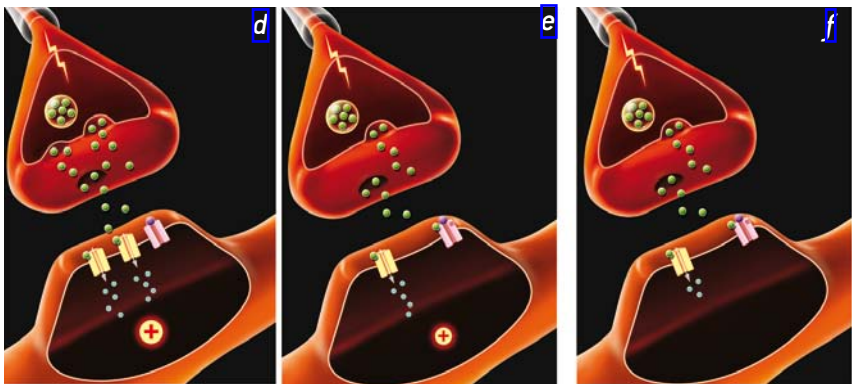
Quand une potentialisation ou une dépression à long terme est induite, des messagers dits rétrogrades sont libérés par le versant postsynaptique, mais agissent sur le versant présynaptique. Ils contrôlent notamment la quantité de neurotransmetteur libéré. Enfin, certaines données expérimentales montrent que le nombre des boutons synaptiques augmente lors d'une potentialisation à long terme et diminue lors d'une dépression à long terme. Ainsi, des modifications fonctionnelles et morphologiques se combinent.

Ces changements sont le support de tout apprentissage et sont renforcés par le phénomène de dépression à long terme. Ce dernier, comme nous l'avons évoqué, participe à l'oubli, mais également à l'apprentissage. Quand un phénomène de potentialisation à long terme renforce certaines synapses et que, simultanément, la dépression à long terme en affaiblit d'autres, le rapport du signal sur le bruit de fond augmente : la potentialisation à long terme renforce le signal, tandis que la dépression à long terme réduit le bruit de fond, de sorte que la trace mnésique est renforcée.

Une fois déclenchées, la potentialisation ou la dépression à long terme peuvent se maintenir plus ou moins longtemps (de quelques minutes à plusieurs mois). Tous les changements à long terme nécessitent la production de nouvelles protéines, par exemple de récepteurs AMPA à intégrer dans la membrane pour une potentialisation à long terme, ou une production d'enzymes phospholipases qui atténuent l'action de ces récepteurs en cas de dépression à long terme. Pour ce faire, les ARN messagers stockés dans la synapse renforcée par la potentialisation doivent être traduits en protéines (les récepteurs AMPA, par exemple). Ces protéines sont produites dans la terminaison synaptique, la machinerie de production des protéines – des ribosomes – y étant présente.

Toutefois, pour remplacer les ARN messagers utilisés et maintenir le stock, il est nécessaire de modifier à long terme la

© Pour la Science - n° 433 - Novembre 2013



transcription, c'est-à-dire la production de ces ARN messagers. Différentes enzymes, des kinases, servent de signaux transmis de la synapse jusqu'au noyau, où a lieu la production de nouveaux ARN. Simultanément, les synapses activées par une potentialisation ou inhibées par une dépression produisent un marqueur, ou tag. Les ARN messagers et les protéines nouvellement produits par le noyau sont redirigés vers les synapses et capturés spécifiquement par celles qui sont « taguées ». On ignore encore la nature précise de ces « tags ».

Différents niveaux de codage

La mémoire nécessite donc plusieurs niveaux de codage : le niveau moléculaire, qu'il s'agisse des neurotransmetteurs, des canaux récepteurs responsables de la dépolarisation, ou des tags signalant les synapses dont le fonctionnement est modifié. Le codage peut se faire aussi à un niveau supérieur, morphologique, par exemple avec la formation de boutons synaptiques, mais aussi à l'échelle des dendrites. On a longtemps admis que les dendrites étaient des éléments passifs qui servaient seulement à faire la somme des entrées excitatrices et inhibitrices, et à déclencher un potentiel d'action au niveau du segment initial de l'axone quand la somme était adéquate. Mais la surface des dendrites porte aussi des canaux sodium, potassium et calcium voltage-dépendants et des récepteurs NMDA qui en font des éléments actifs, capables de produire des potentiels d'action localisés dans les dendrites.

De plus, l'intégration d'une information (sommation des excitations et des inhibitions) ne se fait pas de la même façon tout le long d'une même branche dendritique, ce qui suggère l'existence de sous-unités dendritiques pouvant intégrer les entrées indépendamment (ou presque) les unes des autres. Chaque branche dendritique peut faciliter la plasticité des synapses activées par une information, la création de nouvelles synapses et leur renforcement à long terme. Les dendrites, par les milliers de synapses qu'elles portent, représentent un niveau de codage de l'information plus riche que le codage moléculaire.

Franchissons un niveau supplémentaire : des informations quasi simultanées pourraient être codées par des groupes (ou clusters) particuliers de synapses portés par des dendrites communes. Inversement, les informations mémorisées qui sont séparées par de longs intervalles de temps seraient plutôt codées par des clusters de synapses portés par des branches dendritiques différentes, voire des neurones différents, d'après Athanasia Papoutsis et ses collègues de la Fondation pour la recherche et la technologie, à Héraklion, en Crète. Cette hypothèse a été confortée par le modèle proposé par Robert Legenstein et Wolfgang Mass, de l'Université de technologie de Graz, en Autriche, montrant qu'un neurone peut s'auto-organiser pour stocker des informations différentes sur différentes dendrites.

Ainsi, l'apprentissage et la mémorisation reposent sur la création et le maintien de réseaux de neurones. Pour étudier le fonctionnement de ces réseaux, il faut pouvoir enregistrer simultanément l'activité d'un grand nombre de neurones et être capable d'analyser leurs interactions. L'équipe de Joe Tsien, de l'Université Georgia des sciences de la santé, aux États-Unis, a développé un dispositif pour enregistrer simultanément l'activité individuelle de plus de 250 neurones dans une aire spécifique de l'hippocampe d'une souris libre de ses mouvements (rappelons que l'hippocampe est une aire essentielle pour les mécanismes de mémorisation).

Ces neurobiologistes ont soumis les animaux à différentes expériences comportementales, par exemple un jet d'air, une chute dans une boîte ou une secousse dans une cage. Ces événements sont stockés dans la mémoire épisodique de la souris qui dépend de l'hippocampe. En analysant l'activité des neurones de l'hippocampe lors de ces différentes situations, il est apparu un patron d'activité cellulaire spécifique à chaque situation, c'est-à-dire que pour chaque événement, différents ensembles de neurones de l'aire étudiée sont activés. Ces différents réseaux de neurones sont nommés « cliques neuronales ». Un neurone n'appartient qu'à une seule clique, et les neurones d'une même clique ont la même activité lors d'un événement spécifique. Dans ce modèle, la clique est donc l'unité fondamentale de codage. Chaque clique correspond à un aspect plus ou moins précis d'un événement, et les cliques sont organisées de façon hiérarchique, de la plus générale à la plus spécifique.

Un événement serait donc toujours composé du même ensemble de cliques (un grand réseau de neurones composé de réseaux plus petits codant différentes informations). Par exemple, dans l'expérience menée par J. Tsien, il y a une clique spécifique d'un mouvement perturbateur, une pour un tremblement et une autre encore pour un tremblement dans une boîte noire. Selon ce modèle, une clique est soit active (1), soit inactive (0) : chaque événement serait représenté par un code mnésique binaire. De plus, de nouvelles « cliques » peuvent apparaître et se réorganiser selon les expériences du sujet, grâce aux phénomènes de plasticité synaptique.

Progressons encore d'un niveau vers le macroscopique. L'apprentissage d'une

LA POTENTIALISATION À LONG TERME renforce les signaux, tandis que la dépression à long terme réduit le bruit de fond : le rapport signal sur bruit est amélioré et le système plus efficace.

information ou son rappel nécessite la communication et la coordination de cliques intra- et interstructures. Or quand de vastes cliques déchargent simultanément, le courant résultant peut être détecté sur le cuir chevelu quand on enregistre l'activité par électroencéphalographie, EEG. La communication entre les cliques est probablement liée aux oscillations cérébrales. Celles-ci apparaissent quand on enregistre l'activité globale moyenne d'une structure cérébrale.

Ainsi, quand un animal doit apprendre à associer différents contextes à des emplacements où a été déposée une récompense, on observe dans l'hippocampe des ondes dites thêta, dont la fréquence est comprise entre quatre et dix hertz. Il existe aussi un rythme nommé gamma, dont la fréquence est comprise entre 30 et 150 hertz, et sans doute un code thêta-gamma où les deux rythmes s'imbriqueraient. C'est ce qu'ont montré en 2013 John Lisman, de l'Université Brandeis, aux États-Unis, et Ole Jensen, de l'Université Radboud, à Nimègue, aux Pays-Bas : les cellules qui codent diverses informations déchargent à différentes phases du rythme thêta, mais tous les neurones (qu'il s'agisse de cliques ou de réseaux) qui codent une même information déchargent au même moment du cycle gamma.

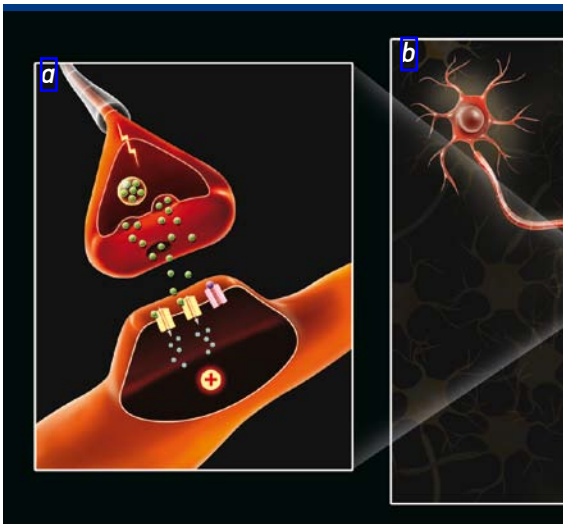
Enfin, ce « couplage oscillatoire » apporte un niveau supplémentaire d'organisation entre les structures de la mémoire. Toutes les structures nerveuses ne sont pas activées en même temps, de la même façon, pour les informations qu'il faut apprendre, mémoriser ou rappeler. L'hippocampe est impliqué dans la mémoire déclarative épisodique qui permet de stocker des informations ayant des repères temporels et spatiaux (mon trajet à moto, par exemple). On peut rappeler consciemment ces informations qui ont des repères contextuels. L'hippocampe est nécessaire pour apprendre et mémoriser les informations épisodiques. La trace mnésique de l'information est ensuite transférée de l'hippocampe vers le cortex pour être stockée à long terme. Simultanément, il y a un désengagement de l'hippocampe, c'est-à-dire que les cliques initialement sollicitées pour la mémorisation de tel ou tel savoir-faire ou connaissance ne sont plus activées (ou le sont plus faiblement) par ce stimulus, et ce sont des cliques présentes dans le cortex qui prennent le relais.

Ainsi, une information épisodique serait d'abord stockée dans l'hippocampe sous forme détaillée, mais, avec le temps, se développerait une version plus schématique de cette information au niveau du cortex. Gordon Winocur et ses collègues, de l'Université Trent, au Canada, ont récemment proposé que l'hippocampe et le cortex étant en interaction permanente, chaque trace mnésique pourrait être réactivée selon le contexte. Dès lors, quand un réseau reliant l'hippocampe et le cortex et correspondant à un souvenir particulier est réactivé, il est légèrement modifié. Les mécanismes de modification – même légère – des cliques neuronales codant les souvenirs expliquent que la mémoire humaine évolue avec le temps.

Mémoire biologique *versus* mémoire numérique

Après avoir brossé ce tableau de la mémoire biologique, peut-on comparer ses avantages et ses inconvénients avec ceux de la mémoire numérique ? Question simple en apparence, qui appelle une réponse complexe, car on ne compare pas deux « entités » similaires. Au sens strict, la mémoire numérique est le stockage pur et simple d'informations, par un code binaire, sur différents supports. D'un point de vue biologique, c'est un ensemble de processus étroitement liés qui assurent aussi bien l'acquisition que la restitution de l'information et dont on ne connaît pas encore les mécanismes précis. Le concept de mémoire biologique est plus complexe, plus large et plus flou (ou méconnu) que celui de mémoire numérique.

La mémoire doit stocker les informations pendant longtemps et sans altération. Sur ces deux aspects, l'avantage semble être à la mémoire numérique. En effet, il n'est pas facile de se rappeler précisément un événement et, pire, on peut se souvenir d'une information fausse, alors qu'il est si facile de faire une recherche sur le Web quand on a oublié une date, un chiffre ou la chronologie des rois de France. On obtient l'information en quelques clics. En revanche, elle n'est pas toujours fiable. De plus, la conservation à long terme d'une information numérique est toute relative. Au contraire, de par son organisation en réseaux de neurones interconnectés, le



2. LA MÉMOIRE EST ORGANISÉE en plusieurs niveaux, de la molécule (neurotransmetteurs, récepteurs, etc.), à la synapse, élément clé de la mémorisation *(a)*, puis aux dendrites qui

cerveau est capable de reconstituer une information (ou un comportement) même s'il lui manque une partie des informations qui s'y rapportent.

La capacité de stockage est un autre aspect de la mémoire. Celle-ci a augmenté exponentiellement pour les supports numériques, ce qui n'est pas le cas de notre cerveau. Toutefois, pour un support donné, la capacité de stockage est fixe et, une fois qu'elle est atteinte, il faut effacer des informations afin d'en stocker de nouvelles (ou prendre un autre disque dur !). Au contraire, la plasticité du système nerveux permet d'augmenter sa capacité de stockage (augmentation du nombre et de l'efficacité des synapses, des dendrites et des axones, du nombre de neurones, ou modification des réseaux de neurones). Autant de processus qui font que plus on apprend, plus il est facile d'apprendre !

Mais cela ne signifie pas que le cerveau ait une capacité de stockage illimitée. Nous ne savons toujours pas précisément comment il stocke et rappelle une information. Par conséquent, il est illusoire de vouloir calculer sa capacité comme on peut le faire pour un ordinateur. La « puissance » de stockage de la mémoire numérique résulte plus de la rapidité de communication entre les différents supports numériques (Web) que de la capacité d'un support (un ordinateur) en particulier. Par ailleurs, les données mémorisées sur Internet ne sont rien de plus que la somme de toutes les connaissances accumulées par l'homme

© Pour la Science - n° 433 - Novembre 2013



portent des milliers de synapses *(b)*. Viennent ensuite des clusters de synapses et de dendrites, et les réseaux locaux de neurones, ou cliques *(c)*, activés par le même événement. Enfin, on trouve des réseaux de neurones plus ou moins vastes, qui mettent en jeu des aires réparties dans tout le cerveau et communiquent. Le *cloud* de l'être humain ?

au fil du temps. La mémoire numérique est-elle supérieure à la « métamémoire » humaine, somme de toutes les connaissances de tous les êtres humains ?

Si l'on fait une comparaison (approximative) entre le cerveau et l'ordinateur, les deux systèmes présentent une mémoire procédurale qui permet de maintenir temporairement une quantité limitée d'informations pour réaliser des opérations. Dans un ordinateur, il s'agit de la mémoire vive (RAM), localisée au niveau des microprocesseurs réalisant les opérations, les résultats étant enregistrés sur le disque dur. Dans le cerveau, la mémoire est distribuée dans un réseau plus ou moins vaste de structures réparties dans les différents lobes du cerveau (préfrontal, pariétal, temporal et occipital) qui servent elles-mêmes à stocker des informations à long terme.

Le numérique perturbe-t-il la mémoire biologique ?

Au-delà des performances respectives des deux types de mémoires, l'utilisation de plus en plus importante des divers médias numériques, et donc des situations de tâches multiples simultanées qui demandent une division de l'attention, n'aura-t-elle pas une influence sur la mémoire biologique ?

L'utilisation de jeux vidéo qui nécessitent de gérer simultanément plusieurs paramètres (flux visuel dans l'environnement produit par le déplacement des personnages, évolution de leurs positions, interactions des personnages, gestion des armes, des munitions, etc.), souvent dans des fenêtres différentes, permet de développer les capacités multitâches des sujets. En revanche, le traitement de l'information selon que l'on réalise une tâche unique ou plusieurs tâches simultanées diffère. La résolution de tâches multiples simultanées entraîne un traitement automatisé et non conscient des informations qui fait intervenir des structures différentes de celles que nécessite une attention focalisée.

Quelle est l'influence sur la mémoire ? Un début de réponse est donné par les résultats de l'expérience réalisée en 2003 par Helene Hembrooke et Geri Gay, de l'Université Cornell, aux États-Unis. Durant un cours, la moitié des étudiants devait utiliser un ordinateur portable pour surfer sur le Web, chercher des pages relatives au cours (ou non), et donc suivre le cours en mode multitâches, pendant que l'autre moitié devait simplement écouter le cours. Lors du rappel du contenu du cours, les performances des étudiants multitâches ont été inférieures à celles des autres étudiants.

De plus, les étudiants ayant consulté plus de pages relatives au cours n'avaient pas de meilleures performances que les autres étudiants ayant utilisé Internet pour des recherches sans lien avec le cours. Ainsi, suivre un cours en situation de multitâches et de multisupports entraîne un déficit mnésique. Les étudiants « monotâche » ont suivi le cours de façon consciente, en focalisant leur attention, alors que les « multitâches » l'ont fait de façon superficielle, partageant leur attention entre l'ordinateur et le professeur.

Si l'on n'a pas encore mis au jour tous les mécanismes qui sous-tendent la mémoire biologique, il est avéré que cette mémoire est performante et pertinente : on sait retrouver, de façon non consciente et en quelques dixièmes de seconde le souvenir adapté à telle ou telle situation. Et même si Internet est plus rapide, donnant des liens en quelques centièmes de seconde, il faut ensuite soi-même faire le tri – conscient – entre les millions de pages correspondant parfois à la requête. L'internaute doit alors décider de la page qui répond vraiment à sa demande, ce qui prend généralement plusieurs minutes, voire bien davantage ! ■

BIBLIOGRAPHIE

A. Papoutsis *et al.*, **Coding and decoding with dendrites,** *J. Physiol.*, à paraître, 2013.

T. Bliss et G. Collingridge, **Expression of NMDA receptor-dependent LPT in the hippocampus : bridging the divide,** *Mol. Brain*, vol. 6(5), 2013.

J. Lisman et O. Jensen, **The theta-gamma neural code,** *Neuron*, vol. 77, pp.1002-1016, 2013.

M. Sanhueza et J. Lisman, **The CaMKII/NMDAR complex as a molecular memory,** *Mol. Brain*, vol. 6(10), 2013.

R. Legenstein et W. Maass, **Branch-specific plasticity enables self-organization of nonlinear computation in single neurons,** *J. Neurosci.*, vol. 31, pp. 10787-10802, 2011.

G. Winocur *et al.*, **Memory formation and long-term retention in humans and animals : Convergence towards a transformation account of hippocampal-neocortical interactions,** *Neuropsychologia*, vol. 48, pp. 2339-2356, 2010.

