

# DU CONTENU ROI AUX DONNÉES REINES

LE 21 JUILLET 2010 FRED CAVAZZA

**Les données sont le nouvel or numérique, aux enjeux considérables. Dans ce contexte, les moteurs de recherche, et en particulier Google, cherchent à se positionner sur le terrain de leur monétisation.**

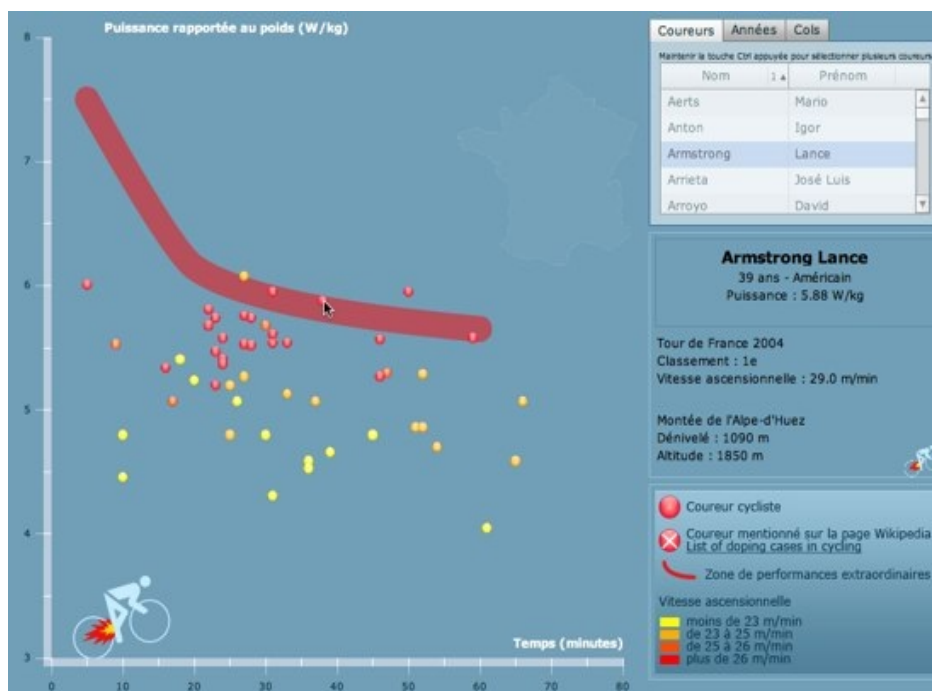
Souvenez-vous... il y a quelques années, le contenu était considéré comme la matière première du web : celui qui maîtrisait le contenu maîtrisait le web (les portails qui agrégeaient de très nombreuses sources de contenu concentraient également l'audience). Puis, il y a eu MySpace, les Skyblogs, Facebook, Twitter, Foursquare... et maintenant il paraît que c'est la communauté qui est reine. Certes, les plateformes sociales sont indéniablement en haut des tableaux d'audience, mais je reste convaincu que, sans contenus, une communauté n'est pas viable. Comprenez par là que ce sont les contenus qui alimentent les conversations et font tourner les communautés. De ce point de vue là, les plateformes sociales ne sont qu'un intermédiaire entre le contenu et les internautes. Un intermédiaire à valeur ajoutée, mais qui présente tout de même une certaine fragilité dans sa pérennisation (cf. **De la qualité des contenus sur Facebook**).

Sans rentrer dans la polémique, je pense ne pas me tromper en disant que **le contenu reste roi, la communauté se nourrit de ce contenu pour générer des interactions sociales** (mais là encore il y a des subtilités : **ne confondez plus communautaire et social**). La grande question que je me pose est la suivante : qu'est-ce qui alimente les rédacteurs de ce contenu ? C'est là où les données entrent en scène ; non pas les données que les rédacteurs possèdent déjà, mais plutôt les données disponibles publiquement que les internautes peuvent interroger et manipuler à loisir.

## Les données à la base du... journalisme de données

Nous parlons bien ici de données brutes en très grande quantité (des chiffres) qu'il serait trop coûteux de traiter. En les exposant publiquement, ce travail de compilation/trituration/interprétation est délégué à la communauté qui va ainsi pouvoir nourrir une réflexion ou appuyer des prises de position. Et à ce petit jeu, certains journalistes en ont fait leur spécialité, cela s'appelle du **journalisme de données** (*datajournalism* en anglais). L'idée est d'extraire des informations pertinentes de quantités importantes de données.

