

<https://lejournal.cnrs.fr/articles/le-big-data-un-enjeu-economique-et-scientifique>

Le Big Data : un enjeu économique et scientifique

15.11.2012, par [Fabrice Demarthon](#) Mis à jour le 29.01.2014



Cette carte réalisée par Olivier de Beuchesne représente des collaborations scientifiques entre 2005 et 2009.

O.H. BEAUCHESNE/ SCIENCE- METRIX /DONNEES DE SCOPUS

Réseaux sociaux, clouds, appareils nomades, grands instruments scientifiques... Depuis quelques années, les nouvelles technologies produisent des masses de données numériques qu'il faut trier, analyser, stocker... L'homme est entré dans l'ère du Big Data.

De nos jours, l'humanité produit chaque année un volume d'informations numériques de l'ordre du zettaoctet. Soit presque autant d'octets qu'il existe d'étoiles dans l'Univers¹ ! En 2010 déjà, Eric Schmidt, le patron de Google, estimait que « *tous les deux jours, nous produisons autant d'informations que nous en avons générées depuis l'aube de la civilisation jusqu'en 2003* ». Des chiffres et des phrases chocs sur le sujet qui agite les mondes scientifique et économique aujourd'hui : les très grands volumes de données. « *Nous sommes confrontés à un déluge d'informations* », constate Christine Collet, chercheuse au Laboratoire d'informatique de Grenoble² et professeur à l'Institut polytechnique de Grenoble. Les chiffres parlent d'eux-mêmes. Chaque seconde, plus d'une heure de vidéo est mise en ligne sur YouTube, et plus de 1,5 million d'e-mails sont envoyés³.

¹ Source : IDC.

² Unité CNRS/UJF/Grenoble INP/UPMF/Univ. Stendhal-Grenoble-III/Inria.

³ Sources : YouTube, Instagram, The Radicati Group.

Un déluge d'informations à gérer

Les scientifiques ne sont pas en reste. En huit ans (2000-2008), le Sloan Digital Sky Survey, un grand programme d'observation astronomique, a enregistré 140 téraoctets d'images. Mais il ne faudra que cinq jours à son successeur, le LSST (Large Synoptic Survey Telescope), auquel participent des équipes de l'Institut national de physique nucléaire et de physique des particules du CNRS (IN2P3), pour acquérir ce volume ! Dans sa quête du boson de Higgs, le grand collisionneur de hadrons (LHC) a amassé de son côté, chaque année, près de 15 pétaoctets de données, l'équivalent de plus de 3 millions de DVD.



Le futur télescope LSST sera installé sur le Cerro Pachon, au Chili (vue d'artiste). Doté de 189 capteurs, il enregistrera des images du ciel d'une taille inégalée.

LSST CORPORATION

Avec le succès de l'économie numérique, la généralisation des appareils mobiles, le boom des réseaux sociaux, l'ouverture au public de certaines bases de données (l'Open Data) ou encore le développement de grands programmes scientifiques internationaux, le phénomène du Big Data va en s'amplifiant. « *On assiste actuellement à une véritable démocratisation des données, précise Christine Collet. Elles sont partout et sont élaborées, commercialisées et consommées comme n'importe quel produit manufacturé.* »

Ces grandes masses de données sont devenues un tel enjeu économique, industriel et scientifique que les gouvernements et les entreprises investissent massivement dans le domaine. Aux États-Unis, le plan Big Data alloue 200 millions de dollars à la recherche dans ce domaine (Big Data Research and Development Initiative). De son côté, l'Europe a inscrit la gestion des

contenus numériques dans ses priorités pour la fin du 7^e Programme-cadre de recherche et de développement technologique. En France, les Investissements d'avenir se sont aussi emparés du sujet. Un programme de 25 millions d'euros est consacré aux technologies d'exploitation des très grands volumes de données.

Un défi considérable pour les scientifiques

« *Le Big Data constitue un défi scientifique considérable qui nécessite des travaux aussi bien en ingénierie que dans les sciences fondamentales* », explique Mark Asch, chargé de mission pour les mathématiques et le calcul numérique à la Direction générale pour la recherche et l'innovation du ministère de l'Enseignement supérieur et de la Recherche. Face à ce constat, la Mission pour l'interdisciplinarité du CNRS a lancé en 2012 le défi Mastodons. L'idée est de soutenir des projets interdisciplinaires afin d'identifier où sont les verrous dans la gestion des grandes masses de données scientifiques.

« *Comment stocker les données et les pérenniser ? Comment les traiter, les analyser, les visualiser, leur donner du sens ? Comment les protéger, empêcher leur usage abusif et aussi les supprimer ? Autant de questions qui se posent aujourd'hui et dont nous sommes encore loin d'avoir les réponses*, détaille Mokrane Bouzeghoub, directeur adjoint scientifique de l'Institut des sciences de l'information et de leurs interactions du CNRS (INS2I), qui coordonne Mastodons. *Ces questions concernent autant les aspects algorithmiques que méthodologiques, sans oublier les infrastructures.* »

Une analyse que partage Farouk Toumani, chercheur au Laboratoire d'informatique, de modélisation et d'optimisation des systèmes ⁴et porteur du projet Petasky dans le cadre de Mastodons. « *Prenez le LSST, suggère-t-il. Ce télescope, qui doit être mis en service en 2020, sera capable d'enregistrer des images du ciel de 3 milliards de pixels toutes les 17 secondes. À la fin du programme, les astronomes disposeront ainsi d'une base de données de 140 pétaoctets, avec des centaines de caractéristiques pour chaque objet du ciel.* » Or, aujourd'hui, les algorithmes de fouille de données les plus efficaces mettraient des dizaines d'années pour explorer la base et répondre à certaines questions des chercheurs. Les scientifiques engagés dans le programme du LSST savent donc déjà que certaines de leurs interrogations les plus complexes resteront sans réponse.

Il faut inventer de nouvelles manières de manipuler les données

Cependant, une telle base de données constitue le champ d'application rêvé pour qui veut faire avancer la recherche fondamentale dans le domaine du Big Data : c'est précisément le cas des chercheurs du projet Petasky. « *Pour contourner les obstacles rencontrés dans la gestion des grandes masses de données, il faudra certes améliorer les technologies de stockage et de calcul, mais aussi inventer de nouvelles manières de manipuler les données* », annonce Farouk Toumani. C'est aussi l'avis de Sihem Amer-Yahia, spécialiste des réseaux sociaux au Laboratoire d'informatique de Grenoble. Elle confirme : « *La révolution des grandes masses de données, portée notamment par l'explosion du Web social où les citoyens produisent eux-*

⁴ Unité CNRS/UBP/UDA/ENSM.SE/Ifma.

mêmes l'information, bouleverse l'architecture classique des bases de données et de leur traitement. »

Tout le monde est concerné. Des biologistes aux astrophysiciens, de Facebook au fisc, plus personne n'échappe au phénomène du Big Data et aux problèmes qu'il engendre. Et l'urgence est là : désormais, le volume d'informations produites dans le monde numérique double tous les deux ans et le rythme s'accélère. « *Les données sont au cœur de l'économie numérique et de la société du savoir, conclut Christine Collet. Elles représentent une matière première à forte valeur ajoutée, sans laquelle rien ne pourra se passer. »*