J. Tsien, **Le langage (codé) de la mémoire,** *Pour la Science*, n° 359, pp. 26-32, 2007.

HISTOIRE DES SCIENCES

Cinquante ans de chaos

En 1963, un météorologiste américain, Edward Lorenz, révéla que la prédictibilité déterministe est dans certains cas illusoire. La théorie du chaos prenait forme…

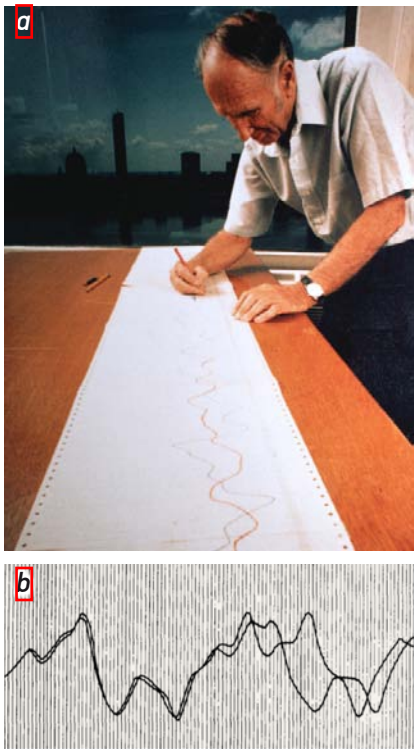
Adilson MOTTER et David CAMPBELL

En physique classique, on apprend qu'à partir des conditions initiales d'un système, on peut calculer tous ses états futurs. Un déterminisme que le mathématicien français Pierre Simon Laplace étendait à tout l'Univers et expliquait en ces termes dans sa *Théorie analytique des probabilités* (1812) : « Une intelligence qui, pour un instant donné, connaîtrait toutes les forces dont la nature est animée et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'Univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir, comme le passé, serait présent à ses yeux. »

Mais dans la réalité, la connaissance exacte d'un état initial est impossible. Jusqu'à la fin du XIXe siècle, l'hypothèse tacite était qu'une connaissance approchée de l'état initial conduisait à une connaissance approchée de l'état final. Au vu de leurs succès dans la description du mouvement des planètes, des comètes, des étoiles et de la dynamique d'innombrables autres systèmes, les physiciens n'avaient pas de raison de supposer le contraire.

Cependant, à partir de la fin du XIXe siècle, une série de travaux ont révélé que la notion de prédictibilité déterministe, bien que séduisante et intuitive, est fausse pour la plupart des systèmes. De petites incertitudes sur l'état initial peuvent conduire à d'énormes erreurs sur l'état final. Même des systèmes dont toutes les forces sont connues peuvent avoir un comportement imprévisible. Le déterminisme n'exclut pas le chaos.

Le point culminant de ces recherches fut un article publié il y a 50 ans par un météorologiste de l'Institut de technologie du Massachusetts (MIT), Edward Lorenz (1917-2008). C'est cette histoire de la découverte du chaos déterministe et de ses implications que nous racontons ici. Ses acteurs sont nombreux – trop pour être tous cités. Nous présentons nos excuses à ceux dont nous n'avons pu mentionner les contributions.



1. EDWARD LORENZ étudiant une série temporelle de solutions numériques de son modèle météorologique, générée par ordinateur *(a)*. En comparant deux séries produites à partir des mêmes équations, mais avec des conditions initiales légèrement différentes, il s'aperçut que les séries divergeaient rapidement *(b)*.

Une galerie de monstres

La théorie du chaos, telle que nous la connaissons aujourd'hui, a pris forme au cours du dernier quart du XXe siècle. Mais dès la fin des années 1880, les physiciens ont été confrontés au phénomène, à commencer par Henri Poincaré lorsqu'il s'intéressa au problème des trois corps en mécanique céleste. Il s'agissait de déterminer la trajectoire de trois corps interagissant selon la loi de la gravitation universelle.

De tels systèmes sont bien décrits dans l'espace des phases. Dans cet espace, les dimensions correspondent aux variables dynamiques – les positions, les quantités de mouvement –, qui permettent de décrire le système par des équations différentielles du premier ordre. L'idée dominante avait longtemps été que, laissé à lui-même, un système classique ordinaire finit par converger soit vers un état stable, correspondant à un point dans l'espace des phases, soit vers un état périodique (une boucle fermée dans l'espace des phases), soit vers un état quasi périodique (un tore dans l'espace des phases).

Or les trajectoires à trois corps calculées par Poincaré ne s'inscrivaient dans aucune de ces catégories : elles formaient un enchevêtrement complexe de courbes, aujourd'hui bien connu des théoriciens du chaos. Pour Poincaré, le déterminisme avait

Avec l'aimable autorisation de la formation Mamon de l'après-guerre Chaos Making a New Science Vikram 1987

© Pour la Science - n° 433 - Novembre 2013

trouvé ses limites : « Il peut arriver que de petites différences dans les conditions initiales en engendrent de très grandes dans les phénomènes finaux [...]. La prédiction devient impossible », écrit-il en 1912 dans l'introduction de son *Calcul des probabilités*.

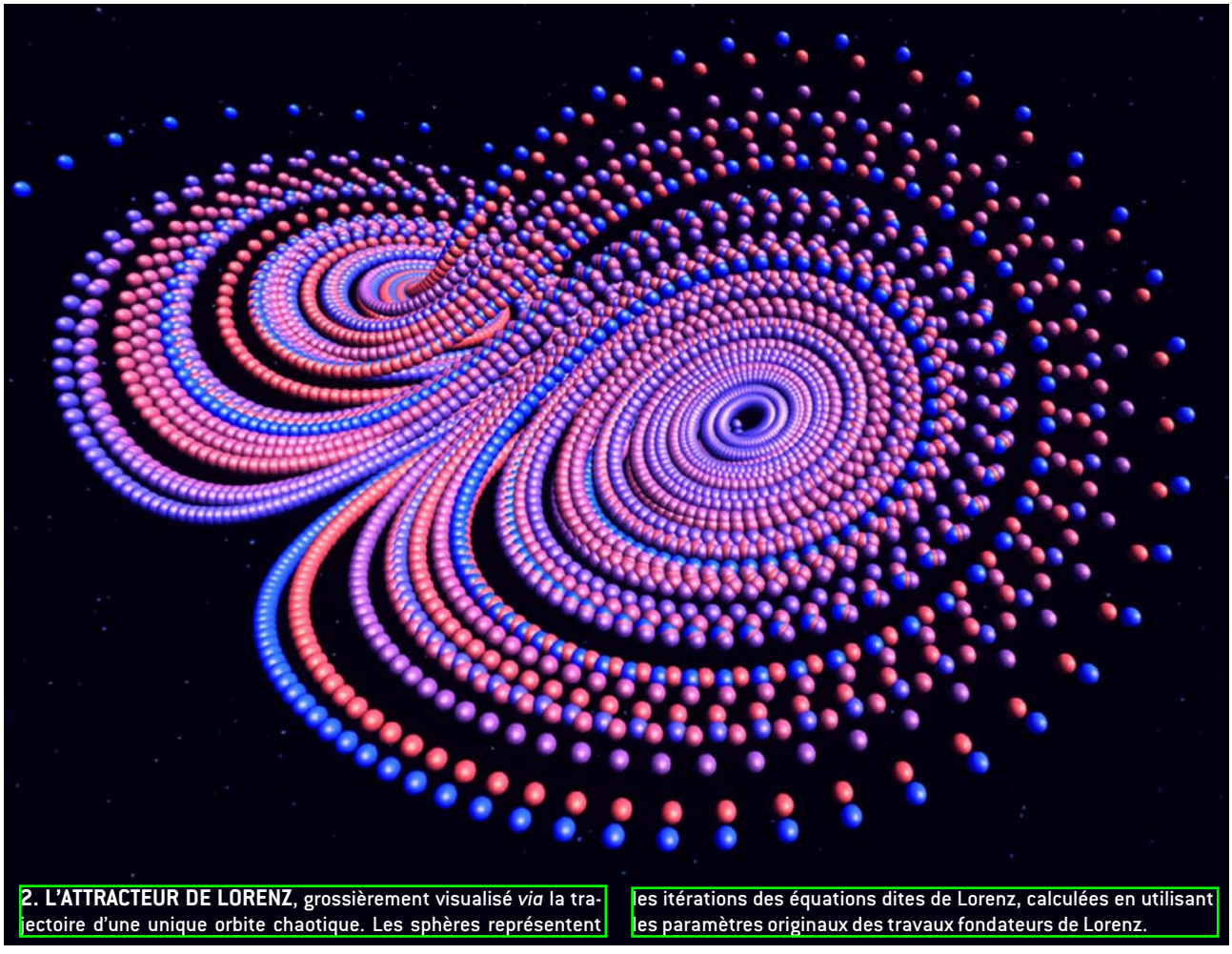
Pour la plupart des scientifiques de l'époque, les résultats de Poincaré étaient des artéfacts ou le fruit d'erreurs méthodologiques. Il en était de même de conclusions similaires indépendantes du mathématicien français Jacques Hadamard et d'indices expérimentaux du chaos observés par leurs contemporains, rangés eux aussi dans la galerie des monstres. La théorie du chaos ne s'imposa qu'un siècle plus tard.

Selon toute vraisemblance, Lorenz ne connaissait pas les travaux de Poincaré quand il s'est lancé dans la météorologie au milieu du XX^{e} siècle. Muni d'une maîtrise de mathématiques, Lorenz avait servi en tant que météorologiste pendant la Seconde Guerre mondiale, avant de finir sa thèse en météorologie au MIT et de rejoindre la faculté en 1955.

Une découverte fortuite

À l'époque, la plupart des météorologistes prévoyaient le temps en utilisant des procédures linéaires, qui reposaient sur l'hypothèse selon laquelle le temps de demain est une combinaison linéaire bien définie des caractéristiques du temps d'aujourd'hui. Mais selon une nouvelle école émergente, le temps pouvait être prédit avec davantage de précision en simulant les équations de la dynamique des fluides décrivant les écoulements atmosphériques. Lorenz, qui venait d'acquérir son premier ordinateur, un *Royal McBee LGP-30* avec une mémoire interne de 4 096 mots à 32 bits (mille fois plus rapide que sa calculatrice de bureau, mais un million de fois plus lent qu'un ordinateur portable actuel), décida de comparer les deux approches en confrontant les procédures linéaires à un modèle dynamique simplifié à 12 variables.

Lorenz se concentra sur la recherche de solutions non périodiques, qui lui semblaient plus difficiles à modéliser à l'aide de procédures linéaires. Pour se restreindre à ces solutions, il imposa à son système – un vaste volume de molécules d'air – une source de chaleur externe variable selon la latitude et la longitude (modélisant la chaleur du Soleil dans l'atmosphère réelle).



2. L'ATTRACTEUR DE LORENZ, grossièrement visualisé *via* la trajectoire d'une unique orbite chaotique. Les sphères représentent les itérations des équations dites de Lorenz, calculées en utilisant les paramètres originaux des travaux fondateurs de Lorenz.

Reproduit avec l'autorisation de A. Mottelet [J. Campbell, Physics Today, vol. 66, pp. 27-33, 2013] © 2013 AIP Publishing LLC

Avec l'aimable autorisation de Sylvain Canet

En d'autres termes, il força son système à ne jamais atteindre d'équilibre.

Les modèles linéaires montrèrent en effet leurs limites. Mais Lorenz ne s'arrêta pas à cette conclusion. Ayant trouvé les solutions non périodiques de son modèle intéressantes en elles-mêmes, il les examina plus en détail. Il reproduisit les données, imprimant cette fois-ci les variables de sortie après chaque jour de météorologie simulée. Pour gagner de la place, il les arrondissait à la troisième décimale, même si les calculs sur l'ordinateur se poursuivaient avec une précision plus élevée. Il obtint alors un curieux résultat, qu'il relate ainsi dans son ouvrage *The essence of chaos* (1993) :

À un moment, j'ai décidé de répéter certains des calculs afin d'examiner plus en détail ce qui se passait. J'ai arrêté l'ordinateur, tapé une ligne de nombres qu'il avait imprimés plus tôt, et je l'ai relancé. Je suis sorti dans le couloir boire un café et je suis revenu environ une heure plus tard, tandis que l'ordinateur avait simulé environ deux mois de prévisions météorologiques. Les nombres qu'il avait imprimés n'avaient rien à voir avec les anciens.

J'ai tout de suite soupçonné une faiblesse d'un tube à vide ou un autre problème matériel, ce qui est fréquent, mais avant d'appeler le service technique, j'ai décidé de regarder à quel endroit l'erreur s'était produite, sachant que cela pourrait faciliter le dépannage. Au lieu d'une rupture soudaine, j'ai trouvé que les nouvelles valeurs répétaient au début les anciennes, puis divergeaient rapidement, d'abord d'une unité sur la dernière décimale, puis de plusieurs, puis sur l'avant-dernière décimale et, ensuite, sur celle d'avant encore.

En fait, les différences doublaient tous les quatre jours à peu près, jusqu'à ce que toute ressemblance avec le résultat initial s'efface au cours du deuxième mois. Cela suffisait pour que je comprenne ce qui s'était passé : les nombres que j'avais entrés n'étaient pas les nombres originaux exacts, mais les valeurs arrondies qui étaient apparues dans le document original imprimé. Ces erreurs par arrondi étaient les coupables ; et elles s'amplifiaient peu à peu jusqu'à dominer la solution.

Un papillon au Brésil, une tornade au Texas ?

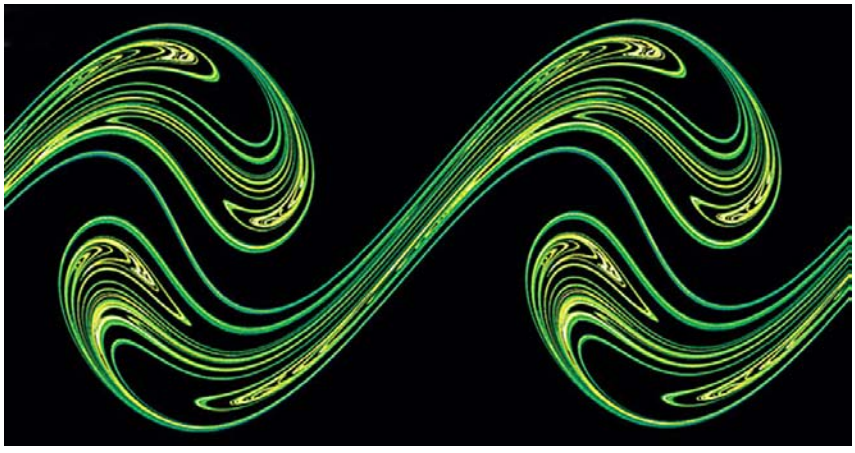
Avec son modèle, Lorenz a en fait observé ce que l'on nomme aujourd'hui la sensibilité aux conditions initiales – une propriété fondamentale du chaos. Dans l'espace des phases, ce phénomène a une signature : la distance entre deux trajectoires proches quelconques croît exponentiellement avec le temps. Ce comportement confère aux systèmes chaotiques un caractère aléatoire. Mais, comme Lorenz lui-même le notait, les apparences sont trompeuses : alors qu'à tout moment dans un système aléatoire (un lancer de dés, par exemple), il peut se produire une chose parmi deux ou plus, dans des systèmes chaotiques, les résultats sont déterministes. (Et, à strictement parler, c'est la même chose pour le lancer de dés.)

Lorenz comprit que si l'atmosphère avait le même comportement que son modèle, il était impossible de prévoir le temps à long terme. En 1972, lors d'un colloque de l'Association américaine pour l'avancement de la science, dans une conférence intitulée *Prédictibilité : le battement des ailes d'un papillon au Brésil déclenche-t-il une tornade au Texas ?*, il présenta sa découverte en utilisant le papillon comme métaphore d'une minuscule perturbation qui pourrait changer les conditions météorologiques. La métaphore prit et la sensibilité aux conditions initiales devint connue du grand public sous le nom de l'effet papillon.

Lorenz présenta les résultats de son modèle à 12 variables en 1960, lors d'un symposium à Tokyo. Il ne mentionna que brièvement l'effet inattendu des erreurs d'arrondi, pensant en parler dans un autre article. Mais la plupart de ses contemporains ne perçurent pas l'importance de ses résultats. De même, les travaux d'autres pionniers du chaos, tel le Japonais Yoshisuke Ueda qui fit en 1961 des expériences numériques analogues, passèrent inaperçus.

Lorenz publia sa découverte fortuite dans un article de mars 1963 intitulé *Écoulement non périodique déterministe*. Entretemps, il avait passé une bonne partie de son temps à rechercher le modèle le plus simple présentant une sensibilité aux conditions initiales, et avait abouti à un système d'équations différentielles non linéaires à trois variables, aujourd'hui connues sous le nom d'équations de Lorenz.

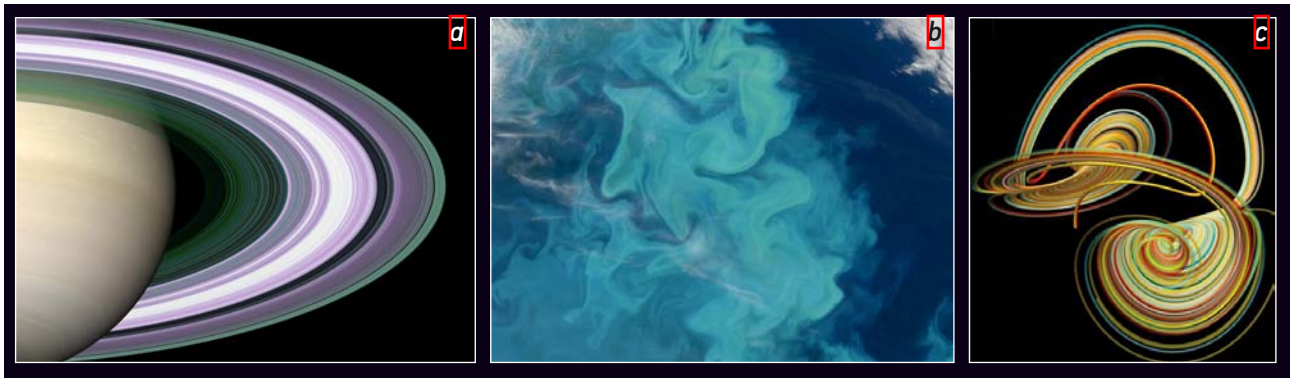
Comme le système à trois corps de Poincaré, les équations de Lorenz donnaient des trajectoires dans l'espace des phases qui ne se recouvrent jamais et n'engendrent pas des surfaces de dimension entière. Au lieu de cela, les trajectoires convergent vers une structure bornée de dimension non entière – fractale –, nommée attracteur de Lorenz *(voir la figure 2)*.



3. LES TRAJECTOIRES DANS L'ESPACE DES PHASES d'un pendule amorti forcé périodiquement (système physique dissipatif) convergent vers un attracteur chaotique dont les points sont tracés ici à des instants qui sont des multiples entiers de la période de forçage. L'étirement et le pliage des volumes dans l'espace des phases confèrent à l'attracteur une structure fractale.

© Pour la Science - n° 433 - Novembre 2013

Habibi | Ch | Ch | Ch | Gruiz | Chaotic Dynamics | Ch | Introduction Based | On Classical | Mechanics | Camondge Univ | Press | Cuvl



4. LE CHAOS SE MANIFESTE dans des contextes variés : les anneaux de Saturne, où les régions vides correspondent à des orbites chaotiques *(a)* ; les efflorescences de plancton (vues ici sur une image satellite de la mer de Barents) qui forment des structures semblables à des fractales en raison du comportement chaotique des écoulements dans le sillage d'une île *(b)* ; le champ géomagnétique de la Terre qui inverse ses pôles à des intervalles chaotiques irréguliers – un comportement visualisé dans l'espace des phases par cette simulation *(c)*.

