L'Open Data, enjeux et outils

Retour au sommaire de la lettre

Domaine :	**Recherche**	Référencement
Niveau :	**Pour tous**	Avancé

Il existe un nombre considérable de bases de données d'informations publiques, appelées "datasets", disponibles sur le Web, regroupées sous le terme générique d'Open Data. Plateformes proposées au niveau d'un pays, d'une région, d'une ville ou d'un organisme, il est tout à fait possible de les utiliser pour vos propres applications. Encore faut-il les détecter et les utiliser à bon escient. Voici quelques conseils pour y arriver...

Les enjeux de l'Open Data

Le mouvement de l'Open Data, ou "données ouvertes", a pour objectif de permettre à tous l'accès et l'usage des données publiques et, plus spécifiquement, des données brutes et structurées (séries permettant le travail statistique, description, coordonnées, horaires, budgets, données en temps réel émises par des capteurs,...).

Les caractéristiques des données ouvertes sont les suivantes :
- Collectées ou produites par les organismes publics ;
- Non-nominatives ;
- Ne relevant pas de la vie privée ;
- Ne relevant pas du domaine de la sécurité (différent pour les pays anglo-saxons).

Le secteur public génère en effet une grande variété de données qui ont vocation à être réutilisées, dans un cadre commercial ou non, afin de produire de nouveaux services ou, plus simplement, de permettre aux citoyens de s'informer et de prendre des décisions. Si par exemple une personne dispose, *via* une seule et unique carte géographique interactive, de données concernant les moyens de transports vers un lieu et les prix des loyers par secteur, elle sera plus à même de choisir son futur lieu d'habitation (voir le site anglais http://where-can-i-live.com/). On retrouve par cet exemple les préoccupations du chercheur en économie Ian Ayres, qui prévoit l'avènement de ceux qu'il nomme les *Super Crunchers*, des individus qui maîtriseront l'analyse statistique des données impactant le monde réel (bourse, météo, paris sportifs,...) et sauront en tirer partie, tant dans leur vie privée que dans leur vie professionnelle.

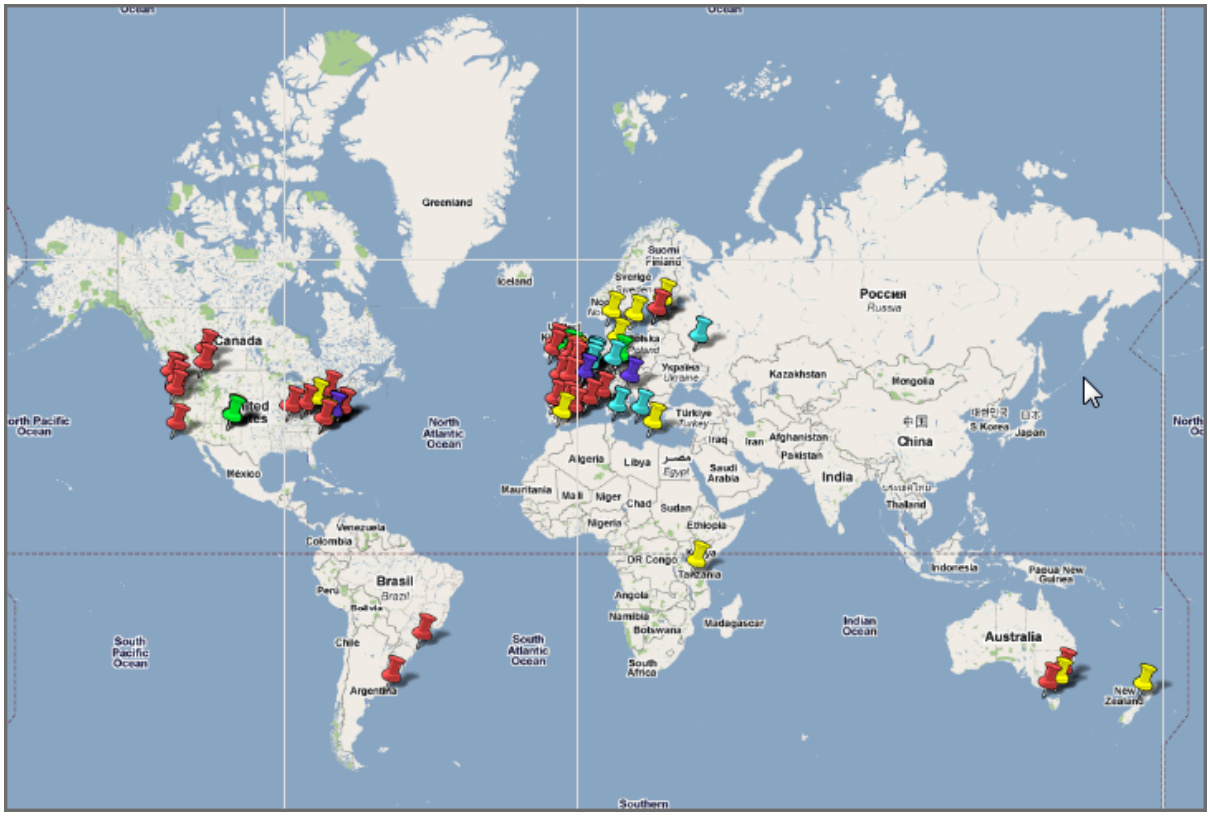
La Fondation Internet Nouvelle Génération (FING) résume la problématique de l'Open Data à six enjeux majeurs :
- Nouveaux services ;
- Développement économique ;
- Culture, création ;
- Prospective, aménagement ;
- Débat public, démocratie ;
- Connaissances et analyses.

