

BIG DATA : FAIRE DU SENS À GRANDE ÉCHELLE

LE 13 JANVIER 2011 HENRI VERDIER

Après l'open-data la nouvelle tendance tout droit venue des US sera-t-elle le « big data » ?

D'un récent voyage dans la *Silicon Valley* (merci aux amis du *Orange Institute*), je rentre avec une conviction : **tout ce que nous connaissons du web va changer à nouveau avec le phénomène des big data**. Il pose à nouveau, sur des bases différentes, presque toutes les questions liées à la transformation numérique.

En 2008, l'humanité a déversé 480 milliards de Gigabytes sur Internet. En 2010, ce furent 800 milliards de Gygabytes, soit, comme l'a dit un jour Eric Schmidt, plus que la totalité de ce que l'humanité avait écrit, imprimé, gravé, filmé ou enregistré de sa naissance jusqu'en 2003.

Ces données ne sont pas toutes des œuvres. Outre les blogs, les textes, les vidéos (35 millions sont regardées sur Youtube chaque minute) ou le partage de musique, il y a désormais les microconversations, les applications géolocalisées, la production de données personnelles, la mise en ligne de données publiques, les interactions de l'Internet des objets...

Naviguer dans ce nouveau web demande une nouvelle science. C'est comme passer d'une navigation fluviale à une plongée en eaux profondes. *Mobilis in mobile*.

Hierarchy Of Visual Understanding? Just playing. Something in this?



David McCandless // v 0.1 // work in progress
InformationIsBeautiful.net

Qu'est-ce qui change avec les big data ?

L'actualité de la semaine nous a donné une petite illustration de ce qui se passe à grande échelle. Au fond, quelle est la différence entre *Wikileaks* et le *Canard Enchaîné* ? On voit bien qu'elle n'est pas seulement de volume. 250.000 documents d'un coup, ce n'est plus tout à fait les 30 rumeurs de la page 2 du *Canard*. Ça ne se traite pas pareil. Ça ne s'exploite pas pareil. Et visiblement, ça ne se combat ni ne se défend pas de la même manière. On sent bien que l'appareil juridique, l'appareil de communication et les stratégies de pouvoir adaptés au *Canard Enchaîné* ne passent pas l'échelle. Le web était globalement transactionnel. Des gens, des objets, des machines échangeaient. Des données, des conversations, des protocoles. De un à un, ou de beaucoup à beaucoup, on restait dans la transaction. Le web, aujourd'hui, produit aujourd'hui des masses de données, des masses de sens, qui échappent complètement aux principaux acteurs. Y compris à ceux qui déversent ces données. Y compris à ceux qui tentent de se les approprier. On sent bien que les questions aujourd'hui posées ne sont pas tout à fait à l'échelle. Tout savoir sur un individu ? Mieux cibler un marketing ? So what ? On est au bord de découvrir la thermodynamique et les gens continuent à regarder la trajectoire des molécules.

On sent bien que l'on est tout près d'un phénomène d'émergence. L'émergence est bien souvent une notion pseudo-mystique, mais c'est en fait une réalité. Lorsque des propriétés nouvelles apparaissent à un niveau d'organisation supérieur. Les propriétés du cerveau ne sont pas incluses dans le neurone. En tous cas, il est impossible de les anticiper simplement en disséquant un neurone. Ce sont des propriétés propres au système de neurones. L'émergence du *big data* sera de même nature. Il ne suffit pas de prolonger ce que nous faisons avec les données d'aujourd'hui pour anticiper ce que sera l'exploitation des données du futur.

Les outils sont-ils prêts ?

Ce qui est fascinant avec la *Silicon Valley*, c'est qu'une telle perspective y devient immédiatement un carburant pour la recherche et la création.

Le professeur **Michaël Franklin**, de Berkeley, a par exemple obtenu des financements importants de Google, Amazon, IBM, mais aussi SAP ou la **Darpa** (eh oui) pour créer son laboratoire **Algorithm, Machine, People**. Son programme de travail est simple : rien n'est prêt pour faire face à ce déluge de données. Les infrastructures profondes du réseau ne passeront pas l'échelle, la logique de bases de données n'est plus adaptée, les algorithmiques utilisées par l'informatique ne sont pas assez parallèles. Mais surtout, nous ne savons pas encore créer du sens à la bonne échelle. Donner cohérence, et valeur, aux informations, ou plutôt aux histoires dissimulées dans ces masses de données est encore une science balbutiante. Transmettre ce sens au plus large public est encore hors de portée.