Les attracteurs chaotiques – dits aussi étranges –, dont celui de Lorenz est un exemple, ont tendance à émerger quand un système dissipatif est régulièrement forcé à compenser la perte d'énergie, comme quand un enfant sur une balançoire accompagne le mouvement de ses jambes pour ne pas s'arrêter. Dans le cas du système de Lorenz, c'est la source de chaleur qui compense la perte d'énergie due à la viscosité du fluide.

Dès le début, Lorenz semble avoir saisi une caractéristique essentielle du chaos : il a reconnu non seulement que la non-périodicité implique une sensibilité aux conditions initiales, mais aussi que cette sensibilité est la cause profonde de la non-périodicité. De fait, on sait aujourd'hui que dans l'espace des phases, une structure chaotique telle que l'attracteur de Lorenz baigne dans un ensemble infini d'orbites périodiques instables : si l'on s'écarte un tant soit peu d'une telle orbite, la trajectoire du point représentatif du système devient chaotique.

La publication fut peu remarquée. Douze ans après, l'article était cité moins de 20 fois. Le tournant a eu lieu en 1975, notamment *via* sa citation dans l'article qui établit le nom du domaine : *Period three implies chaos*, de Tien-Yen Li et James Yorke, de l'Université du Maryland, aux États-Unis. Dès la fin des années 1980, non seulement la recherche sur le chaos a connu un essor considérable, avec des milliers de publications scientifiques à l'appui, mais elle est aussi familière aux non-scientifiques. Aujourd'hui, le chaos est aussi décrit dans les systèmes dynamiques conservatifs – dont l'énergie totale est conservée au fil du temps. Dans de tels systèmes, il ne peut y avoir d'attracteur : le chaos se manifeste comme une sorte de poussière en mouvement dans l'espace des phases (parfois nommée poussière de Cantor).

Le chaos se distingue des autres grandes théories de la physique. Contrairement à la relativité ou à la mécanique quantique, il ne décrit pas un phénomène physique particulier, mais apporte un ensemble de concepts et de méthodes pour analyser un comportement insoupçonné qui peut se manifester dans des disciplines diverses. Ces caractéristiques expliquent en partie l'indifférence avec laquelle ont été accueillies les premières indications du phénomène. Les premières confrontations avec le chaos ont eu lieu dans des disciplines disparates (mécanique céleste, mathématiques, ingénierie), les uns n'étant pas au courant des découvertes des autres.

Aujourd'hui, si nombre de problèmes liés au chaos demeurent irrésolus, tels que ses implications dans les systèmes quantiques et relativistes, la théorie du chaos est utilisée dans de multiples domaines : physiologie, informatique, finance, dynamique des fluides... Comme le résume l'annonce du prix de Kyoto que reçut Lorenz en 1991 : « [Le chaos déterministe] a amené l'un des changements les plus importants de notre vision de la nature depuis Isaac Newton. » ■

LES AUTEURS

Adilson E. MOTTER est professeur de physique et d'astronomie à l'Université du Nord-Ouest à Evanston, dans l'Illinois, aux États-Unis. David K. CAMPBELL est professeur de physique et de génie électrique et informatique à l'Université de Boston.

BIBLIOGRAPHIE

A. E. Motter et D. K. Campbell, **Chaos at fifty**, *Physics Today*, vol. 66, pp. 27-33, 2013.

Fifty years of chaos : applied and theoretical, *Chaos*, numéro spécial, vol. 22, 2012.

E. Ott, **Chaos in Dynamical Systems**, Cambridge University Press, 2002.

Y. Ueda, **The Road to Chaos-II**, Aerial Press, 2001.

J. Gleick, **La théorie du chaos**, Flammarion, 1999.

F. Diacu et P. Holmes, **Celestial Encounters : The Origins of Chaos and Stability**, Princeton Univ. Press, 1996.

E. N. Lorenz, **The Essence of Chaos**, Univ. of Washington Press, 1993.

A, avec l'aimable autorisation de la mission Cassini/NASA ; B, adapté de NASA Ocean Color Web ; C, avec l'aimable autorisation de Christophe Cassinga

La sélection de livres POUR LA SCIENCE

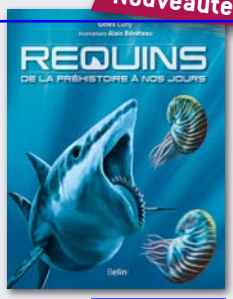
Découvrez la sélection du mois *Pour la Science* avec des ouvrages scientifiques amusants. Ne manquez pas la nouvelle collection « InfoGraphie » Bonne lecture !

Requins

De la préhistoire à nos jours

Gilles Cuny, Illustrations d'Alain Beneteau

Nouveauté



Les requins ? Ils sont presque à égalité avec les dinosaures dans l'imaginaire collectif. Ce livre original, illustré de 200 reconstitutions et photographies, écrit par LE spécialiste français des requins, est riche en informations scientifiques, en sensations et en surprises.

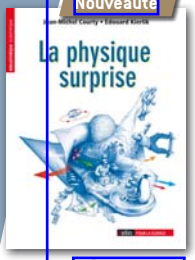
Éditions Belin 2013
224 pages – 27,90 euros – ISBN 978-2-7011-5423-7

Réf. 005423

La science amusante !

La physique surprise

Nouveauté



Réf. 075121

Jean-Michel Courty et Édouard Kierlik

Passionnés de physique et curieux de tout, Jean-Michel Courty et Édouard Kierlik vous invitent à de nouvelles promenades scientifiques dans un langage accessible, simple et imagé, sans jargon ni formules. Avec eux, laissez-vous surprendre par les phénomènes que vous avez déjà côtoyés sans même les remarquer.

Éditions Belin-Pour la Science 2013
224 pages – 26 euros – ISBN 978-2-8424-5121-9

50 expériences pour épater vos amis au jardin

Jacques Guichard, Kamil Fadel et Guy Simonin



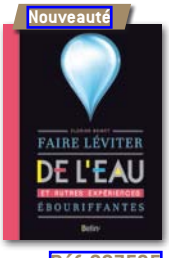
Réf. 090618*

Marcher sur l'eau, tirer la nappe sans faire tomber ce qui est sur la table, créer un double arc-en-ciel, transpercer un ballon sans qu'il explose… Voici 50 nouvelles expériences très faciles à réaliser avec, pour chacune, du matériel simple, le déroulé pas à pas, les gestes précis et les trucs pour la réussir à coup sûr. Et pour ceux qui veulent en savoir plus, des compléments scientifiques et des indications sur les lieux où l'on peut observer ces phénomènes à grande échelle.

Éditions Le Pommier 2013
168 pages – 18,90 euros – ISBN 978-2-7465-0618-3

Faire léviter de l'eau

Nouveauté



Réf. 007535

Florian Briant

Comment provoquer une tempête dans une bouteille ? Sauriez-vous allumer une ampoule sans la brancher, souffler une bulle géante ou bien extraire votre propre ADN ? Découvrez le plaisir de faire de la science, à travers des expériences ludiques et réalisables avec du matériel simple. À vous l'excitation du « labo » !

Éditions Belin 2013
128 pages – 21 euros – ISBN 978-2-7011-7535-5

Mathémagic

Nouveauté



Réf. 007686

David Acheson

De pi au pendule chaotique, de Pythagore à Andrew Wiles, embarquez pour une fascinante balade au pays des mathématiques ! De savoureuses illustrations achèveront de convaincre le profane comme l'initié qu'il s'agit là d'un des livres les plus réjouissants jamais écrit sur le sujet.

Éditions Belin 2013
176 pages – 15 euros – ISBN 978-2-7011-7686-4

En partenariat avec les éditeurs

Belin ÉDITEUR INDÉPENDANT DEPUIS 1777 — Le Pommier

Tous ces livres sont également disponibles en librairie.

Nouvelle collection

InfoGraphie, une collection d'ouvrages résolument graphiques, pour se faire son opinion sur de grandes questions de société ! Un concentré d'informations sous forme de textes, graphes, chiffres clés et avis d'experts.

Le Cannabis ?

Marina Julienne

Qui sont les consommateurs de cannabis ? Quels sont ses effets sur la santé ? Le cannabis peut-il aussi soigner ? Comment aider les jeunes à arrêter ? Qui sont les trafiquants de cannabis ? Etc.

Éditions Belin 2013
80 pages – 19 euros – ISBN 978-2-7011-6380-2



Réf. 006380

La Biodiversité ?



Réf. 005692

Lise Barnéoud

Depuis combien de temps les espèces disparaissent-elles ? Combien de terres sauvages sont-elles chaque année englouties sous l'asphalte ? Reste-t-il encore de la nature « sauvage » ? Etc.

Éditions Belin 2013
80 pages – 19 euros – ISBN 978-2-7011-5692-7

La Procréation assistée ?

Lise Barnéoud

Qui peut recourir à la PMA, en France et dans le monde ? Est-ce une solution pour éviter de transmettre une maladie génétique ? Comment débattre des problèmes éthiques et sociaux soulevés? Etc.

Éditions Belin 2013
80 pages – 19 euros – ISBN 978-2-7011-6144-0



Réf. 006144

À LIRE AUSSI...

Incertitudes sur le climat

Katia Laval et Guy Laval

Éditions Belin 2013
272 pages – 19 euros
ISBN 978-2-7011-7485-3



Réf. 007485

Une brève histoire du magnétisme

Jean-Paul Poirier et Jean-Louis Le Mouël

Éditions Belin 2013
224 pages – 18,50 euros
ISBN 978-2-7011-7532-4



Réf. 007532

Cendrillon et la mort des étoiles

Robert Gilmore

Éditions Le Pommier 2013
358 pages – 19 euros
ISBN 978-2-7465-0706-7



Réf. 090706*

Le Dernier fado de l'androïde

Hugues Bersini

Éditions Le Pommier 2013
256 pages – 17 euros
ISBN 978-2-7465-0705-0



Réf. 090705*

Pour commander ces ouvrages renvoyez le bon de commande ci-dessous

ou connectez-vous sur **www.pourlascience.fr/librairie**
(sauf références signalées par un * sur www.editions-belin.fr)

BON DE COMMANDE

POUR LA SCIENCE

À retourner accompagné de votre règlement à : Groupe Pour la Science • 628 avenue du Grain d'Or • 41350 Vineuil • Tél. : 0 805 655 255 • e-mail : pourlasciencevpc@daudin.fr

❑ Oui, je commande les livres présentés dans la sélection de *Pour la Science*. Je reporte les titres, prix et références à 6 chiffres correspondant aux ouvrages choisis :

Titre	Réf.	Prix
		€
		€
		€
		€
		€
Frais de port (4,90 € France – 12 € étranger)	+	€
Total à régler	=	€

J'indique mes coordonnées :

Nom :
Prénom :
Adresse :
C.P. : Ville :
Pays : Tél. : (Pour le suivi client facultatif)
E-mail : @ (Pour recevoir la newsletter Pour la Science (à remplir en majuscules))

Je choisis mon mode de règlement :

❑ par chèque à l'ordre de *Pour la Science*
❑ par carte bancaire
Numéro
Date d'expiration
Cryptogramme

Signature obligatoire

Belin ÉDITEUR INDÉPENDANT DEPUIS 1777 — Éditions Le Pommier

En application de l'article 27 de la loi du 6 janvier 1978, les informations ci-dessus sont indispensables au traitement de votre commande. Elles peuvent donner lieu à l'exercice du droit d'accès et de rectification auprès du groupe Pour la Science. **Par notre intermédiaire, vous pouvez être amené à recevoir des propositions d'organismes partenaires. En cas de refus de votre part, merci de cocher la case ci-contre ❑.**

DPL433 Offre valable jusqu'au 30/11/2013 dans la limite des stocks disponibles.

LOGIQUE & CALCUL

La quête du pavé apériodique unique

Les spécialistes des pavages et des quasi-cristaux espéraient depuis longtemps découvrir un pavé unique avec lequel le pavage du plan serait nécessairement non périodique. Un tel pavé a été trouvé… enfin, presque.

Jean-Paul DELAHAYE

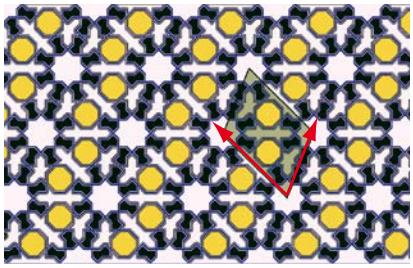
Des contraintes locales ont-elles des conséquences à l'infini ? La théorie des pavages apériodiques a examiné cette question et un domaine de la cristallographie en est né, amenant une interaction nouvelle des mathématiques et de la physique. Dernier résultat : un pavé décoré interdit à lui seul la périodicité ! Ainsi, les aspects ludiques et sérieux de la géométrie se mêlent pour donner de nouveaux outils aux physiciens.

Périodiques ou non ?

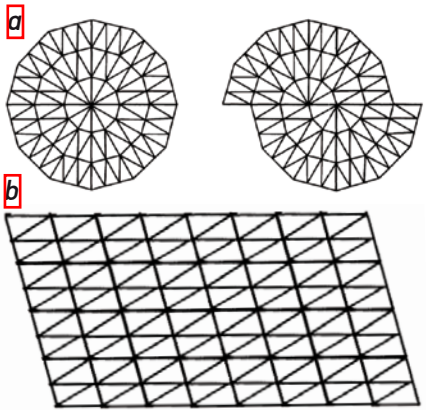
Dans un pavage périodique du plan, une partie finie, en forme de parallélogramme, permet de paver totalement le plan en utilisant deux translations (un pavage est périodique s'il existe deux vecteurs de translation non colinéaires qui le laissent invariant) *(voir la figure 1)*.

Sauf de rares exceptions, les pavages décoratifs utilisent des motifs périodiques, sans doute pour des raisons de facilité d'impression ou d'assemblage. Il est cependant facile de concevoir des motifs non périodiques qui pavent le plan. Deux exemples très simples utilisent un pavé en forme de triangle isocèle, dont l'angle au sommet est égal à 30° ; ce pavé peut bien sûr composer un pavage périodique, comme c'est le cas de tout triangle *(voir la figure 2)*.

On a longtemps pensé que tout ensemble de pavés qui permet de paver le plan autorise aussi un pavage périodique.



1. UN EXEMPLE DE PAVAGE PÉRIODIQUE du plan. Ce pavage peut être construit en dupliquant le parallélogramme en gris et en le translatant selon les deux vecteurs *(en rouge)* construits sur ses côtés.



2. DEUX EXEMPLES de pavage apériodique du plan *(a)* dont le pavé élémentaire est un triangle isocèle d'angle au sommet égal à 30°. Ce pavé permet aussi de recouvrir périodiquement le plan *(b)*.

Personne ne réussissant à démontrer cette affirmation, on s'est demandé s'il existait des ensembles de pavés conduisant uniquement à des pavages non périodiques, ce que nous dénommerons des « ensembles de pavés apériodiques ».

Si un tel ensemble existe, alors des contraintes géométriques locales (que des pavés se joignent parfaitement) imposent une propriété globale, l'apériodicité, ce qui n'est pas une évidence (c'est, d'ailleurs, impossible en dimension 1).

Le pavage, un problème indécidable

La solution est issue d'une question algorithmique. Les logiciens, dont en particulier Hao Wang, un confident de Kurt Gödel, se sont posé une question : un ensemble de formes géométriques élémentaires étant donné, peut-on savoir par un moyen systématique, c'est-à-dire par un algorithme, si oui ou non on pourra couvrir tout le plan avec des exemplaires de ces formes sans chevauchement et sans laisser de vide ? C'est le problème de la « décidabilité du pavage ». Il est aussi dénommé « problème des dominos », car une version simplifiée ne fait intervenir que des dominos carrés à côtés colorés que l'on dispose en tenant compte des couleurs (ce que l'on ramène à des pavés sans couleur en déformant les côtés, pour transformer les contraintes de couleur en contraintes de forme).

© Pour la Science - n° 433 - Novembre 2013

Si un ensemble fini donné de pavés ne pave pas le plan, on s'en aperçoit en un temps fini, mais non connu à l'avance : on essaie toutes les combinaisons pour paver un disque de rayon 1, puis 2, etc. Si l'ensemble de pavés ne permet pas de paver le plan, on finit par rencontrer un blocage (quand le disque n'est pavable avec aucune combinaison de pavés) et on sait alors qu'il n'existe pas de pavage possible.

De même, si l'on sait qu'un ensemble fini de formes pave le plan de manière périodique, on peut réaliser ce pavage en essayant des formes de plus en plus nombreuses de l'ensemble ; on découvre alors nécessairement comment réaliser le pavage périodique, car vient un moment où l'on tombe sur la partie finie qui recouvre le plan par translation.

Il résulte de ces deux résultats algorithmiques positifs que :

– soit « tout ensemble de pavés qui pave le plan peut le faire de manière périodique » et, dans ce cas, le problème du pavage est décidable (on essaie simultanément de paver des disques de plus en plus grands et de trouver un motif périodique composé de pavés pris dans l'ensemble, et on est assuré d'arriver à une conclusion en un temps fini) ;

– soit « il existe des ensembles de pavés qui pavent le plan, mais qui ne peuvent pas le paver de manière périodique ».

Dit autrement, si le problème du pavage est algorithmiquement indécidable, alors il existe des ensembles de pavés apériodiques.

À l'heure de Berger

La résolution de l'énigme est due à Robert Berger. En 1964, dans sa thèse de doctorat à l'Université Harvard, il montra qu'on peut simuler n'importe quel calcul en choisissant soigneusement un ensemble fini de pavés et en tentant d'en recouvrir le plan. Si la tentative de pavage n'est jamais bloquée, cela signifie que le calcul se poursuit indéfiniment ; si l'on tombe sur un blocage, cela signifie que le calcul est terminé. Comme on sait depuis les travaux d'Alan Turing de 1936 que l'arrêt d'un programme est indécidable, il résulte du système de

La fléchette et le cerf-volant de Penrose

Durant la décennie 1970-1980, le physicien britannique Roger Penrose découvrit des ensembles de pavés apériodiques divers : ils permettent de paver le plan, mais uniquement de façon apériodique. Il réussit à en proposer ne comportant que deux pavés.

L'ensemble *kite-dart* (cerf-volant et fléchette) est simple. Il n'y a que deux sortes de côtés dont le rapport des longueurs est le nombre d'or $\tau = (1 + \sqrt{5})/2$. Sans contraintes particulières, ces deux formes pavent le plan de manière périodique : on place la fléchette dans le creux du cerf-volant, ce qui donne un quadrilatère avec lequel on pave périodiquement le plan. En revanche, en colorant les sommets de la fléchette et du cerf-volant et en ne considérant comme acceptables que les pavages où on fait se correspondre des sommets de même couleur, on obtient uniquement des pavages apériodiques. La contrainte des sommets colorés se traduit en contrainte de forme en redessinant les côtés.

On obtient ainsi deux formes qui pavent le plan, mais qui ne le font que d'une manière non périodique : on a un ensemble apériodique à deux éléments. Une question non résolue aujourd'hui est si l'on peut trouver une forme unique qui pave le plan et qui ne le fait que de manière apériodique.

Voici quelques propriétés des pavages convenables obtenus avec la fléchette et le cerf-volant.