Nous y ajoutons l'enjeu transversal de transparence, qui est un challenge nécessaire pour les démocraties actuelles.

En France, les projets viennent pour l'instant essentiellement des villes. La ville de Paris met par exemple à disposition des jeux de données sur le site http://opendata.paris.fr, idem pour Montpellier (http://opendata.montpelliernumerique.fr/). Bordeaux, Nantes, Marseille ont pour leur part annoncé des projets qui aboutiront entre la fin de l'année 2011 et 2013.

L'Etat français n'est pas en reste puisque le projet data.gouv.fr devrait voir le jour en décembre prochain (voir le blog du groupe de travail : http://blog.etalab.gouv.fr/).

Techniquement les sets de données mis à disposition adoptent généralement des standards (XML, RDF, JSON) qui facilitent leur interopérabilité, leur mise à disposition dans des API et la possibilité de les interroger.



Carte mondiale des initiatives Open Data. http://bit.ly/a9D05o

Comment trouver des jeux de données

Les initiatives de mise à disposition de jeux de données structurées sont nombreuses dans le monde. Voici une typologie des services qui vous permettront d'y accéder :

Les plateformes nationales de données publiques

Il s'agit des bases de données en ligne proposées directement par les états, à l'instar de ce que sera data.gouv.fr. Exemples :
http://www.data.gov/ : portail du gouvernement américain ;
http://www.data.gov.uk/ : portail du gouvernement anglais ;
http://www.data.gov.au/: portail : portail du gouvernement australien ;
http://www.data.gc.ca/ : portail du gouvernement canadien.

Les plateformes régionales de données publiques

On y trouvera des données mises à disposition par les régions, départements, états (US),...
Exemples :
http://opendata.euskadi.net/ : portail du pays basque espagnol ;
http://data.ok.gov/ : portail de l'état de l'Oklahoma.