Pour vous aider à comprendre l'intérêt de cette pratique, amusez-vous à compter le nombre d'articles qui font référence à **Google Trends**, les statistiques de recherche sont les données sur lesquelles repose toute l'argumentation de ces articles. Autre illustration avec ce graphique très intéressant qui met en évidence les performances extraordinaires (=suspectes) des coureurs du Tour de France.



Ces données sont extraites du portail **ActuVisu** qui permet justement de manipuler des bases de données (cf. **Datajournalisme : du nouveau en France avec ActuVisu**). **Les données sont, dans ce cas de figure, la matière première d'une réflexion, ou plutôt d'une investigation.** Les possibilités sont nombreuses et la profession se met en marche pour développer de nouvelles compétences dans ce domaine. Pour mieux comprendre ce phénomène, je vous recommande les trois articles suivants : **pourquoi le data-journalisme, c'est l'avenir en marche**, **Quatre voies du datajournalism** et **Illusions et malentendus sur le journalisme de données.**

## Après les portails de contenus, les portails de données

L'exemple français d'ActuVisu illustre une tendance de fond initiée il y a cinq ans avec la fondation **Gapminder** qui fournit justement un accès à de très nombreuses données et statistiques (leur credo : *"Unveiling the beauty of statistics for a fact based world view"*).

Tout l'intérêt de ce portail est d'une part d'agrèger le plus grand nombre de données possible (de préférence en les rendant exploitables et compatibles) ainsi que de **fournir un outil simple pour manipuler et visualiser ces données.** Il existe d'autres initiatives comme **Many Eyes** d'IBM, **Socrata**, ou, plus modestement, **Worldmapper**. Notez que ces interfaces pour données sont une notion chère à Tim Bernes-Lee (cf. **ReadWriteWeb Interview with Tim Bernes-Lee, part 2 : search engines, user interfaces for data, Wolfram Alpha, and more...**), preuve que ce sujet est important.

Un créneau très porteur qui intéresse les moteurs de recherche de Google, qui a racheté en 2007 l'outil de visualisation qui propulse Gapminder et qui propose également **Google public data explorer** dans son labo. Ce rachat fait sens dans la mesure où Google est très certainement un des mieux placés pour collecter les données éparpillées aux quatre coins du web. Reste encore le problème des données non-publiques.

## Libération des données publiques avec Open Data

Les initiatives d'Open Data consistent à **libéraliser les données publiques pour apporter plus de transparence** (à l'image du portail anglais **WhereDoesMyMoneyGo?**) et pour nourrir des réflexions et projets sociétaux (lire à ce sujet **Open Data : des licences libres pour concilier innovation sociale et économique**). L'administration américaine a été la première à se lancer en ouvrant le portail **Data.gov**, suivie par d'autres pays comme l'Angleterre, l'Australie et la Nouvelle-Zélande (cf. **Quel modèle pour le data.gov français ?**).

The screenshot shows the homepage of data.gov.uk. At the top, there is a navigation bar with links for Home, Blog, Data, SPARQL, Apps, Ideas, Forum, Wiki, Resources, and About. Below this is a search bar. The main content area features a banner with the text "Unlocking innovation Working with UK Public Sector information and data" and a blue molecular structure graphic. To the right, there are buttons for "Subscribe by RSS", "Community Log in / Sign up", and "Local Data Panel". Below the banner is a "Latest datasets" section listing several datasets with their dates and descriptions. Further down is a "What we do" section with a paragraph about the site's purpose. Below that are two columns: "Search Data" with a search input field and "Browse for Data" with links for "Random dataset", "List all datasets", "By Public Body", and "Common tags". At the bottom, there is a "Most Recent Apps" section displaying three app cards: "Nottingham Information Prescriptions", "Jobseekers' Allowance claimant data", and "FTP Jobcentreplus feed". On the right side of the page, there are additional sections: "What is the Semantic Web?", "Digital Engagement Twitter stream" with a tweet from @DFID\_UK, and "Submit an app".