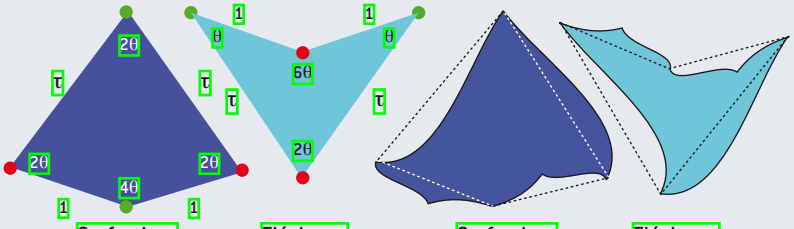
– Un pavage convenable ne possède jamais de symétrie de translation (non seulement il n'existe pas deux vecteurs de translation le laissant invariant, mais il n'en existe pas même un seul).

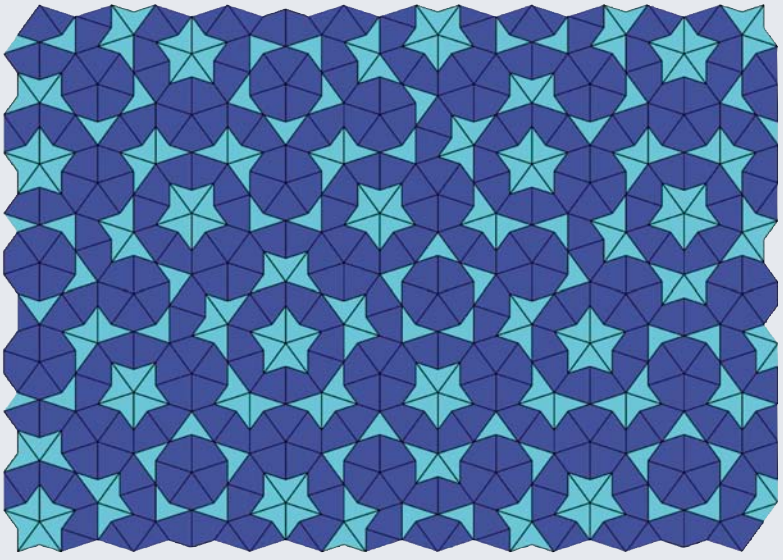
– Toute partie finie de tout pavage convenable se retrouve une infinité de fois dans tout autre pavage ; en ce sens, tous les pavages convenables se ressemblent.

– Il existe une infinité non dénombrable de pavages convenables possibles (tous apériodiques) ne se ramenant pas l'un à l'autre.

– Dans tout pavage convenable, il existe des parties finies arbitrairement grandes qui possèdent une symétrie de rotation d'ordre 5 (rotation d'angle $2\pi/5$).

– Dans tout pavage convenable, si l'on prend des disques de plus en plus grands, le rapport entre le nombre de cerfs-volants contenus dans le disque et le nombre de fléchettes tend vers le nombre d'or τ.





simulation de R. Berger que le problème du pavage est indécidable… et qu'il existe donc des ensembles de pavés apériodiques.

N'est-il pas remarquable que la démonstration de l'existence d'ensembles de pavés apériodiques provienne d'une préoccupation théorique concernant la décidabilité algorithmique, théorie que certains pensent sans intérêt pratique ?

Mais on attend des théoriciens qu'ils sachent traduire concrètement leurs préoccupations abstraites. R. Berger, exploitant sa démonstration d'indécidabilité, en a tiré la définition effective d'un ensemble de pavés forçant la non-périodicité. Dans sa thèse, il construisit un tel jeu de 20 426 formes. Dès lors, on commença à rechercher des jeux de pavés apériodiques comportant un plus petit nombre de formes de base.

R. Berger réduisit ses 20 426 pavés à 104. Donald Knuth (l'informaticien créateur du traitement de texte mathématiques *TeX*) amena alors le record à 92. D'année en année, des ensembles apériodiques plus petits ont été découverts. On passa ainsi à 40 (H. Läuchi en 1966), à 35 (Raphael Robinson en 1967), à 34 (Roger Penrose en 1973), à 16 (Robert Ammann en 1978). En 1971, Robinson découvrit un ensemble de six pavés apériodiques *(voir la figure 3a)*. La démonstration de l'apériodicité, dans ce cas comme dans bien d'autres, se fait en ajoutant des tracés colorés sur les pavés *(voir la figure 3b)*.

Quand on compose un pavage, ces tracés forment deux systèmes de carrés emboîtés, en orange et en vert *(voir la figure 3c)*. Les carrés ainsi créés sont de taille aussi grande qu'on le veut et on montre que c'est nécessairement le cas pour tout pavage réalisé avec les six pavés de Robinson. Si un pavage périodique était possible avec ces pavés, il y aurait une taille maximale de carrés. Le fait qu'il existe des carrés de taille aussi grande qu'on le veut prouve que tous

Des pavages plus ou moins compliqués

Quasi-périodique signifie presque périodique et cela suggère une hiérarchie des pavages non périodiques. Un pavage est quasi-périodique si, pour tout motif *M* qui y apparaît, il existe un entier *n* tel que *M* apparaît dans tout motif couvrant un disque de rayon *n*. Les pavages quasi-périodiques sont donc ceux qui font apparaître tout motif de manière régulière et pas trop espacée. Ils sont un peu moins simples que les pavages périodiques, mais leur désordre reste modéré. Tous les pavages réalisables avec les pavés de Penrose sont quasi-périodiques. Dans le cas de pavages quasi-périodiques de Penrose *(photographies ci-dessous)*, tout motif fini que l'on trouve dans l'un se retrouve dans tout autre. Il est facile de concevoir des pavages qui ne sont pas quasi-périodiques. C'est par exemple le cas des deux pavages non périodiques à base de triangles ayant une sorte de point central *(voir la figure 2)*. Une question naturelle se pose : puisqu'on peut forcer la non-périodicité, peut-on forcer la non-quasi-périodicité ?

Autrement dit : peut-on trouver un ensemble de pavés qui pavent le plan, mais seulement en dessinant des pavages qui ne sont pas quasi-périodiques (contenant donc une sorte de grand désordre) ? La réponse est négative : si un ensemble de pavés recouvre le plan, alors il peut en particulier le faire de manière quasi-périodique. Ce beau théorème, que le chercheur français Bruno Durand a établi en 1999, signifie que le type de désordre exprimé par la propriété de non-quasi-périodicité ne peut pas être forcé par des règles locales.

La non-quasi-périodicité est une forme de désordre, mais la non-calculabilité aussi, et cette fois on peut la « forcer ». William Hanf et Dale Myers ont en effet démontré en 1974 qu'il existe des ensembles de pavés qui pavent le plan, mais ne le font que de manière non calculable : aucun algorithme n'indique comment le faire ; pour paver le plan, il faut procéder au hasard et avoir de la chance !

Dans cette veine, un autre résultat a été démontré en 2008 par B. Durand, Leonid Levin et Alexander Shen. Il pousse à son extrême cette étude du forçage par un ensemble fini de pavés. Son énoncé indique qu'il existe des ensembles de pavés qui forcent une croissance rapide de la complexité des dessins produits par tout pavage.

Tout pavage possible est tel que pour en décrire un motif couvrant un disque de rayon *n*, il faut un programme de longueur au moins proportionnelle à *n* (la complexité de Kolmogorov du dessin est au moins proportionnelle à *n*, c'est-à-dire au moins égale à *Cn* pour une certaine constante *C*). Un tel ensemble de pavés force donc une grande complexité des dessins.





© Pour la Science - n° 433 - Novembre 2013

les pavages du plan réalisables avec les six formes de Robinson sont non périodiques.

Grâce à R. Penrose, ce bel ensemble de six pavés apériodiques fut d'abord battu par un ensemble de trois pavés, puis par un ensemble de deux pavés *(voir l'encadré page 125)*.

En 1979, R. Penrose, fier de sa découverte, fit breveter les motifs produits par ses pavés apériodiques – une procédure d'appropriation de formes géométriques sur laquelle on peut s'interroger ! En 1997, l'épouse de R. Penrose acheta par hasard des rouleaux de papier hygiénique produits par la firme *Kleenex* dont le motif décoratif était l'un des pavages brevetés. Sans doute furieux de voir sa merveilleuse découverte mathématique chargée d'une tâche indigne, R. Penrose engagea des poursuites contre la firme indélicate. Il gagna. Carl Morris raconte qu'en 2002, un de ses professeurs de mathématiques de l'Université de Cardiff, qui avait pu acheter quelques précieux rouleaux avant leur retrait des magasins, offrait une feuille décorée aux étudiants obtenant les trois meilleures notes à ses examens.

Des pavages aux quasi-cristaux

Les pavages apériodiques ont longtemps été considérés comme une simple curiosité mathématique sans applications autres que ludiques ou esthétiques. Cependant, en 1982, Dan Shechtman, chercheur israélien alors invité à l'Institut américain des normes et de la technologie (le NIST), s'aperçut qu'un alliage d'aluminium et de manganèse produit, quand il est éclairé par des rayons X monochromatiques, des figures de diffraction ayant une symétrie d'ordre cinq, c'est-à-dire analogue à celle d'un pentagone régulier *(voir la figure 4)*. C'était fort étonnant, car aucune structure cristalline périodique ne peut produire de telles figures de diffraction.

Cette découverte des « quasi-cristaux » était si inattendue que le chercheur a eu du mal à publier sa découverte. Finalement admise et confirmée, elle conduira en 1992 à la modification de la définition des cristaux par l'Union internationale de cristallographie, définition qui ne retient que le caractère discret de la figure de diffraction des rayons X, sans imposer d'ordre aux symétries observées. Et en 2011, D. Shechtman reçut le prix Nobel de physique pour sa découverte.

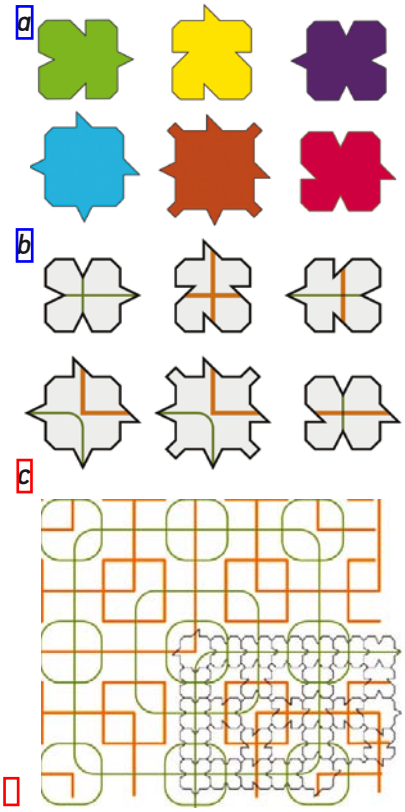
Les premiers quasi-cristaux étaient tous artificiels, mais en 2009 un quasi-cristal naturel a été découvert dans les montagnes de Koriakie, en Russie. Son origine extraterrestre ne semble faire aucun doute !

Le lien avec les pavages apériodiques, dont certains sont construits à partir de pavés ayant une symétrie d'ordre 5, ou liés au pentagone régulier (comme c'est le cas pour le pavage *kite-dart* de Penrose à deux pavés), s'est révélé essentiel pour comprendre la nature des quasi-cristaux. Des centaines d'articles scientifiques ont été publiés sur le sujet et, aujourd'hui, tout livre de cristallographie consacre des pages aux quasi-cristaux et aux pavages apériodiques. Parmi les points importants, mentionnons la compréhension que les pavages quasi-périodiques proviennent parfois de la projection de structures périodiques appartenant à des espaces de dimension supérieure.

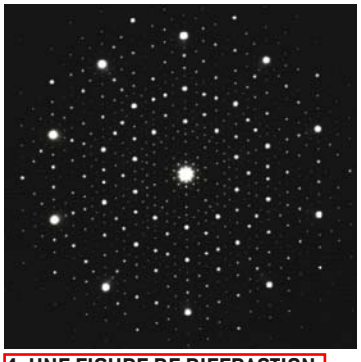
En informatique, les pavages apériodiques se sont révélés intéressants dans le domaine du graphisme et de la synthèse d'images. Pour engendrer des textures ou pour disposer des séries de nombreux objets dans une image, la solution la plus simple consiste à recopier périodiquement une structure de taille limitée. La répétition périodique, trop visible, donne un aspect peu naturel au résultat *(voir la figure 5)*. D'où l'idée d'utiliser un pavage apériodique qui évite d'avoir à utiliser de l'aléa ou du pseudo-aléa, comme l'ont montré Adrien Peytavie, Éric Galin, Jérôme Grosjean et Stéphane Merillou en 2009.

Forcer l'apériodicité avec un pavé unique ?

Indépendamment des applications à la cristallographie et à l'infographie, une question simple et fondamentale est restée posée depuis la découverte par R. Penrose d'un ensemble de deux pavés forçant la non-périodicité : peut-on forcer la non-périodicité avec un seul pavé ? Ce problème porte le



3. LES SIX PAVÉS DE ROBINSON *(a)* ne peuvent construire qu'un pavage apériodique. Pour le prouver, on leur superpose des tracés colorés *(b)* qui, lorsqu'on assemble les pavés *(c)*, forment des carrés aussi grands que l'on veut.



4. UNE FIGURE DE DIFFRACTION présentant une symétrie d'ordre 5.

nom de « problème *ein stein* » (une pierre en allemand, clin d'œil à Einstein).

En dimension 3, c'est possible, comme Peter Schmitt et John Conway l'ont montré assez facilement au début des années 1990 à l'aide d'un biprisme *(voir la figure 6)*. Chaque couche est périodique, mais en choisissant bien l'angle créé entre les couches (angle déterminé par la forme du biprisme), l'ensemble du pavage de l'espace n'est pas périodique.

En dimension 2, le polygone coloré découvert en 1996 par l'Allemande Petra Gummelt, de l'Université de Greifswald, est une petite merveille assez proche de la solution : il s'agit d'un décagone régulier dont la surface est colorée avec deux couleurs *(voir l'encadré ci-contre)*. On cherche à recouvrir le plan avec le décagone. On autorise les chevauchements, à condition que les parties chevauchantes soient colorées de la même façon, ce qui est une contrainte locale proche de la contrainte de recouvrement sans chevauchement.

Des solutions avec chevauchements

On montre que tout le plan peut ainsi être recouvert et qu'alors, c'est nécessairement de manière apériodique. Des liens ont été établis entre ce décagone et les pavages de Penrose. De plus, ce décagone est d'une utilité directe pour comprendre certains quasi-cristaux. Malheureusement, les chevauchements autorisés interdisent qu'on le considère comme la solution attendue au problème *ein stein*.

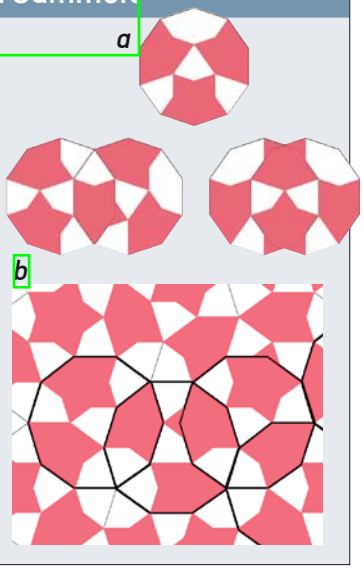
En 2010, Joan Taylor, une habitante de la lointaine Tasmanie (île célèbre pour son loup disparu), après de longs essais s'étalant sur plusieurs années, mit au point un hexagone coloré et des règles locales pour paver le plan avec des copies de cet hexagone. Elle prouva que tout pavage qui respecte les règles est non périodique *(voir l'encadré page ci-contre)*.

Le fait que les règles soient locales a laissé espérer un moment qu'on pourrait, comme pour les tuiles de Penrose, traduire les règles d'assemblage en déformant les côtés de l'hexagone pour obtenir un pavé unique (sans dessin, ni couleur) dont la seule règle « recouvrir sans chevauchement ni espace vide » forcerait la non-périodicité. Personne n'a réussi cette traduction, que l'on considère maintenant impossible.

Les tentatives de traduction ont tout de même donné deux « quasi-solutions ». La première propose un pavé unique (sans dessin ni couleur) en plusieurs morceaux disjoints considérés comme liés rigidement comme par une force magique, qui pave le plan et ne le fait que de manière apériodique. La seconde introduit une épaisseur (non constante) et donne une sorte de pavé tridimensionnel qui pave le plan uniquement de façon apériodique. Ce succès remarquable laisse malheureusement un goût d'inachevé et l'on s'interroge aujourd'hui pour savoir si, en définitive, un « pavé pur » (sans dessin ni couleur) forçant la non-périodicité est possible. Une course est donc lancée entre ceux qui tentent encore de trouver un pavé unique apériodique et ceux qui tentent de démontrer que ce pavé n'existe pas. ■

Le polygone de Petra Gummelt

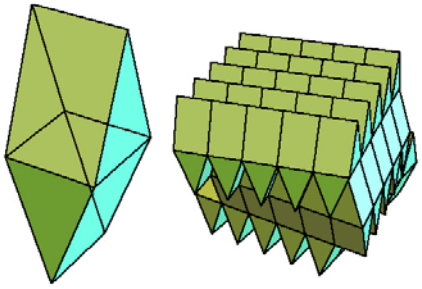
Le polygone coloré de Petra Gummelt est presque une solution de l'énigme du pavé unique apériodique (problème *ein stein*). C'est bien une forme qui force la non-périodicité, mais, pour l'utiliser, on change un peu les règles du jeu. On cherche à recouvrir le plan (aucun espace ne doit être oublié) par des copies du décagone en s'imposant que les parties colorées de deux décagones qui se chevauchent (dans ce cas, c'est autorisé) soient de la même couleur. Cette règle simple locale force les recouvrements du plan à être non périodiques. C'est un exemple aussi simple que ceux de Penrose (avec lesquels il n'est pas sans liens) de règle locale ayant un effet à l'infini.



Le décagone de Petra Gummelt avec deux façons convenables de faire se chevaucher des copies de ce décagone *(a)* et un exemple de recouvrement du plan par des copies se chevauchant convenablement *(b)*.



5. UNE TEXTURE FORMÉE avec des éléments disposés périodiquement *(en haut)* a un aspect peu naturel, contrairement à une disposition apériodique *(en bas)*.



6. POUR L'ESPACE À TROIS DIMENSIONS, Peter Schmitt et John Conway ont montré qu'avec un biprisme de forme bien choisie, on peut remplir tout l'espace, mais uniquement de façon apériodique. L'analogue en deux dimensions existe-t-il ?

Aider la science et simuler le hasard

Une expérience est organisée par l'Algorithmic Nature Group qui réunit des chercheurs français (dont l'auteur de cette rubrique), espagnols, hollandais et mexicains. Le but est d'évaluer l'aptitude des humains à reconnaître et à simuler le hasard. Le

© Pour la Science - n° 433 - Novembre 2013

SciLogs

La nouvelle communauté de blogueurs scientifiques francophones

Grâce à ***Scilogs.fr***, dialoguez avec des chercheurs et acteurs du monde des sciences aux points de vue riches et variés. De blog en blog, explorez la science avec ceux qui la font, la décryptent, l'analysent : résultats de recherche, applications pratiques, perspectives insolites, débats éthiques et politiques… Pour vivre la science au quotidien !