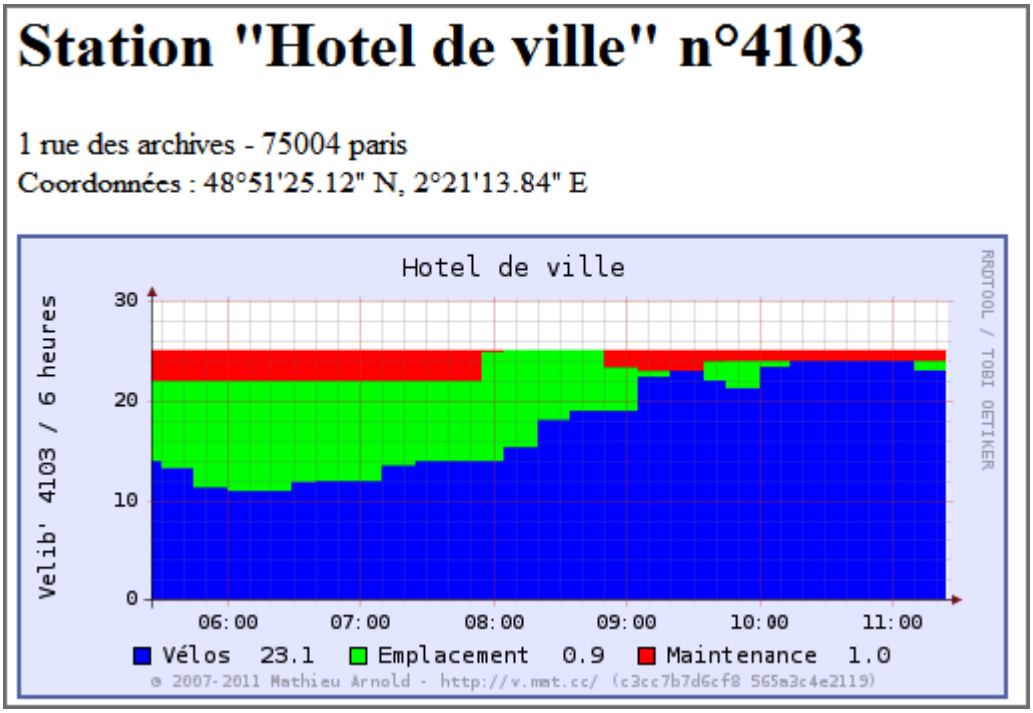
En France, le département de Saône et Loire devrait être le premier à se doter d'un portail de données ouvertes selon les vœux formulés par son président du conseil général, Arnaud Montebourg.

Les plateformes locales de données publiques

Données mises proposées par les municipalités.
Exemples :
http://opendata.paris.fr/ : portail de la ville de Paris ;
http://opendata.montpelliernumerique.fr/ : portail de Montpellier ;
http://www.data.rennes-metropole.fr/ : portail de Rennes.

Les plateformes d'initiative citoyenne

Des citoyens créent également des plateformes pour lancer des services répondant à des besoins qui pourraient/devraient être adressés par des collectivités locales ou des services publics.
Exemples :
http://www.montrealouvert.net/ : initiative ayant pour but de promouvoir l'accès ouvert aux données civiques de la région de Montréal ;
http://www.nosdonnees.fr/ : plateforme de mise à disposition de jeux de données d'organismes publics français ;
http://www.nosdeputes.fr/ : observatoire citoyen de l'activité parlementaire ;
http://v.mat.cc/ : statistiques d'utilisation des Velib par station.



Les annuaires de jeux de données (datasets)

Il existe également de nombreux annuaires répertoriant les fichiers Open Data disponibles de par le monde :

Data Publica (http://www.data-publica.com/) : projet privé dont l'objectif est de recenser et de donner accès au maximum de donnés publiques (et privés) françaises. Les données sont téléchargeables au format .xls

Gapminder (http://www.gapminder.org/) : jeux de données d'organisations publiques téléchargeables au format xls. Gapminder est un des plus anciens services de mise à disposition de données. Il propose par ailleurs de télécharger gratuitement le *Gapminder desktop*, un logiciel qui permet de visualiser les jeux de données.

Infochimps (http://www.infochimps.com/) : portail donnant accès à des milliers de jeux de données gratuits et payants provenant d'organisations publiques ou privés du monde entier.

Data Market (http://datamarket.com/) : un concurrent direct d'Infochimps. Accès gratuit ou payant (59$/mois). Les jeux de données sont les mêmes mais la version payante propose des fonctionnalités supplémentaires comme l'envoi de rapports périodiques ou l'export selon différents formats.