Il est important de comprendre que ces initiatives ne sont pas tant une manœuvre politique ou un outil de surveillance qu'un levier d'innovation pour **accélérer l'émergence de nouveaux modèles sociétaux ou de nouveaux projets relatifs à l'environnement, l'éducation, la santé...**

Pour le moment le chantier est toujours en cours en France mais des initiatives locales permettent déjà d'accéder à des poches de données : **État des lieux de l'OpenData en France**.

## Les données comme trésor de guerre des moteurs

Comme nous venons de le voir, les données sont donc une matière première particulièrement convoitée. À partir de ce constat, il n'est pas surprenant de voir que les grands moteurs de recherche s'intéressent de près à ces données et cherchent à les exploiter pour **apporter une couche d'intelligence aux résultats de recherche**. Illustration avec le tout nouveau **Bing Shopping** qui propose des pages de résultats structurées.

Web Images Videos Shopping News Maps More | MSN Hotmail Sign in

**bing**  
Shopping

Sort by: [best match](#) | [best user ratings](#) | [best expert ratings](#) | [price](#)

**BROWSE**  
 All Products  
 Cameras & Optics  
 Cameras  
 Digital Cameras

**BRAND**  
 Canon  
 Nikon  
 Olympus  
 Kodak  
 Sony  
 More >

**PRICE**  
 \$10 - \$25  
 \$25 - \$50  
 \$50 - \$75  
 \$75 - \$100  
 \$100 - \$200  
 above \$200

**TYPE**  
 Point and shoot  
 Prosumer


**RESOLUTION**  
 0.3 - 1.3 MP  
 2 - 3.3 MP  
 4 - 7.1 MP  
 7.2 - 10.19 MP  
 12 - 12.2 MP

**SCREEN SIZE**  
 0.9 - 1.1 in  
 1.3 - 1.75 in  
 1.8 - 2.4 in  
 2.48 - 3 in  
 3.5 - 3.7 in

**MEMORY TYPE**  
 Integrated  
 xD-Picture Card  
 Memory Stick Duo  
 xD-Picture Card Type H  
 Microdrive  
 More >


**VIEWFINDER TYPE**  
 Optical  
 None  
 Electronic  
 LCD  
 SLR

**SHOW ONLY**  
 Available products




**Canon PowerShot A550 - digital camera, 7.1MP, 4x Optical Zoom, 4x...**  
 Easy-to-use and easy-to-hold, the PowerShot A550 gives you 7.1 Megapixels, a 4x optical zoom lens, an ISO 800 feature for expanded low-light shooting and a DIGIC II Image... more...  
 ★★★★★ User reviews(252)  
 ★★★★★ Expert reviews(1)

**\$145** (1 store)  
 cashback · 3%  
[View price](#)




**Sony Cyber-shot DSC-T33 - digital camera**  
 You can be sure of a perfect picture with the Sony Cyber-shot DSC-T33 series. Though amazingly slim (about the size of a pack of cards) it's packed with features to make life... more...  
 ★★★★★ User reviews(59)

**\$179** and up (2 stores)  
 cashback · 3 - 8%  
[Compare prices](#)




**Panasonic Lumix DMC-FZ28K - digital camera**  
 The Lumix DMC-FZ28 digital camera boasts a premium 27mm wide-angle LEICA lens with an 18x optical zoom, ideal for tight indoor shots and long-distance action photos. The 10.1 mega... more...  
 ★★★★★ User reviews(6)

**\$600** (1 store)  
 cashback · 3%  
[View price](#)




**Canon PowerShot SD790 IS Digital ELPH - digital camera**  
 Chiseled edges with a subtle gleam give this PowerShot SD790 IS Digital ELPH distinctive sculptural appeal. Just as attractive are its high-end specifications, including 10... more...  
 ★★★★★ User reviews(3)

**\$301** (1 store)  
 cashback · 8%  
[View price](#)




**Nikon Coolpix S550 - digital camera, 10MP, 5x Optical Zoom, 4x Digital...**  
 Smart, and fast 10 megapixel camera with the smallest body in the world for its class. The 5x Zoom-NIKKOR lens brings distant subjects up close and Vibration Reduction means there... more...  
 ★★★★★ User reviews(45)