Best of Bestioles
Loïc Mangin
Rédacteur en chef adjoint
Pour la Science

Cerveau aux hormones
William Rostène et Jacques Epelbaum
Neuroscientifiques, directeurs de recherche à l'Inserm

Signal sur bruit
Richard Taillet
Astrophysicien, professeur chercheur au LAPTh

Complexités
Jean-Paul Delahaye
Mathématicien, informaticien, et auteur de la rubrique *Logique et calcul* dans *Pour la Science*

Retrouvez aussi les autres communautés « SciLogs » : déjà plus 140 blogueurs scientifiques à l'international !

Psycho Info
Jérôme Palazzolo
Médecin psychiatre, à Nice et à l'Université Internationale Senghor, à Alexandrie, en Égypte

Alterscience
Alexandre Moatti
Polytechnicien, chercheur en histoire des sciences au laboratoire SPHERE

Prospective Spatiale
Christophe Bonnal
Expert en astronautique, président des commissions « Débris spatiaux » à l'IAA et l'IAF

Connectez-vous maintenant pour découvrir www.Scilogs.fr !

Suivez les dernières actualités des blogueurs également sur les réseaux sociaux

Proposé par POUR LA SCIENCE

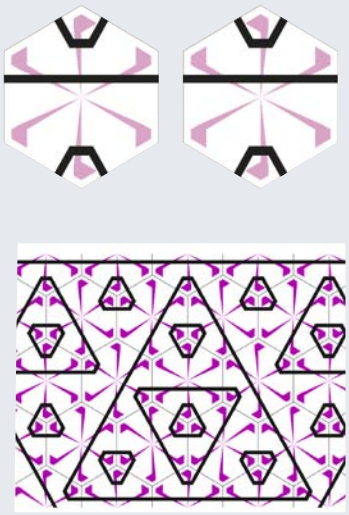
Le pavé de Joan Taylor

Le pavé de Taylor est une simple tuile hexagonale décorée. On peut la retourner, mais les motifs décoratifs sont les memes des deux cotes, comme si la tuile était en verre et les décorations incorporées dans le verre.

Les contraintes à respecter pour paver le plan sont les suivantes :

– les tuiles doivent être mises côté contre côté ;
– les traits noirs doivent se joindre ;
– lorsqu'on considère le côté d'un pavé quelconque, les deux dessins violets sur les tuiles dont le diamètre est aligné avec le côté doivent avoir des orientations identiques.

La démonstration que tout pavage respectant les règles précédemment édictées est apériodique se fait, comme pour les pavés de Robinson *(voir le texte de l'article)*, en considérant les triangles de toutes tailles que dessinent les traits noirs sur les tuiles jointes.



L'AUTEUR



J.-P. DELAHAYE est professeur à l'Université de Lille et chercheur au Laboratoire d'informatique fondamentale de Lille (LIFL).

BIBLIOGRAPHIE

• J.-Y. Lee et R. Moody, **Taylor-Socolar hexagonal tilings as model sets,** *Symmetry*, vol. 5, pp. 1-46, 2013.

• J. Socolar et J. Taylor, **Forcing non periodicity with a single tile,** *The Mathematical Intelligencer*, vol. 34(1), pp. 18-27, 2012.

• J. Socolar et J. Taylor, **An aperiodic hexagonal tile,** *J. Comb. Theory A*, vol. 118, pp. 2207-2231, 2011.

• J. Taylor, **Aperiodicity of a functional monotile,** 2010, www.math.uni-bielefeld.de/sfb701/preprints/view/420

• A. Musesti et M. Poalini, **The Penrose Tessellation,** 2010 [magnifique film sur les pavés apériodiques de Penrose] : http://penrose.dmf.unicatt.it/

• W. Steurer et S. Deloudi, **Crystallography of Quasicrystals,** Springer, 2009.

• B. Durand *et al.*, **Complex tilings,** *The Journal of Symbolic Logic*, vol. 73(2), pp. 593-613, 2008.

test prend trois ou quatre minutes. Merci d'avance d'y participer (http://complexitycalculator.com/hrng/).

Si vous acceptez de donner une adresse électronique, une fois l'expérience terminée et les données recueillies interprétées, vous recevrez un message avec les résultats et un lien vers l'article scientifique qui en sera tiré.

A propos de la persistance des nombres

Yves Césari m'a fait parvenir la solution de plusieurs questions posées dans l'article sur la persistance des nombres (Pour la Science n° 429, juillet 2013). Il s'agit des conjectures concernant le nombre de cas où la suite tombe sur un point final donné. La première affirme que les seuls entiers dont le point final (dans le calcul de la suite persistante) est 1 sont les nombres repunit 1, 11, 111, ... Y. Césari propose les deux résultats suivants :

1) L'ensemble des nombres dont le chiffre des unités est 1, 3, 7 ou 9, et dont le chiffre des dizaines est pair, est stable par produit.

2) Les nombres obtenus par produits itérés de 3 et 7 ont pour chiffre des unités 1, 3, 7 ou 9, et un chiffre des dizaines pair.

Ces énoncés limitent l'inventaire et le dénombrement de tous les nombres qui ont pour point final 1, 3, 7 ou 9, et conduisent à la solution des conjectures proposées au sujet des colonnes du tableau de la page 83.

Une autre solution pour le problème des repunits a été envoyée par Yves Lohmann. Marc Lapierre a aussi proposé un nombre qui, pour la version d'Erdős du problème de la persistance, n'atteint un point stable qu'après 21 étapes (au lieu de 18 pour le record cité dans l'article).



7. UN HOMMAGE À MARTIN GARDNER, réalisé par Bruce Torrence avec un pavage non périodique constitué de cerfs-volants et de fléchettes convenablement colorés. M. Gardner introduisit ces pavages dans sa rubrique de *Scientific American* en janvier 1977. Ce pavage par des fléchettes et des cerfs-volants a une symétrie d'ordre 5.

© Pour la Science - n° 433 - Novembre 2013

Vient de paraître

Les sens trompés

Des anomalies du cerveau aux comportements étranges

par Patrick Verstichel



Cet ouvrage est un recueil des articles que le neurologue Patrick Verstichel a publiés dans la rubrique « Le cas clinique » du magazine *Cerveau & Psycho*. En tant que médecin clinicien, il est confronté à des personnes qui présentent des symptômes inquiétants, surprenants, souvent inexpliqués. Après un examen clinique minutieux et les indispensables scanners IRM, il pose son diagnostic et propose une prise en charge adaptée.

Patrick Verstichel mène ses enquêtes médicales et scientifiques avec perspicacité. En nous présentant ces patients et leurs maladies sous forme de petites histoires, il nous fait comprendre comment fonctionne le cerveau, et pourquoi, parfois, il dysfonctionne. Un voyage étonnant au cœur du cerveau !

Éditions Belin/Pour la Science 2013
160 pages • 25 euros • ISBN 978-2-84245-111-0

BON DE COMMANDE

Cerveau & Psycho

à découper ou à photocopier et à retourner accompagné de votre règlement à : Groupe Pour la Science • 628 avenue du Grain d'Or • 41350 Vineuil • Tél. : 0 805 655 255 • e-mail : pourlasciencevpc@daudin.fr

❑ Oui je commande l'ouvrage *Les sens trompés* de Patrick Verstichel.

..... exemplaires × 25 € (Réf. 075111) = ... €
Frais de port (4,90 € France – 12 € étranger) + ... €
TOTAL À REGLER = ... €

Je choisis mon mode de règlement :
❑ par chèque à l'ordre de *Pour la Science*
❑ par carte bancaire
Numéro
Date d'expiration
Cryptogramme
(les 3 chiffres au dos de votre CB)
Signature obligatoire

J'indique mes coordonnées :
Nom :
Prénom :
Adresse :
C.P. : Ville :
Pays : Tél. : (pour le suivi client (facultatif))

En application de l'article 27 de la loi du 6 janvier 1978, les informations ci-dessus sont indispensables au traitement de votre commande. Elles peuvent donner lieu à l'exercice du droit d'accès et de rectification auprès du groupe Pour la Science. Par notre intermédiaire, vous pouvez être amené à recevoir des propositions d'organismes partenaires. En cas de refus de votre part, merci de cocher la case ci-contre ❑.

E-mail pour recevoir la newsleter *Cerveau & Psycho* (à remplir en majuscules).
@

VERPL433
Offre valable jusqu'au 31.12.2013 dans la limite des stocks disponibles

SCIENCE & FICTION

Le sexe chez les extraterrestres

Les extraterrestres existent – du moins dans la science-fiction ! –, mais comment ont-ils été conçus ? Puisque les informations sur cette question sont rares, il ne reste qu'à essayer de les déduire des mécanismes que l'évolution a sélectionnés… sur Terre.

J.-Sébastien Steyer et Roland Lehoucq

Les histoires de science-fiction ne s'attardent pas souvent sur le sexe ou la reproduction des extraterrestres. Certaines œuvres évoquent pourtant des coups de foudre intergalactiques ou des kidnappings et des reproductions forcées. Comme souvent, les extraterrestres sont le reflet de nos peurs et de nos fantasmes. D'ailleurs, les hybrides entre humains et extraterrestres posent d'intéressantes questions sur la classification et l'évolution des espèces.

Si la reproduction de certains extraterrestres a été décrite en détail – par exemple, les cycles biologiques et les différents stades de croissance d'Alien –, ce n'est pas le cas de la plupart des autres personnages de science-fiction : même lorsque le troisième type est de forme humanoïde, très peu de détails concernent sa reproduction, qu'elle soit sexuée ou non.

Rappelons que sur Terre, le sexe n'est pas obligatoire pour perpétuer l'espèce : la plupart des organismes (les bactéries, des plantes, des champignons, mais aussi des vers plats ou annelés, etc.) se reproduisent par multiplication « végétative », c'est-à-dire par simple division cellulaire entraînant une duplication du patrimoine génétique porté par les chromosomes. Par rapport à la reproduction sexuée, ce type de reproduction paraît plus simple : il suffit d'un petit bourgeon ou d'un greffon. La multiplication végétative serait aussi plus efficace en termes de nombre de rejetons. En revanche, le brassage génétique est plus limité, car les descendants ne sont que des clones de l'organisme initial. C'est aussi le cas de la parthénogenèse, où une unique femelle non fécondée – souvent nommée reine – donne naissance à des descendants (mâles ou femelles) tous identiques. Ce type de reproduction adopté par certaines plantes, insectes, ou lézards, est souvent à l'origine de sociétés matriarcales complexes.



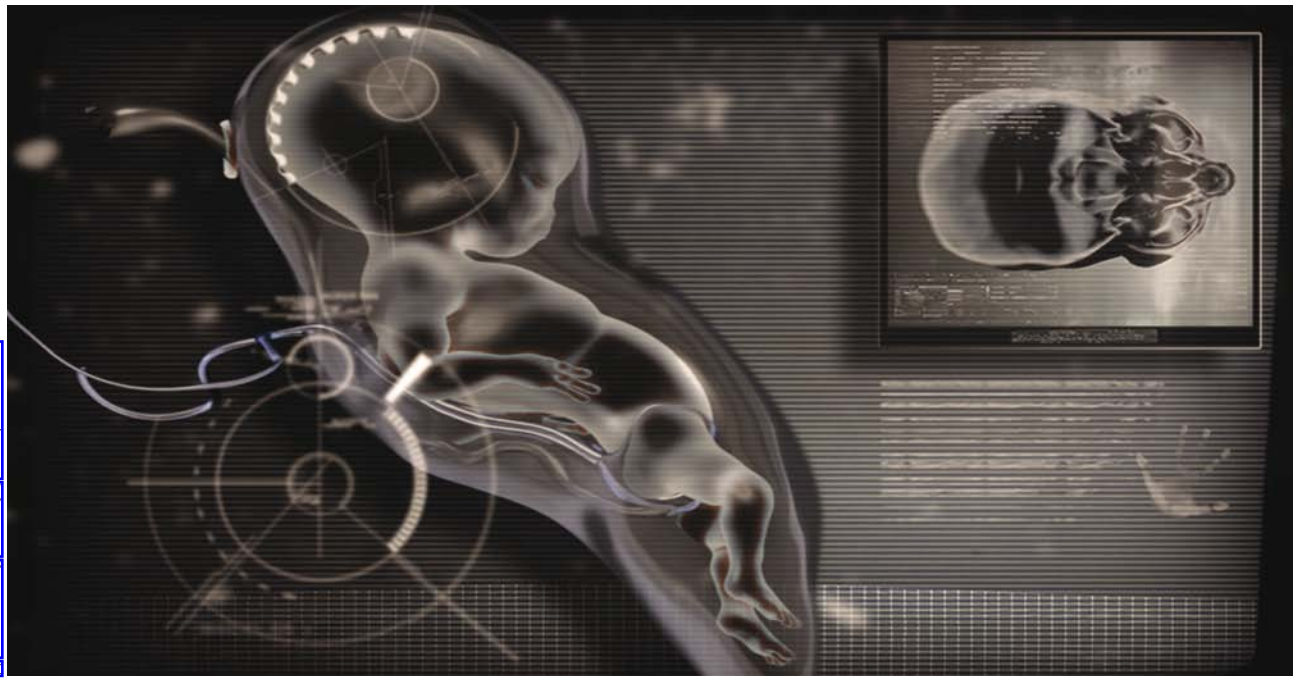
1. CAPITAINE SPOCK *(au second plan à droite)* **est l'hybride de la science-fiction le plus connu. Ce serait un « vulcarrien ».**

Homme et femme à la fois

Qu'il s'agisse de parthénogenèse et de multiplication végétative, les clones sont souvent obtenus par voie naturelle. Lorsque le clonage devient artificiel, apparaissent les craintes que la méthode ne soit un jour appliquée aux humains à des fins eugéniques : dans l'épisode II de *Star Wars* (*L'Attaque des Clones*, George Lucas, 2002), les Kaminoans, créatures expertes des biotechnologies, fabriquent secrètement une armée de clones à partir des gènes d'un redoutable chasseur de primes. Plus proche de nous, *The Island* (Michael Bay, 2005) dépeint une société souterraine, aseptisée et prétendue idéale, composée d'humains clonés, conditionnés et « élevés » comme banques d'organes destinés aux riches vivant en surface.

Quand la reproduction est sexuée, elle n'est pas forcément binaire (mâle ou femelle), comme en témoignent de nombreux hermaphrodites sur Terre ; plantes à fleurs (avec leurs étamines – organes mâles – et

© Pour la Science - n°433 - Novembre 2013



© 2013 Marc Boutin marcboutin possess-productions.com

2. LA REPRODUCTION ARTIFICIELLE évoquée dans diverses œuvres de science-fiction attise nos craintes, notamment celle de clones humains destinés à fournir des organes, ou celle de l'eugénisme, sélection des embryons sur des critères inacceptables.

leur pistil – organe femelle – sur le même pied), animaux terrestres (tel l'escargot) ou marins (le poisson clown commence sa vie comme mâle avant de devenir femelle). Ce poisson a-t-il inspiré John Varley pour son roman *Le Système Valentine* (2003), où un artiste original et génétiquement modifié change de sexe à volonté pour jouer seul *Roméo et Juliette* ? La rencontre du troisième type est aussi celle du troisième sexe dans *Sexomorphose* du Français Ayerdhal (1994), où Aimelin(e), humain capable de changer de sexe, part à la recherche de son identité.

Le sexe étant aussi organe de plaisir, certains auteurs n'hésitent pas à imaginer des scènes torrides entre humains et extraterrestres ! Philip José Farmer est l'un des pionniers de cette science-fiction « érotique » : dans *Les Amants étrangers* (1952) – nouvelle refusée par ses éditeurs et finalement adaptée en roman en 1961 –, l'auteur dépeint les liaisons dangereuses entre un homme fuyant une théocratie implantée sur Terre et une belle extraterrestre de forme humaine, mais de nature insectoïde. Avec ce plaidoyer pour le métissage intergalactique, Farmer utilise, en pleine Amérique puritaine, le sexe comme arme de critique sociale.

Ainsi la forme, le type, le nombre et l'utilisation des organes génitaux varient en fonction des mondes et des espèces : sur Terre, les « génitalias » des insectes sont tellement particuliers qu'ils servent à identifier les espèces ! Sur les autres planètes, tout est permis à partir du moment où les tailles des organes sont compatibles.

Et quand, de ces liaisons intergalactiques, naissent des hybrides, ceux-ci jouissent de différents statuts : l'enfant stellaire, symbole de paix dans la série *V* (Kenneth Johnson, 1983-1985), est le fruit des amours d'une résistante humaine et d'un « lézard » venu de Sirius pour envahir la Terre. Mais l'hybride le plus célèbre de la science-fiction, et le plus adulé, reste Spock, de la série *Star Trek*, symbole du métissage interplanétaire et du respect d'autrui.

Un hybride fécond

De mère terrienne et de père vulcain, il pose d'intéressantes questions en termes de phylogénie et de classification des espèces : comment nommer Spock ? Il n'y a pas de règles en zoologie, bien que la tendance consiste à contracter les syllabes des noms des parents : ainsi le « zébrâne », issu de l'âne et du zèbre (tous deux appartenant au genre *Equus*), est un hybride naturel que Darwin lui-même a observé en Afrique australe. Précisons que le préfixe du nom hybride provient parfois du parent mâle : ainsi le tigron est un croisement entre un tigre (*Panthera tigris*, originaire d'Asie) et d'une lionne (*Panthera leo*, originaire d'Afrique), tandis que le ligre est le rejeton d'un lion et d'une tigresse.

Ainsi Spock pourrait être un « vulcarrien ». Notons aussi qu'il a eu un fils : c'est donc un hybride naturel fécond. Cela montre (du moins d'après la biologie terrienne classique) que les gènes des parents de cet hybride sont compatibles. Vulcains et Terriens appartiendraient donc, soit à la même espèce, soit au même genre. Supposons qu'*Homo sapiens* ait priorité sur l'espèce vulcaine. Tous les Vulcains, père de Spock compris, seraient donc des *Homo sapiens vulcanensis* (s'ils sont de la même espèce) ou des *Homo vulcanensis* (s'ils appartiennent au même genre). Reste à expliquer la présence du genre *Homo* sur la planète Vulcain... Les Vulcains seraient-ils des humains génétiquement modifiés, adaptés aux conditions de vie sur Vulcain ? À moins que nous ne soyons des Vulcains installés sur Terre ?

Cette dernière idée évoque l'hypothèse, plus sérieuse, de la panspermie, selon laquelle les premières formes de vie seraient apparues sur Terre à partir d'éléments extraterrestres, par exemple des acides aminés apportés de l'espace dans certaines météorites. ■

J.-S. STEYER est paléontologue au CNRS-MNHN, à Paris. R. LEHOUCQ est astrophysicien au CEA, à Saclay.

Retrouvez les articles de J.-S. Steyer et R. Lehoucq sur www.pourlascience.fr

ART & SCIENCE

Promenade dans un paysage… épigénétique

Lors du développement embryonnaire, le destin de chaque cellule est dicté à la fois par les gènes et par l'environnement. On peut représenter ces contraintes par un paysage de pics et de vallées où se déplace la cellule. Une vidéo a été réalisée pour illustrer cette métaphore.

Loïc MANGIN

EpiGeneSys est un réseau d'excellence qui regroupe 150 laboratoires de recherche à travers l'Europe. Financé par le 7e programme cadre de la Communauté européenne, il a son siège à l'Institut Curie, à Paris. Son objectif est de bâtir un pont entre la recherche dans le domaine de l'épigénétique, la biologie des systèmes et le grand public. Mais qu'est-ce que l'épigénétique ? Elle correspond à la façon dont l'environnement et l'histoire individuelle influent sur l'expression des gènes. En d'autres termes, si l'ADN qui contient les gènes est un livre, l'épigénétique en est l'interprétation par la cellule.