Kul Wadhwa, le directeur de la **Wikimedia Foundation**, m'a raconté comment il réfléchit à la prochaine évolution de Wikipedia, d'un Wikipédia qui ne serait plus seulement une vaste encyclopédie, mais un immense système d'information, interfacé à toutes sortes d'API, utilisable par toutes sortes de services

Roger Magoulas, directeur de recherche chez O'Reilly pense tout simplement que les *big data* vont être le sujet le plus sexy des dix prochaines années. Et qu'il va nous falloir forger un langage commun entre les données, le quantitatif et le storytelling.

Ce nouveau rève Californien est holistique. Il organise la convergence de toutes sortes de travaux : datavisualisation, algorithmique parallèle, bases de données, Mapreduce, Machine Learning, sécurité et vie privée, datacenter, statistiques, analyse en langage naturel, sémantique. Il mobilise, et bouleverse, le journalisme, la communication, les microsimulations, les politiques publiques, l'urbanisme... Il est notamment stimulé par l'ambition d'un traitement en temps réel.

Il commence à avoir des résultats concrets. On ne parle plus tant serveurs et bases de données, ni même ERP, que *Mapreduce*, *NoSQL*, *GraphDatabase*, *Hadoop*. On voit naître un marché de ces bases de données d'un nouveau genre, encore immature mais dont les fleurons sont Neo4J, Allegrograph, HypergraphDB.

Big data, big business ?

Naturellement, une telle bascule induit également des transformations stratégiques majeures. Pour être honnête, on sent bien que le business n'est pas encore tout à fait au rendez-vous. Si *making sense at scale* est encore un problème ouvert, *making money at scale* est encore plus problématique. Sauf que, vue la vitesse d'évolution du web, on peut gager sans crainte qu'il ne faudra que quelques années avant que ces nouveaux business ne voient le jour. Il n'en demeure pas moins que la bataille industrielle a commencé. Et notamment la bataille pour la captation monopolistique de ces données. Ce n'est pas un

hasard si Tim Wu, professeur à Columbia, « gauchiste » au vu des Américains, vient de sortir un nouveau livre consacré à ces nouveaux monopoles. Nouveaux, car d'un genre nouveau. Google, Facebook, Twitter littéralement parlant, ne bloquent aucun marché. Ils sont, d'un certain point de vue, non rivaux du point de vue du marché des données personnelles. Ils ne monopolisent par ailleurs aucun marché, puisqu'ils créent les marchés au fur et à mesure de leur croissance. Et pourtant, ils sont visiblement engagés dans une lutte à mort pour la captation, le contrôle et l'exploitation de masses de données personnelles. Et comme le dit **Tim Wu**, ils créent de ce fait de nouvelles stratégies monopolistiques. Car posséder les données personnelles de 500 millions d'individus, ce n'est plus avoir un fichier illicite sur la clientèle de sa petite boutique. On est entré dans un ordre nouveau. On est dans les big data. Tim Wu propose d'imposer, mondialement, une séparation drastique entre les opérateurs d'infrastructures profondes, les fournisseurs d'accès Internet et les fournisseurs de contenus et services. C'est une idée qui mérite examen.

En même temps, je me demande si les nouveaux démiurges ne seront pas très vite dépassés par leurs créatures. Les données prolifèrent à une telle vitesse ; on apprend tellement vite à les interpréter. J'ai ainsi rencontré une start-up, **Sense Network**, fondée par des anciens de Columbia, qui réussit, avec la seule trace de déplacement des téléphones portables (d'obtention facile aux Etats-Unis), à prédire votre « lifestyle », à vous catégoriser et à prédire vos goûts, vos habitudes et même vos risques de santé (diabète, par exemple).

A quoi leur servira le monopole si l'on sait reconstituer le sens avec les moindres traces ?

—

Article initialement publié sur **le blog de Henri Verdier**

>> photos flickr CC **Tom Woodwar** ;

ÉTIENNE

le 13 janvier 2011 - 17:25 • SIGNALER UN ABUS - PERMALINK



*Je m'en veux de jouer au vieux c*n de service en bas d'un article intéressant, mais ça me démange beaucoup trop pour résister: cet anglicisme "faire sens" en titre, je trouve ça absolument HIDEUX.*

C'est vrai qu'il est difficile de ne pas se laisser déborder par l'anglais dans le domaine des technologies de l'information. C'est vrai qu'il est difficile de tout traduire en gardant un style fluide. Mais en français, on dit "avoir du sens". Alors, "Rester sensé à grande échelle" ?