**\$78** and up (5 stores)  
 cashback · 2 - 12%  
[Compare prices](#)



**Nikon D300 Digital SLR Camera**  
 - 12.3 Megapixel DX-format CMOS sensor- The 3.0-inch super density 920,000-dot VGA color monitor- Continuous shooting from 6 fps up to 8 fps- Low-Noise ISO from 200-3200- Fast,... more...  
 ★★★★★ User reviews(252)  
 ★★★★★ Expert reviews(2)

**\$1,270** (1 store)  
 cashback · 3%  
[View price](#)



**Sony Cyber-shot DSC-H1 - digital camera**  
 A five megapixel, Super HAD CCD imager, 12X optical zoom lens and a large 2.5-inch LCD are just a few of the features on this model that are sure to please even the choicest... more...  
 ★★★★★ User reviews(252)  
 ★★★★★ Expert reviews(2)

**\$155** (1 store)  
 cashback · 3%  
[View price](#)

L'idée derrière tout ça est de proposer non pas un moteur de recherche, mais un outil d'aide à la décision (cf. **New version of Bing Shopping**). Et pour structurer des résultats, que faut-il ? Des données ! Autant Microsoft a opté pour des partenariats, autant Google est passé à la vitesse supérieure avec notamment l'acquisition d'ITA, un fournisseur de données touristiques spécialisé sur l'aérien qui va permettre à Google de faire de l'intégration verticale sur ce créneau : **With ITA purchase, Google now owns the skies.**

La vente de billets d'avion en ligne est un business très juteux, il est donc normal que Google casse sa tirelire pour blinder sa position. Il y a par contre des secteurs a priori moins rémunérateurs mais pour lesquels **un outil de consolidation/manipulation /visualisation des données offrirait une position dominante à son éditeur** : l'immobilier, l'emploi, les loisirs (**IMDB** est un bon exemple de données structurées à valeur ajoutée) ou encore le sport (citons l'exemple de **Footballistic**). Je vous recommande à ce sujet l'article de GigaOm qui détaille ces exemples : **Who will Google buy next for structured data ?**

L'idée ici est d'investir dans une base de donnée verticale et de monétiser son exploitation. Constituer une base de données de référence est un chantier titanesque, et seuls les acteurs avec les plus gros moyens peuvent y parvenir. Mais une fois le monopole établi, les possibilités sont nombreuses pour rentabiliser cet investissement. **Google Maps** est un autre exemple intéressant d'une gigantesque base de données (géographiques) dont nous avons maintenant beaucoup de mal à nous passer et dont le propriétaire a tout le temps pour trouver des solutions de monétisation viables.

Plus intéressant, un article de GigaOm nous révèle que ITA ne se restreint pas au secteur du tourisme aérien mais édite également **une solution de manipulation de données accessible sur Needlebase.com** : **Meet the web database company Google just bought.** Cette solution ne permet pas de manipuler des données publiques mais de groupes de données dont l'utilisateur a les droits. Toujours est-il que cette solution est à la fois puissante et intuitive, tout ce dont nous avons besoin pour faire du journalisme de données.

World Cup History

Index

- 2010 Finals
- 2010 Single Table
- 2010 Scorers
- 2010 Veterans
- 2010 Players Used
- 2010 Goals
- 2010 Ejections
- A History in Flags
- Cup Overview
- Cup Metrics
- Complete History
- About This Data

Match Stats -

- Country/Cup Grid
- Finals
- Highest Scoring Matches
- Most Cards
- Goals in the First 5 Minutes
- Goals in Injury Time
- Goals by Advantage
- 2-Goal Comebacks
- Result/Margin
- Result/Margin Totals
- Lead Changes
- Game Winning Goals
- Game Winning Own Goals

Country Stats -

- Country Table
- Finishing Places by Cup
- Points/Cup Grid
- Goal Diff/Cup Grid
- Repeat Selections
- Number of Goal Scorers
- All Scoring Leaders
- One-Timers
- Region Table
- Region Grid
- Europe vs South America
- Hosts Triumphant
- Hosts Beaten
- Beat by Eventual Winner
- Win Then Crash
- Goals to Reach the Final
- Defense Wins
- Championship
- Lost Then Won
- Bisulines
- Group Pairs
- Group Triplets

Country Table Country

group by

You can sort the World Cup country table a lot of different ways, and in most of them Brazil are #1. Most of these metrics are obvious, but W-P-D-L is a modified Win/Draw/Loss breakdown that calls out penalty-kick wins as the P. Points, just for fun, assigns 3 for regular wins, 2 for penalty-kick wins, 1 for draws, 0 for losses. PPG divides that point score by the number of games, x is total goals for and against. Flukiness is the difference between the country's best and second best PPGs per tournament (if they made it to at least 3).