Après une collaboration avec Geneviève Almouzni, directrice du Centre de recherche de l'Institut Curie, Paul Harrison, du Centre de recherches visuelles de l'Université de Dundee, en Écosse, est aujourd'hui en résidence artistique auprès de *EpiGeneSys*. Il s'intéresse aux méthodes de représentation des idées en biologie et, pour ce faire, développe des méthodes alternatives pour montrer et faire comprendre comment fonctionnent les êtres vivants.

Avec Mhairi Towler, de la société de production *Vivomotion*, également installée à l'Université de Dundee, et l'artiste Link Li, il s'est attelé au concept de « paysage épigénétique ». De quoi s'agit il ? Lors du développement embryonnaire, les cellules se divisent et s'engagent progressivement dans des voies de différenciation : telle cellule deviendra un neurone, telle autre un hépatocyte, une cellule musculaire, un globule blanc… À chaque type cellulaire correspond un profil d'expression des gènes. De fait, toutes les cellules disposent du même génome, mais, par exemple, seules les cellules *bêta* du pancréas fabriquent de l'insuline. Ce profil d'expression relève de l'épigénétique. En effet, grâce à des modifications chimiques de l'ADN lui-même et des protéines autour desquelles il s'enroule (les histones), certains gènes sont accessibles à la machinerie cellulaire de fabrication des protéines, tandis que d'autres, à l'inverse, sont dissimulés et enfouis.

Une cellule n'est donc pas complètement libre. Son devenir obéit à des contraintes imposées par les gènes, l'expression de ces derniers dépendant de l'environnement. Pour rendre compte de ces conditions, le biologiste britannique Conrad Hal Waddington (1905-1975), à qui l'on attribue la paternité du terme *épigénétique* dans son sens moderne, a développé une métaphore, celle du paysage épigénétique. Elle apparaît dans plusieurs de ses publications, dont *The strategy of the genes*. P. Harrison, M. Towler et L. Li se sont approprié cette idée et ont conçu une vidéo pour l'illustrer *(voir la figure ci-contre)*.

Le paysage *(la nappe rayée en noir et blanc)* est un ensemble de « vallées » et de « montagnes » qui correspondent à l'expression ou non de certains gènes *(les picots noirs)*. L'influence des modifications



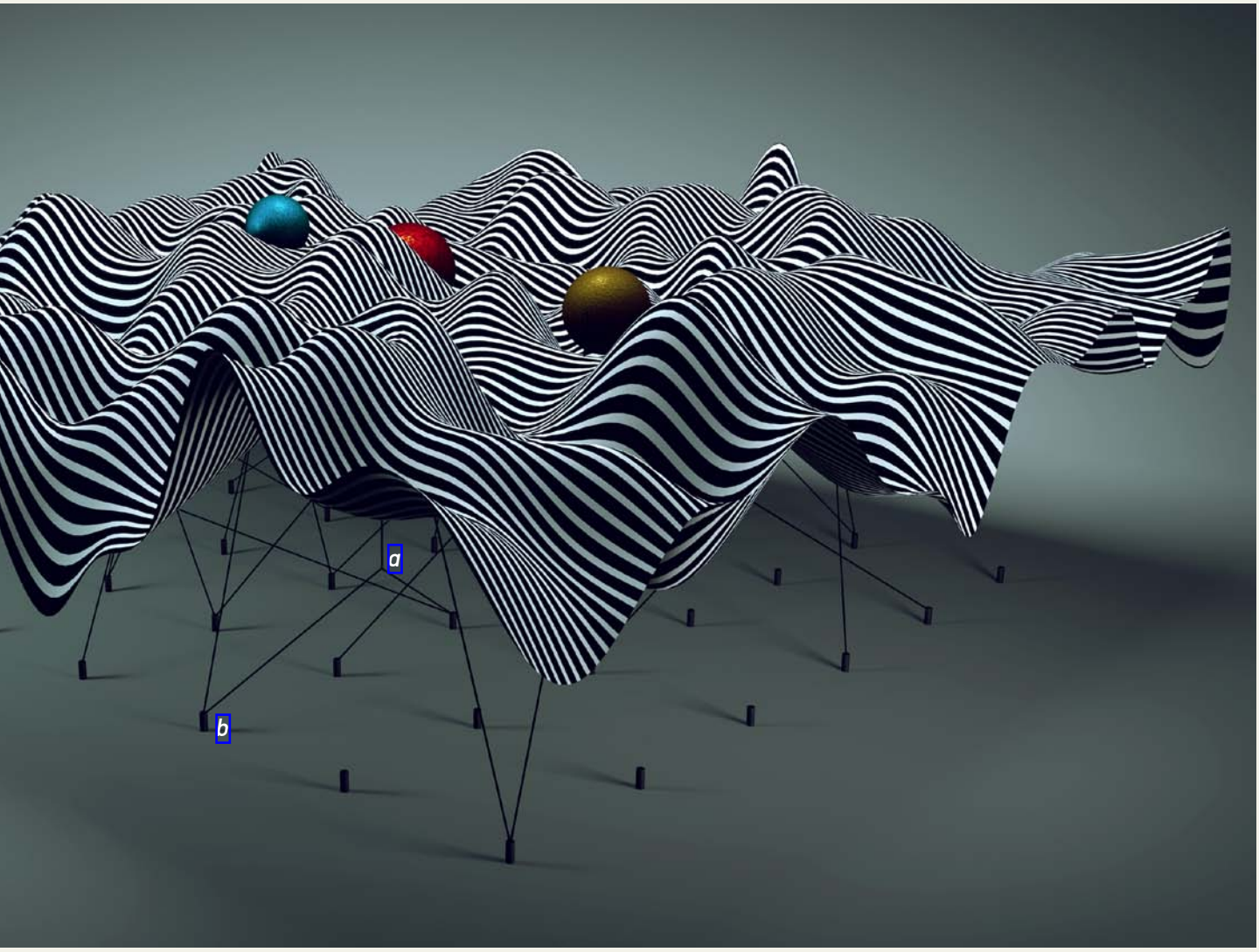
© Pour la Science - n° 433 - Novembre 2013

épigénétiques d'un gène donné est illustrée par un fil qui, selon qu'il est tendu ou non, se traduit par une dépression ou un pic. Ces modifications exercent en quelque sorte une « force chimique » qui déforme le paysage *(la nappe est élastique)* en raccourcissant ou non les fils. Certaines agissent de concert *(a, les fils sont réunis en un nœud)*, d'autres ont plusieurs effets *(b, plusieurs fils partent d'un même picot)*.

La différenciation des cellules d'un organisme est ici représentée par des billes *(en rouge, bleu et vert)* qui cheminent dans les vallées du paysage. Une bille ne peut que descendre vers son destin ! À chaque bifurcation, un choix est fait, dépendant des gènes et de l'environnement, engageant un peu plus irrémédiablement la cellule vers sa fonction finale. Waddington nommait *créodes* les chemins possibles, ce néologisme étant issu de deux termes grecs qui signifient « chemin nécessaire ».

Pour Waddington, le paysage épigénétique n'était qu'une image approximative que l'on ne pouvait interpréter de façon rigoureuse. Ce n'est plus le cas. Plusieurs équipes ont su déterminer expérimentalement le paysage épigénétique d'une cellule (adipocyte, cellule musculaire…) et son évolution. La dynamique révélée correspond bien aux animations de P. Harrison. La métaphore n'en est plus tout à fait une. ■

La vidéo **The Epigenetic Landscape** est visible sur : www. vivomotion.co.uk/folio.html
Le site de *EpiGeneSys* : www.epigenesys.eu



© Pour la Science - n° 433 - Novembre 2013

IDÉES DE PHYSIQUE

Des gouttes bien calibrées

Les gouttes se détachant d'un compte-gouttes sont de taille uniforme, tout comme celles formées par un filet d'eau qui se fragmente. Pourquoi ?

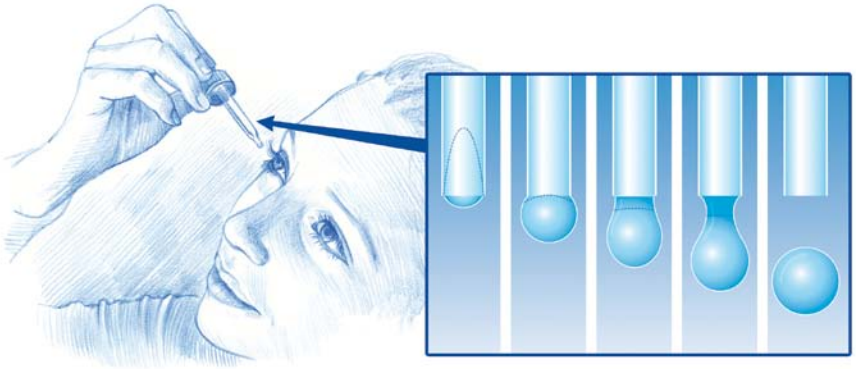
Jean-Michel COURTY et Édouard KIERLIK

Quand un robinet est mal fermé, les gouttes qui tombent une à une sont toutes identiques. Il en est de même pour celles qui se créent quand un filet d'eau se brise. Cette uniformité de taille est telle qu'elle permet un dosage précis de certains médicaments, administrés à l'aide d'un compte-gouttes. Elle ne repose pas sur le remplissage d'un récipient calibré, mais sur des phénomènes liés à la tension superficielle des liquides, c'est-à-dire à leur propension à prendre des formes dont la superficie est minimale.

Pour créer des gouttes de taille déterminée, il existe un outil simple et bien commode : le compte-gouttes. Lorsqu'on appuie doucement sur la poire qui surmonte le tube en verre, du liquide émerge et prend la forme d'une ampoule. Celle-ci s'allonge et se déforme à mesure qu'elle grossit, jusqu'à se détacher en une goutte à peu près sphérique *(voir la figure 1)*.

Que se passe-t-il ? Entre les molécules du liquide agissent des forces de cohésion. Il faut alors fournir de l'énergie pour amener des molécules du cœur du liquide à la surface, là où elles ont beaucoup moins de voisines, donc beaucoup moins de liens de cohésion. Cette propriété est quantifiée par la « tension superficielle », souvent notée γ (gamma), qui est l'énergie à fournir au fluide pour créer une unité de surface supplémentaire. Elle signifie aussi que l'on peut assimiler la surface du liquide à une mince pellicule élastique, qui doit être la moins étendue possible.

Lorsqu'on appuie sur la poire d'un compte-gouttes, on pousse du liquide hors du tube. Pourquoi le liquide ne tombe-t-il pas tout de suite ? Parce qu'il est retenu par la force capillaire exercée au niveau de la circonférence du tube. Cette force résulte de la tension superficielle (qui est de même dimension qu'une force par unité de longueur) et est égale à $2\pi r\gamma$, où r est le rayon du tube.



1. UNE GOUTTE SE DÉTACHE d'un compte-gouttes quand le poids du liquide émergeant du tube devient supérieur à la force de capillarité qui la retient. Cette force de capillarité est égale au produit de la tension superficielle du liquide par la circonférence de l'orifice du tube.

Force capillaire sur goutte pendante

À mesure que la goutte grossit, le volume de liquide augmente, mais pas la force capillaire. Lorsque le poids de liquide dépasse cette dernière, une goutte se détache et tombe – et son volume est déterminé, comme il se doit pour un compte-gouttes. Et si la goutte est assez petite, son poids ne suffit pas à la déformer : elle reste sphérique, forme qui présente une surface minimale pour un volume donné.

Avec un tube de un millimètre de rayon, on calcule, connaissant la valeur de la tension superficielle, que le volume d'une goutte d'eau qui tombe devrait être d'environ 50 millimètres cubes. Cependant, les considérations précédentes nécessitent des corrections. On constate en effet que la goutte pendante s'étire peu à peu et présente un col avec un rétrécissement important. Au moment de son détachement, une partie du liquide sorti hors du tube reste alors accrochée.

À toutes les étapes de sa croissance, la goutte a une forme parfaitement déterminée : celle qui réalise l'équilibre des pressions en tout point de sa surface. Quelles sont

Dessins : Bruno Vacaro

© Pour la Science - n° 433 - Novembre 2013

ces pressions ? Il y a d'abord la pression hydrostatique, due au poids du liquide et qui ne dépend que de sa hauteur. S'ajoute une différence de pression entre l'intérieur et l'extérieur de la goutte, due à la tension superficielle et qui apparaît lorsque la surface du liquide n'est pas plane.

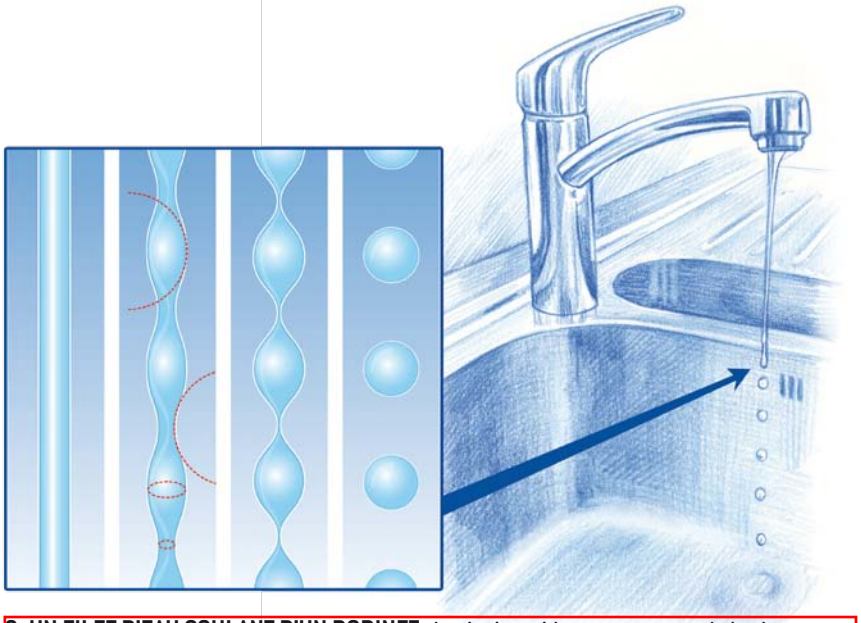
En effet, quand la surface d'un fluide est convexe, la pellicule élastique évoquée plus haut « presse » le fluide vers l'intérieur, comme le fait la paroi d'une bulle de savon ou d'un ballon de baudruche. Le fluide est donc en surpression par rapport à l'extérieur. Quand sa surface est concave, il est au contraire en dépression. Plus précisément, la différence de pression est proportionnelle à la tension superficielle et inversement proportionnelle au rayon de courbure moyen de la surface : c'est la loi dite de Laplace.

Avec ces ingrédients, il est possible de calculer la forme de la goutte pendante. Les prédictions sont d'une telle qualité qu'il est possible d'en déduire la valeur de la tension superficielle d'un liquide juste en examinant les photos d'une goutte pendante !

Des jets fragmentés par une instabilité

Intéressons-nous maintenant à l'évolution d'un jet d'eau régulier coulant d'un robinet. À mesure de sa chute, le jet s'amincit vite et fait apparaître des ondulations jusqu'à se fragmenter en petites gouttes de taille bien déterminée *(voir la figure 2)*. L'amincissement se comprend aisément : l'eau, en chute libre, gagne en vitesse ; le débit étant constant le long du jet, la section de ce dernier doit alors diminuer. Mais pourquoi le jet se fragmente-t-il ?

Les plus astucieux argueront que la forme approximativement cylindrique du jet n'est pas la forme optimale du liquide, à savoir la sphère, qui est de surface minimale pour un volume donné. Certes, mais selon quels mécanismes se développe l'instabilité qui transforme le cylindre en gouttes sphériques ? Une bonne idée est d'imaginer que le jet d'eau est un cylindre présentant déjà des ondulations périodiques *(voir le cartouche de la figure 2)*, et de se demander ce qui pourrait faire croître ou décroître ces déformations.



2. UN FILET D'EAU COULANT D'UN ROBINET s'amincit rapidement au cours de la chute et se fragmente en gouttes. Cette fragmentation résulte du développement d'une instabilité. Supposons qu'une perturbation quelconque déforme la surface du jet *(voir le cartouche)*. La déformation, considérée comme sinusoïdale pour faciliter les analyses, s'amplifie et finit par fragmenter le jet lorsque, dans les bourrelets, la pression du liquide est inférieure à celle du liquide des parties resserrées. Les différences de pression dépendant des courbures *(en rouge)* de la surface, les calculs montrent que l'instabilité se développe pour des longueurs d'onde supérieures à environ six fois le rayon du jet.

Les sections horizontales de ce cylindre ondulé sont des disques de rayons variables, plus grands au niveau des bourrelets du jet et plus petits au niveau de ses creux. D'après la loi de Laplace, la surpression est donc plus élevée dans les creux que dans les bourrelets. Par conséquent, le liquide va se déplacer des creux vers les bourrelets, de même qu'une petite bulle se vide de son gaz dans une bulle plus grande. Les creux vont disparaître, les bourrelets grossir et le jet se fragmenter. CQFD.

Ce raisonnement est hélas incomplet : il oublie que le cylindre ondulé est doté d'une autre courbure, celle que l'on visualise en faisant des coupes verticales. On s'aperçoit alors que l'on doit ajouter une surpression dans les bourrelets, tandis que les creux subissent une dépression en raison de la forme concave de leur surface. Les creux devraient donc aspirer le liquide et résorber les bourrelets !

Des deux phénomènes antagonistes, lequel l'emportera ? On peut déjà remarquer que pour des petites périodicités d'ondulation (à amplitude donnée), les

LES AUTEURS




Jean-Michel COURTY
et Édouard KIERLIK
sont professeurs de physique
à l'Université Pierre
et Marie Curie, à Paris.
Leur blog : http://blog.idphys.fr

BIBLIOGRAPHIE

T. C. Lipscombe et C. E. Mungan, **The physics of shot towers,** *The Physics Teacher*, vol. 50(4), pp. 218-220, 2012.

P.-G. de Gennes *et al.*, **Gouttes, bulles, perles et ondes,** Belin (2e édition), 2005.

Retrouvez les articles de J.-M. Courty et E. Kierlik sur www.pourlascience.fr



3. DES BILLES DE PLOMB sphériques et calibrées étaient fabriquées à l'aide d'une tour, comme à Couëron (Loire-Atlantique). Au sommet, on faisait fondre le plomb, que l'on versait dans une cuve percée de trous de taille précise. Le plomb s'écoulait en filets qui se brisaient en gouttes au cours de la chute. Les gouttes se refroidissaient en tombant et étaient freinées par un récipient rempli d'eau.

rayons de courbure dans le plan vertical sont plus petits, d'où des différences de pression plus élevées, ce qui favorise la résorption de l'ondulation. Il semble donc que seules les ondulations de grandes longueurs d'onde par rapport au rayon R du cylindre peuvent croître.

Le calcul complet est dû à Lord Rayleigh (1842-1919), qui identifia non seulement la valeur minimale de la longueur d'onde λ qui permet le développement de l'instabilité ($\lambda \approx 2\pi R$), mais aussi la valeur qui donne la croissance la plus rapide de l'ondulation ($\lambda \approx 9{,}02\,R$). Cette dernière est particulièrement intéressante, car elle fixe la longueur des segments de liquide qui vont se découper dans le jet et donc le volume des gouttes. En pratique, il faut ajouter l'effet de l'amincissement du jet, mais la conclusion est la même : la fragmentation conduit à des gouttes dont la taille ne dépend que du rayon initial du jet.