VOUS AIMEZ



0

VOUS N'AIMEZ PAS



0

LUI RÉPONDRE

NPERPETUITE

le 13 janvier 2011 - 17:59 • SIGNALER UN ABUS - PERMALINK



Merci pour cet article ! Mais d'un point de vue purement stylistique, il faudrait vraiment veiller à faire des phrases courtes et pertinentes, histoire de ne pas perdre le lecteur et de mettre en perspective les "big data" :D

VOUS AIMEZ



0

VOUS N'AIMEZ PAS



0

LUI RÉPONDRE

NPERPETUITE

le 13 janvier 2011 - 18:01 • SIGNALER UN ABUS - PERMALINK



oops, sorry ! J'avais pas vu le "Article initialement publié sur le blog de Henri Verdier"

VOUS AIMEZ



0

VOUS N'AIMEZ PAS



0

LUI RÉPONDRE

STANISLAS JOURDAN

le 13 janvier 2011 - 18:01 • SIGNALER UN ABUS - PERMALINK



Bonjour Étienne,

Il faut parfois faire des choix...

Le titre en anglais avait l'inconvénient de rebuter beaucoup de lecteurs potentiels, tandis que la traduction française que vous suggérez ne me convient pas non plus...

Je suggère : "Générer du sens à grande échelle" ou encore "créer du sens" ???

VOUS AIMEZ



0

VOUS N'AIMEZ PAS



0

LUI RÉPONDRE

UPTOMONTREAL

le 13 janvier 2011 - 18:21 • SIGNALER UN ABUS - PERMALINK



Je connaissais open data mais big data c'est pertinent. Je m'intéresse beaucoup au crowd sourcing et à son potentiel de rupture. Mais quand on parle de crowd ou de data c'est avant tout des individus particuliers dont il faut parler comme le rappelai ce matin Jimmy Wales;)

VOUS AIMEZ



0

VOUS N'AIMEZ PAS



0

LUI RÉPONDRE

HENRIVERDIER

le 13 janvier 2011 - 18:35 • SIGNALER UN ABUS - PERMALINK



@Etienne. Vous avez raison. Mon titre initial était carrément en Anglais, encore pire ;)

@nperpetuite : je blogue, et j'écris comme je le sens :D

@Stanislas : "Donner du sens à grande échelle" ?

@uptomontreal : Oui, des individus, mais peut-être d'autres structures aussi. Une civilisation, par exemple, est un phénomène résultant des actions des individus, mais dont ils n'ont pas toujours conscience.

A tous : merci des commentaires et suggestions, c'est très stimulant.

VOUS AIMEZ



0

VOUS N'AIMEZ PAS



0

LUI RÉPONDRE

VISITORQ

le 14 janvier 2011 - 9:43 • SIGNALER UN ABUS - PERMALINK



Très intéressant et une véritable performance de ne pas citer Ilya Prigogine et ses travaux sur l'auto organisation et les propriétés émergentes!

VOUS AIMEZ



0

VOUS N'AIMEZ PAS



0

LUI RÉPONDRE

SHEILA BIANCHI

le 10 septembre 2011 - 17:38 • SIGNALER UN ABUS - PERMALINK



Could you publish the design of your blog page? I actually would truly like to have it

VOUS AIMEZ



0

VOUS N'AIMEZ PAS



0

LUI RÉPONDRE

4 pings

Numérique en France : pourquoi pas une Fondation ? » Article » OWNI, Digital Journalism le 11 février 2011 - 15:35

[...] Big Data : faire du sens à grande échelle [...]

Nous y étions : Conférence Les Echos "L'intimité client dans l'ère du social marketing" | Actualité | Locita le 6 juillet 2011 - 16:32

[...] une remise en cause de la Gestion de la Relation Client, et le parallèle avec l'enjeu des Big Data est tentant : "avec les réseaux sociaux, on récupère des milliers de conversations en [...]"

SAS wordpres le 16 septembre 2011 - 13:51

[...] « passer d'une navigation fluviale à une plongée en eaux profondes ». Lire la suite : <http://owni.fr/2011/01/13/big-data-faire-du-sens-a-grande-echelle/> Tags: Analyse de données, big [...]

Chat vidéo : comment les technologies impactent-elles le « business model » des entreprises ? | Buzz Modedemploi le 19 janvier 2012 - 7:28

[...] sociaux d'entreprise, au « social CRM » en passant par le « Big data » (en 2010, l'humanité aurait produit plus de données en un an que de sa [...])