Country	Wins	Trips	Rounds	Matches	W-P-D-L	Points	PPG	x	Flukiness
Brazil	5	19	58	97	67-2-12-16	217	2.237	210-88	0
Italy	4	17	47	80	44-1-17-18	151	1.888	126-74	0.4
Germany	3	17	63	99	60-4-15-20	203	2.051	206-117	0
Argentina	2	15	37	70	37-3-9-21	126	1.8	133-80	0.314
Uruguay	2	11	28	47	18-1-11-17	67	1.426	76-65	0.5
England (UK)	1	13	32	59	26-0-16-17	94	1.593	77-52	0.467
France	1	13	32	54	25-2-7-20	86	1.593	96-68	0.714
Spain	1	13	28	56	28-1-9-18	95	1.696	88-59	0.321
Sweden	0	11	25	46	16-1-12-17	62	1.348	74-69	0.31
Netherlands	0	9	25	43	22-0-9-12	75	1.744	71-44	0.286
Mexico	0	14	22	49	12-0-11-26	47	0.959	52-89	0.25
Yugoslavia	0	9	19	37	16-0-7-14	55	1.486	60-46	0
Belgium	0	11	19	36	10-1-8-17	40	1.111	46-63	0
Hungary	0	9	18	32	15-0-3-14	48	1.5	87-57	0.15
Czechoslovakia	0	8	17	30	11-0-5-14	38	1.267	44-45	0.45
Soviet Union	0	7	15	31	15-0-6-10	51	1.645	53-34	0.25
Austria	0	7	15	29	12-0-4-13	40	1.379	43-47	0.9
Poland	0	7	14	31	15-0-5-11	50	1.613	44-40	0.857
Switzerland	0	9	14	29	9-0-5-15	32	1.103	38-52	0.25
United States	0	9	14	29	7-0-5-17	26	0.897	32-56	0.6
Chile	0	8	13	29	9-0-6-14	33	1.138	34-45	0
South Korea	0	8	13	28	5-1-7-15	24	0.857	28-61	0.381
Paraguay	0	8	13	27	7-1-9-10	32	1.185	30-38	0.1
Portugal	0	5	13	23	12-1-2-8	40	1.739	39-22	0.5
Bulgaria	0	7	12	26	3-1-7-15	18	0.692	22-53	0.905
Romania	0	7	11	21	8-0-3-10	27	1.286	30-32	0.05
Scotland	0	8	8	23	4-0-7-12	19	0.826	25-41	0.333
Cameroon	0	6	8	20	4-0-7-9	19	0.95	17-34	0.467
Denmark	0	4	8	16	8-0-2-6	26	1.625	27-24	0.5
Croatia	0	3	7	13	6-0-2-5	20	1.538	15-11	1.143
Ireland	0	3	7	13	2-1-6-4	14	1.077	10-10	0.25
Peru	0	4	6	15	4-0-3-8	15	1	19-31	0.333
Japan	0	4	6	14	4-0-2-8	14	1	12-16	0.25

Tout récemment, Google a fait une acquisition qui va dans ce sens, en mettant la main sur **Metaweb**, une gigantesque base de donnée "ouverte" où sont répertoriés douze million d'entités sémantiques (visibles sur [Freebase.com](http://Freebase.com)) : **Google acquires 'open database' company Metaweb to enrich search results.**

## Vers des systèmes auto-alimentants

Voici donc la stratégie de Google : acheter des données avec l'idée de la monétiser une fois que le marché sera devenu dépendant de leur exploitation. Mais sommes-nous réellement dépendant des données ? Vous particulièrement, probablement pas, mais de nombreux aspects de votre quotidien reposent sur une exploitation fine de données. Nous pourrions même aller plus loin en disant que **l'exploitation des bonnes données pourrait améliorer votre quotidien** (cf. **Nos vies gérées par les données**) ou la

productivité d'une entreprise.