Il est remarquable que ce résultat ait été exploité bien avant son étude scientifique, pour fabriquer… des billes de plomb bien sphériques ! En faisant s'écouler du plomb fondu à travers une cuve percée de trous calibrés, on crée des jets liquides qui, par fragmentation, donnent des billes de la taille désirée. Si on le fait dans une tour assez haute (de l'ordre de 50 mètres), la chute permet de refroidir suffisamment la bille pour qu'il n'y ait ni choc thermique ni déformation lors de son arrivée et son freinage dans le réservoir d'eau, au bas de la tour *(voir la figure 3)*. Ce procédé a fait la richesse de son inventeur anglais, William Watts, au XVIII^{e} siècle. Il reste utilisé aujourd'hui pour la fabrication de billes d'urée, destinées surtout à l'industrie chimique. ■



© Pour la Science - n° 433 - Novembre 2013

ABONNEZ-VOUS À POUR LA SCIENCE

Offre liberté

19 € par trimestre

- ➜ *Pour la Science* (12 n^{os} par an)
- ➜ *Dossier Pour la Science* (4 n^{os} par an)





+ En cadeau le coffret de deux DVD *Énigmes des planètes*

Six documentaires produits par *National Geographic* pour explorer les splendeurs et phénomènes du Système solaire.

Coffret de deux DVD à paraître le 6 mai – valeur : 19,99€

© Vidéo 2013 Showshank Films. Tous droits réservés.

BULLETIN D'ABONNEMENT

POUR LA SCIENCE

À retourner accompagné de votre règlement dans une enveloppe non affranchie à : Service Abonnements Pour la Science • Libre réponse 90382 • 75281 Paris cedex 06

1 ❑ **Oui, je m'abonne à *Pour la Science* + *Dossier Pour la Science*** au prix de **19 €/trimestre** ou 76 € pour 1 an.
En cadeau, je reçois **le coffret de deux DVD *Énigmes des planètes*.**



608086

Votre cadeau de bienvenue !

2 **J'indique mes coordonnées :**

Nom : _____ Prénom : _____

Adresse : _____

Code postal : _____ Ville : _____ Pays : _____ Tél. : _____
Pour le suivi client (facultatif)

Je souhaite recevoir la newsletter *Pour la Science* à l'adresse e-mail suivante (à remplir en majuscule) :

_____@_____

3 **Je choisis mon mode de règlement :**

❑ Je règle en douceur **19 €/trimestre.** Je remplis la grille d'autorisation de prélèvement ci-dessous en joignant impérativement 1 RIB (mode de paiement valable en France métropolitaine uniquement).

Autorisation de prélèvement automatique (France métropolitaine uniquement). J'autorise l'établissement teneur de mon compte à prélever le montant des avis de prélèvements trimestriels présentés par *Pour la Science*. Je vous demande de faire apparaître mes prélèvements sur mes relevés de compte habituels. Je m'adresserai directement à *Pour la Science* pour tout ce qui concerne le fonctionnement de mon abonnement.

Titulaire du compte à débiter

Nom, Prénom : _____

N° : _____ Rue : _____

CP : _____ Ville : _____

Établissement teneur du compte à débiter

Établissement : _____

N° : _____ Rue : _____

CP : _____ Ville : _____

Compte à débiter

Établissement | Code guichet | N° de compte | Clé RIB

Date et signature (obligatoire)

Organisme créancier
Pour la Science
8 rue Férou – 75006 Paris
N° national d'émetteur : 426900

❑ Je préfère régler mon abonnement **en une seule fois 76 €** (+ frais de port en sus pour l'étranger : Europe 16 € - autres pays 35 €.)

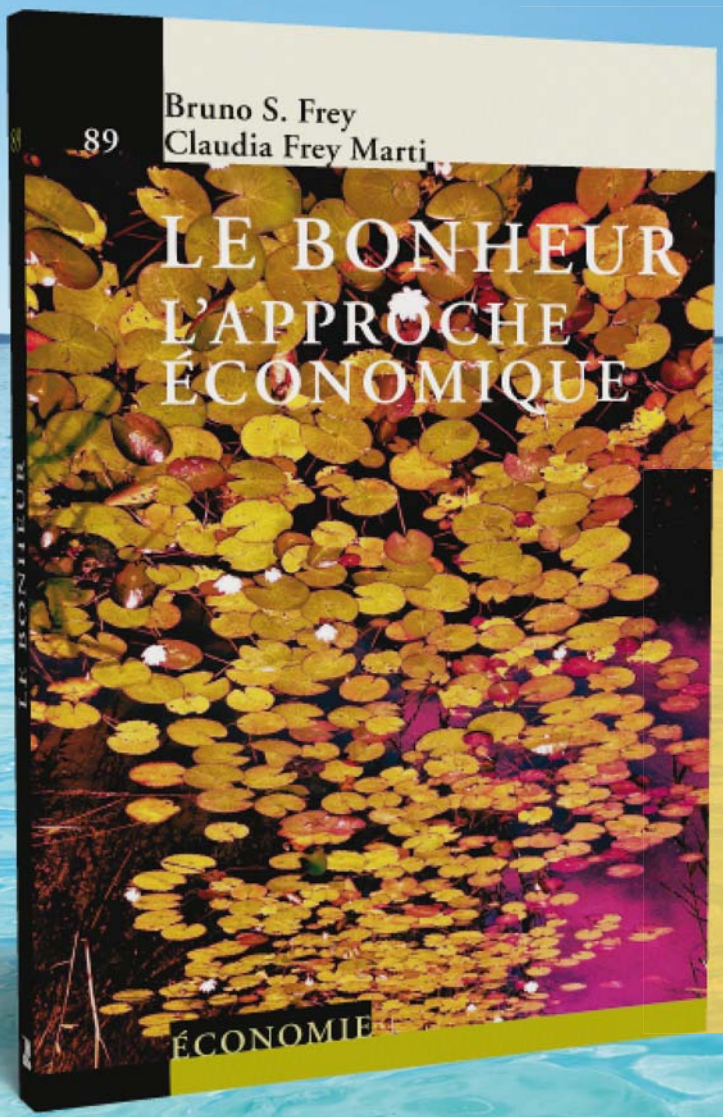
❑ par chèque à l'ordre de Pour la Science ❑ par carte bancaire N° : _____

Expire fin : _____ Clé (les 3 chiffres au dos de votre CB) : _____ Signature obligatoire :

PLS433 – Offre réservée aux nouveaux abonnés, valable jusqu'au 30.11.2013

En application de l'article 27 de la loi du 6 janvier 1978, les informations ci-contre sont indispensables au traitement de votre commande. Elles peuvent donner lieu à l'exercice du droit d'accès et de rectification auprès du groupe *Pour la Science*. Par notre intermédiaire, vous pouvez être amené à recevoir des propositions d'organismes partenaires. En cas de refus de votre part, merci de cocher la case ci-contre ❑

le bonheur, à quel prix ?



elle est la recette
Qu bonheur ?
du
pend-il de notre revenu, de notre
Dép ition sociale, de notre relation
pos loisirs, de notre statut civil ?
aux d'autres facteurs encore ?
Ou r la première fois, des économistes
Pou enchent sur ces questions, et
se pondent dans un petit ouvrage
y ré renant, clair et sans jargon.
sur

Bruno S. Frey, Claudia Frey Marti
Le bonheur, l'approche économique
13,50 €, 152 pages, 978-2-88915-010-6

PRESSES POLYTECHNIQUES ET UNIVERSITAIRES ROMANDES
Editeur scientifique de référence depuis 1980 www.ppur.com

SCIENCE & GASTRONOMIE

Il faut faire sauter le riz *après* l'avoir cuit à l'eau

Dans une cuisson à l'eau, le sort des amidons diffère selon leur nature. Pour la santé, il est bon d'en tenir compte.

Hervé THIS



De tes aliments, tu feras ta médecine, proclamait Hippocrate. Contrairement à Platon qui, dans le *Gorgias*, méprise la cuisine, Hippocrate, connaisseur de notre enveloppe corporelle, pensait que la transformation des aliments est essentielle pour déterminer les propriétés de ce que nous mangeons. Une étude récente de la cuisson du riz exemplifie cette intuition.

Le riz est souvent bouilli ou cuit à la vapeur ; parfois, dans le riz pilaf par exemple, la cuisson à l'eau est précédée par une friture plate (« sauté »). L'amidon du riz est alors plus facilement hydrolysé par les enzymes amylases présentes dans la bouche et dans le système digestif, ce qui provoque une forte augmentation postprandiale de la concentration en glucose dans le sang. Comment les personnes souffrant d'un diabète peuvent-elles l'éviter ? Des méthodes particulières de cuisson du riz seraient-elles préférables ?

L'hydrolyse enzymatique de l'amidon des riz cuits à la vapeur s'apparente à celle de l'amidon des pommes de terre bouillies. Or, pour les pommes de terre, la friture semble ralentir l'hydrolyse. Pourquoi ? Les amidons des divers féculents diffèrent. Ils sont constitués de deux sortes de polysaccharides (polymères dont les « maillons » élémentaires sont le glucose) : les amyloses, polymères linéaires de glucose, et les amylopectines, polymères ramifiés. L'hydrolyse enzymatique d'un amidon dépend de son contenu en amyloses, de la longueur des chaînes des amylopectines et des interactions de l'amidon avec les lipides, les protéines, les celluloses. Notamment, on nomme « amidon résistant » la fraction de l'amidon qui résiste à l'hydrolyse enzymatique dans l'intestin grêle.

Selon les riz, par exemple des variétés *japonica*, *indica* et *waxy*, les proportions d'amyloses, d'amylopectines et d'amidon résistant diffèrent. Le riz *japonica* a une température de gélatinisation plus faible que le riz *indica*, en raison d'une longueur de ramifications inférieure. Le riz *waxy* n'a quasiment pas d'amylose.

À l'Université de l'Iowa, Michael Reed et ses collègues ont examiné comment les méthodes de cuisson modifiaient ces riz. Ils ont d'abord mesuré que le contenu en amidon variait entre 80 % *(waxy)* et 84 % *(japonica)*. L'amidon d'*indica* contenait plus d'amylose (18 %) que celui du riz *japonica* (13,5 %) et, évidemment, que l'amidon de riz *waxy* (0,9 %). Les longueurs de ramifications ne sont pas fixes : les riz *japonica* et *waxy* ont une plus grande proportion de courtes ramifications (environ 40 %) que le riz *indica* (environ 20 %). En revanche, le riz *japonica* a davantage de longues ramifications que les riz *indica* et *waxy*.

Évidemment, ces riz gélatinisent diversement, ce qui leur donne des consistances différentes, à la cuisson dans l'eau ou dans la vapeur, mais ils sont surtout hydrolysés différemment. Les physico-chimistes de l'Iowa ont montré que, pour une cuisson dans l'eau, la proportion d'amidon hydrolysé était supérieure pour l'amidon *waxy*, qui n'a pas d'amyloses : la structure lâche des grains d'amidon permet un gonflement supérieur et une exposition accrue des amylopectines aux enzymes. Par ailleurs, alors que le riz *indica* a un contenu en amylose supérieur au riz *japonica*, le riz *indica* était plus hydrolysé. Cela s'explique sans doute par les particularités du gonflement, lequel est également déterminé par les ramifications des amylopectines.

Surtout, le riz sauté après cuisson dans l'eau était le moins hydrolysé : comme pour les pommes de terre frites, il semble que les triglycérides de l'huile de maïs forment des complexes avec l'amylose et que les complexes formés réduisent à la fois le gonflement de l'amidon et l'hydrolyse. Le degré d'hydrolyse était inférieur à celui mesuré lors d'une préparation de type riz pilaf, où le sautage précède la cuisson dans l'eau. Les différences résultent notamment de la formation de complexes amylose-lipide durant le stockage qui sépare l'ébullition et le sautage.

Ainsi, en cas de diabète comme avant un marathon (où une très lente hydrolyse de l'amidon favorise un apport régulier de glucose à l'organisme), il vaut mieux cuire le riz à l'eau, le laisser reposer, puis le faire sauter avant de le consommer. Et c'est ainsi qu'Hippocrate est grand ! ■



Hervé THIS est physico-chimiste dans le Groupe INRA de gastronomie moléculaire, professeur à AgroParisTech et directeur scientifique de la Fondation Science & Culture Alimentaire (Académie des sciences).

Retrouvez les articles de H. This sur www.pourlascience.fr

© Pour la Science - n° 433 - Novembre 2013

MÉDECINE

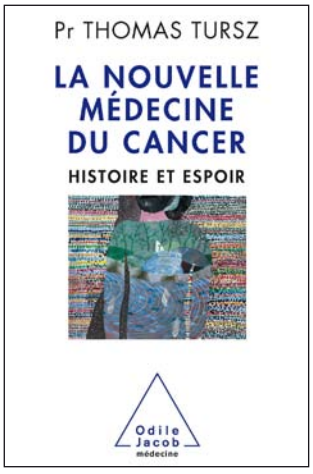
La nouvelle médecine du cancer

Thomas Tursz
Odile Jacob, 2013
(256 pages, 22,90 euros).

Oncologue renommé en Europe, bien connu en France pour ne pas manier la langue de bois, l'auteur pense que la cancérologie est aujourd'hui dans la situation des maladies infectieuses il y a un siècle. Pour l'ancien directeur (1994-2010) du premier centre de soins, de recherche et d'enseignement en cancérologie d'Europe – l'Institut Gustave-Roussy ou IGR, qui fait partie de l'Hôpital Paul Brousse, à Villejuif –, elle serait quelque part entre la microbiologie de Pasteur (la compréhension du problème) et les antibiotiques de Fleming (une solution universelle). Aujourd'hui, ses plus beaux succès, comme ceux enregistrés dans la lutte contre le cancer du sein, ne se mesurent plus seulement en termes de durée, mais aussi en terme de qualité de vie. Pour autant, et malgré d'énormes efforts thérapeutiques, peu de progrès ont été accomplis sur les cancers du poumon, du pancréas ou du cerveau. Et face à cela, les espoirs actuels, estime-t-il, sont à rechercher dans les perspectives qu'il nous brosse des développements récents de la génomique médicale.

Toutefois, l'intérêt du livre de ce professeur de cancérologie de l'Université Paris-Sud tient aussi aux deux chapitres qu'il consacre à l'histoire de sa spécialité médicale. Th. Tursz rend justice à son prédécesseur Gustave Roussy, immense médecin qui fut l'inventeur de la cancérologie moderne, c'est-à-dire d'une spécialité pluridisciplinaire, qui a supplanté la médecine organiciste. Sur le plan institutionnel, Th. Tursz nous explique qu'après la création de l'IGR en tant que fédérateur des Centres de lutte contre le cancer (CLCC) au lendemain de la guerre, la cancérologie s'est trouvée en porte à faux en 1961, quand l'immunologue et cancérologue Georges Mathé vint ajouter son propre Institut de cancérologie à l'Hôpital Paul Brousse et obtint son rattachement à l'Assistance publique - Hôpitaux de Paris. Le directeur de l'IGR d'alors, Paul Denoix, résolut ce problème en obtenant la construction du bâtiment de l'IGR, forme posée sur le plateau de Villejuif, familière aux automobilistes empruntant l'A1.

Th. Tursz ne nous en cache pas les inconvénients techniques. Il importerait, pense-t-il, d'en moderniser les infrastructures, afin de mieux mener la cancérologie dans les nouvelles voies de la médecine personnalisée. La recherche a montré en effet que chaque tumeur est aussi particulière que la personne qu'elle affecte, ce que le séquençage à haut débit du génome permet d'analyser finement. De là découlent les nouvelles voies thérapeutiques que veulent emprunter les cancérologues, ceux de l'IGR en particulier.



Jean-François Picard
HISTRECMED, CNRS, Villejuif

ASTRONAUTIQUE

Le rire d'Icare

Jacques Arnould
Cerf, 2013
(128 pages, 15 euros).

Janvier 1986. Un joint défectueux et la navette *Challenger* se désintègre en faisant sept morts, dont Christa Mc Auliffe, institutrice. Cet accident marque sans doute la fin du temps des pionniers astronautiques, pendant lequel presque tout était admis. Aujourd'hui, l'opinion publique ne veut plus que, pour faire quelques tours autour de la Terre ou aller sur la Lune, on risque trop facilement des vies. C'est pourquoi la réflexion que nous propose Jacques Arnould sur les risques spatiaux est dans l'air et dans l'azur…

Certes, on peut admettre que l'exploration fasse partie de la nature humaine, mais cet instinct ne constitue en aucun cas une justification sans retenue aux activités spatiales. Jusqu'à l'époque moderne, en Occident du moins, la Providence était chargée de prévoir, de prévenir et de pourvoir. Pour Copernic, la force de son système héliocentrique était de proposer l'ordre et l'harmonie. C'était rassurant, du moins plus qu'aujourd'hui, où lorsque notre espèce excède ses limites planétaires, elle se crée des risques. Un individu qui pénètre l'espace ne s'engage pas seul, mais il engage l'humanité entière. Les risques de contamination d'origine exoplanétaire sont, par exemple, à prendre au sérieux. Cela ne veut pas dire qu'il faut cesser les explorations spatiales, mais plutôt qu'il faut tout mettre en œuvre pour éviter cette contamination.

De la même manière, J. Arnould nous invite à réfléchir sur différentes notions qu'il est souhaitable d'associer à l'activité spatiale, telle



la prudence, le principe de précaution, la confiance, l'intime conviction, la justice, etc., et il discute de nombreux cas qu'il analyse souvent dans leurs contextes historiques. Un type de réflexion nécessaire dans de nombreuses activités humaines…

L'essai de J. Arnould ne sort pas du néant, puisque ce philosophe et théologien est chargé d'éthique au Centre national d'études spatiales (CNES), où il a développé aussi une réflexion en relation avec l'espace. Pour autant, son livre va bien au-delà du domaine spatial : tous ceux qui sont engagés dans une activité humaine potentiellement dangereuse pourront y trouver des éléments pertinents. Dans beaucoup d'entreprises en effet, le travail n'est pas que technique, et la notion de risque en fait partie, ne serait-ce que pour des raisons légales, sinon de responsabilité à l'égard des employés et des actionnaires. Une réflexion éthique s'impose donc, et le livre de J. Arnould la nourrit à partir du cas des activités spatiales, qui sont particulièrement dangereuses pour leurs acteurs… Voici donc une excellente occasion de s'arrêter pour réfléchir à ce que l'on veut, peut et doit faire.

Jean Cousteix
ONERA, Toulouse

© Pour la Science - n° 433 - Novembre 2013

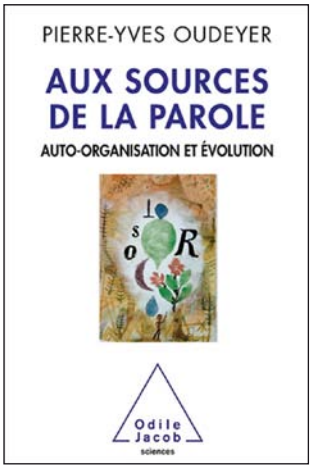
➜ LINGUISTIQUE

Aux sources de la parole

Pierre-Yves Oudeyer
Odile Jacob, 2013
(240 pages, 24,90 euros).

Qu'est-ce que le langage ? Si l'on convient que c'est la capacité d'exprimer sa pensée et de la communiquer de manière intelligible, comprendre sa forme verbale – la parole – est une problématique levier pour appréhender cet immense sujet. Ce que ce livre met à portée de toutes et de tous, c'est une approche encore inouïe : l'approche systémique. On y attaque le problème dans sa globalité, et on invite les sciences humaines, les sciences naturelles et les sciences formelles dans le même laboratoire d'idées.

Afin de rendre plus sensible cette démarche, évoquons le cas d'un autre objet d'étude : celui de l'évolution biologique. Nous savons qu'elle découle de la sélection naturelle, processus qui produit que les individus les mieux adaptés à leur environnement survivent mieux, et donc se reproduisent plus souvent, transmettant leurs gènes. Cette seule approche suffit-elle à tout expliquer ? Dans l'œuvre du pionnier de l'analyse mathématique des formes et des systèmes D'Arcy Thompson (1860-1948), on lit que les abeilles construisent des cellules qui prennent une forme hexagonale, ce qui est effectivement la forme optimale pour produire des parois avec le moins de cire possible. Le point clé est que dès que les insectes tentent de construire des cellules de même taille et de même forme, la structure hexagonale émerge, ce qui résulte avant tout d'une auto-organisation dictée par les lois physiques régissant la situation. Dès lors, considérer que la forme hexagonale sophistiquée des nids d'abeilles n'a émergé que des seuls hasards et de la nécessité de l'évolution n'est pas l'hypothèse la plus plausible. Comme nous le montre l'auteur, c'est bien plutôt l'interaction entre l'évolution et les phénomènes d'auto-organisation qui semble avoir fait émerger les niveaux d'organisation biologiques complexes que nous observons.



Abordant le langage sous ce point de vue, l'auteur attaque les questions suivantes : comment des systèmes de vocalisations (voyelles, consonnes, syllabes) peuvent-ils se former dans une population d'individus, puis en venir à être partagés ? L'hypothèse explorée dans ce livre est que ces systèmes seraient des macrostructures auto-organisées au cours d'interactions vocales pair à pair. L'auteur décrit de manière très didactique comment les expérimentations numériques et robotiques fournissent les méthodes scientifiques nécessaires pour véritablement comprendre les phénomènes les plus complexes du langage difficiles à décrire par des mots. Très clair, son texte est plaisant à lire. Un bel ouvrage de vulgarisation.

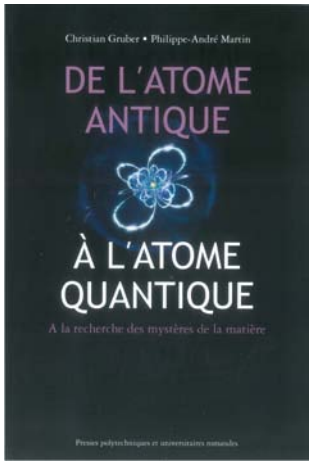
➜ Thierry Vieville
INRIA, Sophia-Antipolis

➜ PHYSIQUE

De l'atome antique à l'atome quantique

Christian Gruber et Philippe-André Martin
PPUR, 2013
(352 pages, 37,50 euros).

L'ambition est vaste et le pari réussi. Il y a sept ans, l'Université de Lausanne proposait aux étudiants en sciences humaines et sociales un enseignement facultatif de culture scientifique nommé Sciences au carré, (Sciences)^2, sous forme de modules. L'essentiel du module consacré à la physique de la matière et de la lumière est proposé dans cet ouvrage.



Sans transiger sur la rigueur scientifique, les auteurs soulignent les apports essentiels et les points d'achoppement, en ajoutant ponctuellement les outils mathématiques ou conceptuels nécessaires. Le propos déroule le temps, des premières idées sur la discontinuité de la matière chez les Anciens jusqu'au renouveau de cette hypothèse au début du XIX^e siècle. Un quart de l'ouvrage est consacré à l'histoire de la matière de Démocrite à Lavoisier, et se termine sur les premières grandes lois (gravitation, conservation de la quantité de mouvement et du moment cinétique, etc.). Tout au long du XIX^e siècle, de

Brèves

Océans
La grande alarme



Callum Roberts
Flammarion, 2013
(492 pages, 24 euros).

Ce livre traite des changements subis par les océans. Professeur de conservation marine à l'Université d'York, son auteur est l'un des créateurs du réseau mondial de sanctuaires marins. Dans cet essai vif et bien construit, il passe en revue les horreurs que l'humanité fait subir à la mer, avant de conclure, sombrement, par un chapitre intitulé « Se préparer au pire ». Un prophète de malheur, certes, mais qui rapporte des faits aussi indéniables que terrifiants, et qui, au terme de son épilogue, finit par lâcher quelques mots d'espoir.

L'être et l'écran



Stephane Vial
PUF, 2013
(324 pages, 19 euros).

Philosophe, l'auteur vise à un renouvellement conceptuel dans l'analyse de la technique numérique. Pour lui, les phénomènes numériques ne sont pas que virtuels (existant en puissance), mais réalisent toute une ontophanie numérique (une apparition de l'être sur l'écran), qui est à la fois noumène (au-delà de l'expérience sensible), programmable, interactive, simulatrice (virtuelle), instable, « autrui-phanique » (sociale), copiable, annulable, destructible, fluide, jouable (on peut jouer avec). De quoi comprendre ce qui, dans ce que l'on désigne par la « révolution numérique », est vraiment révolutionnaire ?

L'avion pour les trouillards



Brian Clegg
EDP Sciences, 2013
(214 pages, 16 euros).

Quelle idée heureuse et simple pour faire des sciences : partir d'un vécu partagé, s'en servir comme fil conducteur pour nouer moult billets vulgarisateurs sur les phénomènes expérimentés. Le vécu choisi ici est le voyage en avion, et l'auteur, un écrivain scientifique anglais, explore sans complications pas moins de 108 situations, phénomènes et objets techniques rencontrés à l'aéroport, à bord pendant l'envol, puis le vol, au-dessus des paysages ou des nuages, en cabine… Le tout bien tourné et éclectique. Idéal pour s'endormir serein lors du prochain voyage aérien.

© Pour la Science - n° 433 - Novembre 2013

Brèves

Le développement durable à découvert



A. Euzen, L. Eymard et F. Gaill (dir.)
CNRS éditions, 2013
(364 pages, 39 euros).

L'homme a un impact indéniable sur son environnement, mais comment s'assurer que cet impact ne soit pas nocif? On regroupe cette préoccupation sous le terme de développement durable, qui semble assez flou, mais dont les enjeux sont cruciaux. Pour y voir plus clair, plus de 190 chercheurs, de spécialités diverses, évoquent dans cet ouvrage les différentes facettes du développement durable: l'urbanisme, l'eau potable, le mix énergétique, l'agriculture, etc. Ils mettent en évidence le rôle essentiel de la science pour construire la société de demain.

Physique de la matière condensée



Michel Héritier
EDP Sciences, 2013
(614 pages, 70 euros).

Le sous-titre de ce manuel, « Des atomes froids aux supraconducteurs à haute température critique », reflète la modernité de son contenu, ainsi que... son niveau relativement élevé. Groupe de renormalisation, transitions de Kosterlitz-Thouless, fermions corrélés, ondes de spin, etc.: cet ouvrage traite de thèmes qui sont au cœur de la physique des solides d'aujourd'hui. Il nécessite de ses lecteurs une formation universitaire en physique correspondant à environ quatre ans d'études, et profitera aussi aux jeunes chercheurs.

Les femmes du laboratoire de Marie Curie



Natalie Pigeard-Micault
Glyphe, 2013
(300 pages, 18 euros).

Après la mort de Pierre Curie en 1906, Marie, à la tête de leur laboratoire, n'était pas seule. Outre de nombreux collaborateurs masculins, 45 femmes ont travaillé à ses côtés. Elles venaient en général de milieux aisés et ouverts à l'instruction féminine. En autant de notices, l'auteure, historienne des sciences, retrace les itinéraires de chacune à partir des archives Curie et de recherches généalogiques. À travers ces multiples parcours fort instructifs sur la condition féminine à l'époque, se compose par petites touches une description vivante du quotidien au laboratoire.

nouveaux champs d'investigation (électrolyse, chimie-physique…) et l'élaboration de nouvelles lois permettent à l'hypothèse atomique de se développer, de John Dalton à J. J. Thomson. Puis, avant d'aborder les grands changements du XX^{e} siècle, les auteurs consacrent un chapitre aux grands principes de conservation, aux symétries et invariances et un chapitre à la relation masse-énergie d'Einstein. Nous sommes à la moitié du livre.

Tout le reste porte sur l'histoire de l'atome, de J. J. Thomson à P. Higgs et son boson. La part belle est faite à la physique quantique, la radioactivité, la matière condensée, pour arriver au monde des particules élémentaires et au modèle standard. Mais les auteurs veulent aussi donner à rêver à leurs étudiants et présentent dans un chapitre final les dernières propositions des scientifiques sur les énigmes de l'Univers. Nous n'avons accès qu'à quatre pour cent de la matière de l'Univers, matière ordinaire soumise à la gravitation et susceptible d'absorber ou d'émettre de la lumière. Les théories physiques indiquent que l'Univers contient aussi 21 pour cent de matière noire et 75 pour cent d'énergie noire qui expliquerait l'expansion accélérée de l'Univers. Le propos est toujours clair et à la portée d'un bachelier scientifique. Avec cette impression d'avoir compris, on ferme le livre avec un sentiment de jubilation.

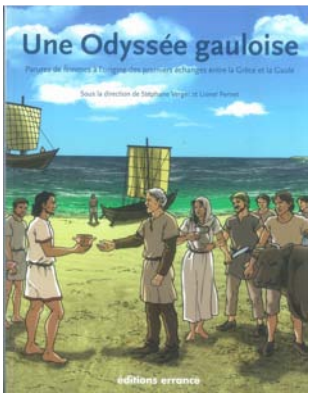
L'ouvrage propose de nombreux schémas, des encadrés, des annexes, une bibliographie et, à chaque fin de chapitre, un résumé des notions à retenir. Il peut se lire de façon continue, mais aussi dans le désordre, chaque chapitre se suffisant à lui-même. Une belle prouesse de didactique!

➜ Danielle Fauque
GHDSO, Université Paris-Sud

➜ ARCHÉOLOGIE

Une odyssée gauloise

S. Verger et L. Pernet (dir.)
Errance, 2013
(400 pages, 35 euros).



Dirigé par Stéphane Verger (Archéologie et philologie d'Orient et d'Occident, ENS) et Lionel Pernet (Lattara, Montpellier), ce bel ouvrage présente dans le détail une enquête archéologique menée à travers la Méditerranée par les meilleurs spécialistes du monde celtique.

Organisée en six chapitres et un épilogue, cet imposant volume présente la problématique, l'état de la recherche et les sites présentés dans le cadre de l'exposition qui se tient à Lattes jusqu'en janvier 2014, avec le détail des objets de chaque site. Les chapitres traitent avec minutie de questions complexes, bien éloignées des clichés trop souvent colportés à propos des Gaulois du Midi qui auraient été « civilisés » par les Grecs. Le lecteur apprendra l'état de la recherche sur les premiers contacts commerciaux avant la fondation de Marseille, l'économie du bronze, les femmes du premier âge du fer, et pourra suivre dans le détail l'hypothèse proposée de déplacements d'objets et de cultes féminins entre Languedoc et Sicile et leur interprétation comme étant les marqueurs de rites solaires féminins.

Abondamment illustré de planches de mobiliers (parfois un peu difficiles à reconnaître pour le novice), cet ouvrage est de nature à intéresser au plus haut point tous les amateurs d'archéologie. Son caractère exhaustif, sa précision et la grande qualité des interventions de spécialistes de l'Europe entière le destinent aussi tout naturellement aux professionnels, qu'ils soient spécialistes de métallurgie, de commerce, de Gaulois du Midi occidental, de colonisation grecque ou de religion… On y trouve en effet d'excellentes synthèses de sujets et de sites complexes, mais également beaucoup de données inédites ou difficiles d'accès, issues entre autres de fouilles préventives récentes, de dépôts anciennement connus et d'archéologie funéraire.

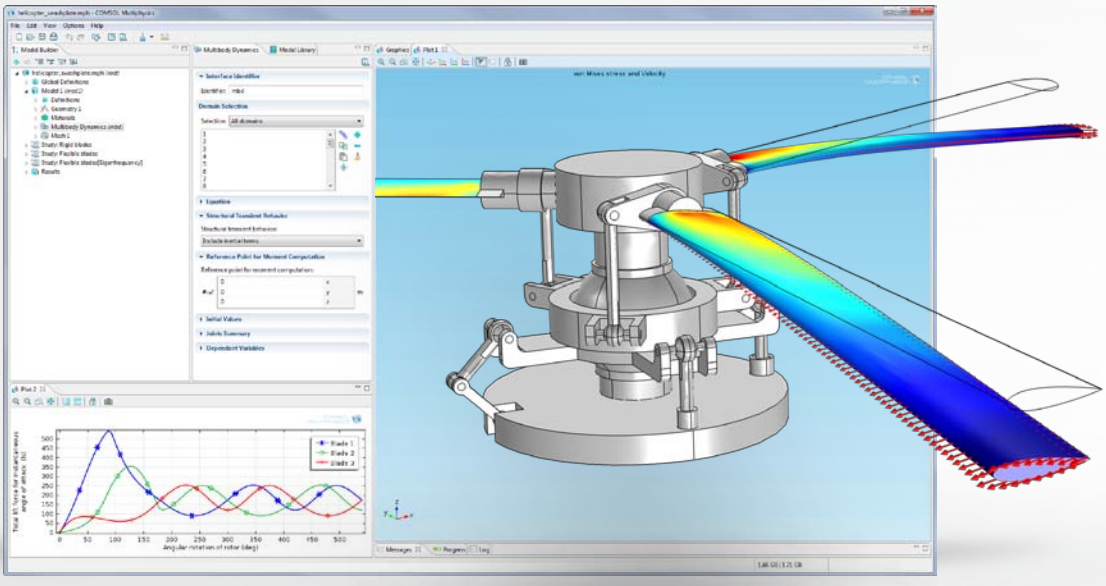
D'un point de vue plus général, cet ouvrage semble aussi intéressant pour le public éclairé, dans la mesure où il montre bien comment, pour des périodes pour lesquelles les sources écrites sont pour ainsi dire absentes, les archéologues parviennent aujourd'hui à reprendre des collections anciennes et à réécrire ainsi la fin de la préhistoire. Pas de doute: un ouvrage de valeur est né.

➜ Loup Bernard
Université de Strasbourg

Retrouvez l'intégralité de votre magazine et plus d'informations sur
www.pourlascience.fr

Imprimé en France – Maury Imprimeur S.A. Malesherbes – Dépôt légal 5636 – Novembre 2013 – N° d'édition 077433-01 – Commission paritaire n° 0917 K 82079 – Distribution : NMPP – ISSN 0 153-4092 – N° d'imprimeur 101/184 776 – Directrice de la publication et Gérante : Sylvie Marcé.

MULTIBODY DYNAMICS: L'orientation des pales du rotor d'un hélicoptère est controlée par le plateau cyclique.



Testez et optimisez vos designs avec COMSOL Multiphysics®

Pour reproduire avec précision le comportement de vos designs, nos simulations multiphysiques intègrent tous les effets physiques du monde réel. Pour plus d'information rendez-vous sur **www.comsol.fr/introvideo**

Product Suite

COMSOL Multiphysics

ELECTRICAL
- AC/DC Module
- RF Module
- Wave Optics Module
- MEMS Module
- Plasma Module
- Semiconductor Module

MECHANICAL
- Heat Transfer Module
- Structural Mechanics Module
- Nonlinear Structural Materials Module
- Geomechanics Module
- Fatigue Module
- Multibody Dynamics Module
- Acoustics Module

FLUID
- CFD Module
- Microfluidics Module
- Subsurface Flow Module
- Pipe Flow Module
- Molecular Flow Module

CHEMICAL
- Chemical Reaction Engineering Module
- Batteries & Fuel Cells Module
- Electrodeposition Module
- Corrosion Module
- Electrochemistry Module

MULTIPURPOSE
- Optimization Module
- Material Library
- Particle Tracing Module

INTERFACING
- LiveLink™ *for* MATLAB®
- LiveLink™ *for* Excel®
- CAD Import Module
- ECAD Import Module
- LiveLink™ *for* SolidWorks®
- LiveLink™ *for* SpaceClaim®
- LiveLink™ *for* Inventor®
- LiveLink™ *for* AutoCAD®
- LiveLink™ *for* Creo™ Parametric
- LiveLink™ *for* Pro/ENGINEER®
- LiveLink™ *for* Solid Edge®
- File Import *for* CATIA® V5

COMSOL

© Copyright 2013 COMSOL. COMSOL, COMSOL Multiphysics, Capture the Concept, COMSOL Desktop, and LiveLink are either registered trademarks or trademarks of COMSOL AB. All other trademarks are the property of their respective owners, and COMSOL AB and its subsidiaries and products are not affiliated with, endorsed by, sponsored by, or supported by those trademark owners. For a list of such trademark owners, see http://www.comsol.com/tm

10 DÉCEMBRE 2013

ORLÉANS
LES TURBULENCES - FRAC CENTRE

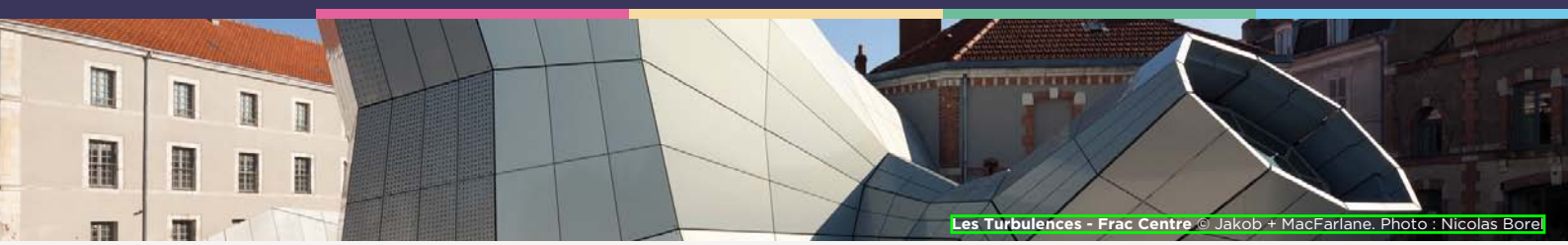
COSMETIC VALLEY
FRANCE

CONNEXIONS R&D

RENCONTREZ LES ACTEURS DE LA RECHERCHE EN COSMÉTIQUE

INITIEZ DES PROJETS DE RECHERCHE INNOVANTS

→ AU PROGRAMME DES CONNEXIONS R&D 2013



Les Turbulences - Frac Centre © Jakob + MacFarlane. Photo : Nicolas Borel

DEUX ATELIERS - CONFÉRENCES :
× Peptides et applications cosmétiques
× La cosmétopée Québécoise

DES RDV PERSONNALISÉS :
× Créez votre agenda en ligne
× Rencontrez des entreprises et des laboratoires publics

UNE SESSION EXPERTISE :
× Exposez vos expertises scientifiques et techniques

POUR S'INSCRIRE : WWW.CONNEXIONS-RD.COM

En partenariat avec

Société Française de COSMETOLOGIE

LES TURBULENCES FRAC CENTRE

QUÉBEC INTERNATIONAL Développement économique

Avec le soutien de

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Région Centre

Relations internationales, Francophonie et Commerce extérieur
MINISTÈRE DES AFFAIRES ÉTRANGÈRES
Québec

COOPÉRATION FRANCE-QUÉBEC