Les objets de notre quotidien pourraient ainsi capter un grand nombre de données vous concernant et fournir ainsi des statistiques très précieuses sur votre mode de vie et la façon d'optimiser votre alimentation, vos trajets, votre budget, votre suivi médical... Imaginez alors l'intérêt d'un coach qui serait à même d'interpréter ces données et de vous offrir de précieux conseils pour améliorer votre quotidien. **Ces conseils, et les données qui en sont à l'origine deviendraient rapidement une drogue pour des hommes et des femmes soucieux de leur bien-être : The upcoming Internet pandemic : data addiction.**

Reste encore à régler le problème de la collecte : seule une minuscule minorité des habitants de cette planète serait d'accord pour s'équiper des outils de mesure de son quotidien (sommeil, alimentation, exercices physiques, trajets, dépenses...). Une minorité de geeks, sauf si un acteur industriel avec de gros moyens décide de fournir gratuitement les outils de mesure et de collecte en faisant un pari sur l'avenir (et sur la monétisation de ces données). Et cet industriel avide de données, encore une fois c'est Google avec son projet de compteur intelligent **PowerMeter**.



Et même si Google ne peut pas remplacer tous les compteurs électriques des pays occidentaux, il peut fournir la plateforme pour consolider les données et les re-publier : **Google releases API for energy tool PowerMeter**. La promesse de Google est simple : vous aider à mieux comprendre vos habitudes de consommation pour optimiser vos dépenses... Tout en revendant les statistiques aux industriels pour qu'ils puissent développer des appareils ménagers plus en phase avec le mode de vie de leurs clients.

Loin de moi l'idée de jouer les paranoïaques et de dénoncer ces pratiques, car si tout le monde y trouve son intérêt il n'y a pas de raison de s'en priver. Il n'empêche que si je fais la somme de tout ce que Google peut potentiellement savoir sur moi, ça commence à faire beaucoup :

- Mes contacts avec Gmail ou Android (carnet d'adresse + historique des appels) ;
- Mon profil (âge, parcours...) avec Google Me ;
- Mes achats avec Checkout ;
- Mes centres d'intérêt avec l'historique de mes recherches ;
- Mes déplacements avec Latitude ;
- Mes loisirs (les programmes TV que je regarde) avec Google TV ;
- Mes lieux de vacances avec Picasa...

Et ce n'est qu'un début, car avec la sémantisation progressive du web, le moteur d'indexation pourra consolider toujours plus de données sur les internautes, mobinautes et même tvnauts. Les données seront donc la matière première à une nouvelle génération d'outils, services et prestations en rapport avec l'amélioration du quotidien de chacun. Des données qui seront l'objet d'une bataille acharnée pour en contrôler la possession, la collecte ou l'exploitation.

J'anticipe donc **un web, dominé par les contenus et données, où Google jouera un rôle prépondérant**. Facebook ou Twitter peuvent-ils prétendre à un rôle important dans ce tableau ? J'en doute car il faut des moyens considérables et surtout des appuis industriels et politiques, tout ce qui leur fait défaut actuellement. Longue vie au couple royal !

Billet initialement publié chez **Fred Cavazza** ; image Loguy /-)

Consultez nos articles **sur les données** et le **datajournalisme**, dont nos productions en la matière, et l'**opendata**

### DESIRADE

le 21 juillet 2010 - 14:25 &bullet; SIGNALER UN ABUS - PERMALINK



*Les données sont au contenu ce que les abeilles sont au miel (en suçant mon pouce)*

VOUS AIMEZ



0

VOUS N'AIMEZ PAS



0

LUI RÉPONDRE

### 1 ping

Les tweets qui mentionnent Du contenu roi aux données reines » Article » OWNI, Digital Journalism -- Topsy.com le 21 juillet 2010 - 10:36

*[...] Ce billet était mentionné sur Twitter par Elodie Moreels, Henri-Paul Roy, Michel Guillou, Matthieu Catillon, Kantken et des autres. Kantken a dit: RT @nbenyounes: RT @owni: [#owni] Du contenu roi aux données reines <http://goo.gl/fb/jBGZZ> [...]*