Aggdata (http://www.aggdata.com/) : beaucoup de jeux de données relatifs aux enseignes commerciales américaines et à leur localisation (données type .kml). Payant.

Factual (http://www.factual.com/) : Jeux de données gratuits. Permet l'accès aux données sur son smartphone. Orienté vers les développeurs avec la mise à disposition d'API.

Numbrary (http://www.numbrar.com/) : à la fois annuaire de sites proposant des datasets et service d'hébergement de jeux de données

Open data search (http://opendatasearch.org/) : jeux de données classés par lieux, entités créatrices ou formats de fichiers.

Data360 (http://www.data360.org/) : portail permettant d'accéder à de très nombreux datasets provenant de services publics et d'organisations internationales.

Kngine stats (http://www.kngine.com/Stats): portail qui agrège et permet de visualiser des jeux de données publiques. Il s'agit d'un projet parallèle au moteur sémantique **Kngine** (cf. lettre R&R n° 110, décembre 2009)

CKAN (http://ckan.net/) : répertoire qui recense plus de 1 800 datasets mais c'est aussi, à l'instar de l'application Mediawiki utilisée par la Wikipedia, un logiciel serveur qu'une organisation peut utiliser gratuitement pour mettre à disposition ses données.

Open Data Directory (http://open.mflask.com/) : annuaire proposant plus de 300 000 datasets privés et publics. Recherche par mots-clés ou par sources (pays, organisations).

Données publiques sur Amazon web services (http://aws.amazon.com/fr/publicdatasets/) : essentiellement orientés vers des jeux de données à caractère scientifique. Exemple : projet *Ensembl Annotated Human Genome data.*

Windows Azure Marketplace (http://windowsazure.pinpoint.microsoft.com/) : service donnant accès à des jeux de données payants et gratuits dans de très nombreux secteurs d'activités.

Populationdata (http://www.populationdata.net/) : portail centré sur les données démographiques.

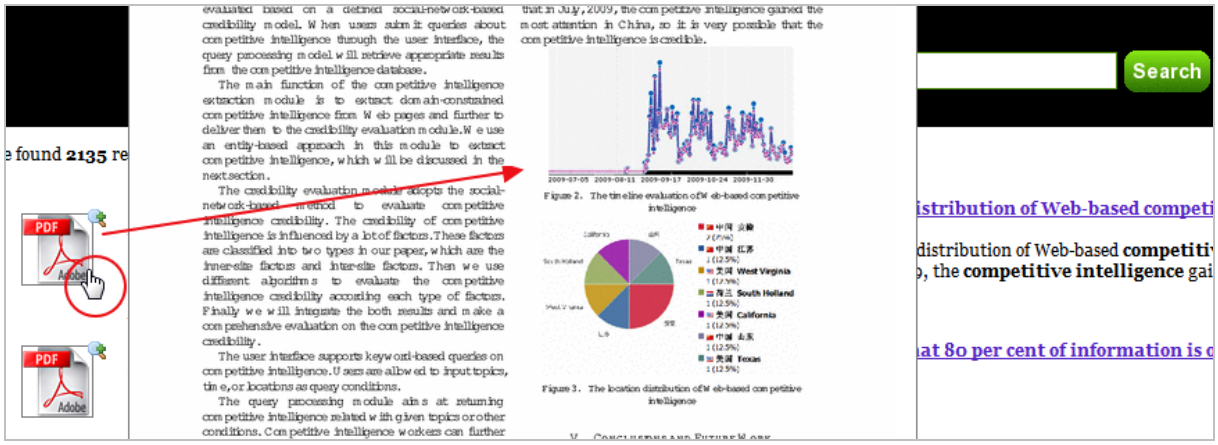
Google Internet Stats (http://www.google.co.uk/intl/en/landing/internetstats/) : service proposé par Google.uk, donnant accès à des données statistiques sur internet (taille, usages, technologies, secteurs d'activité,...).

Google public data explorer (http://www.google.com/publicdata/home) : service proposé par Google qui donne accès et permet de visualiser une quarantaine de jeux de données publiques.

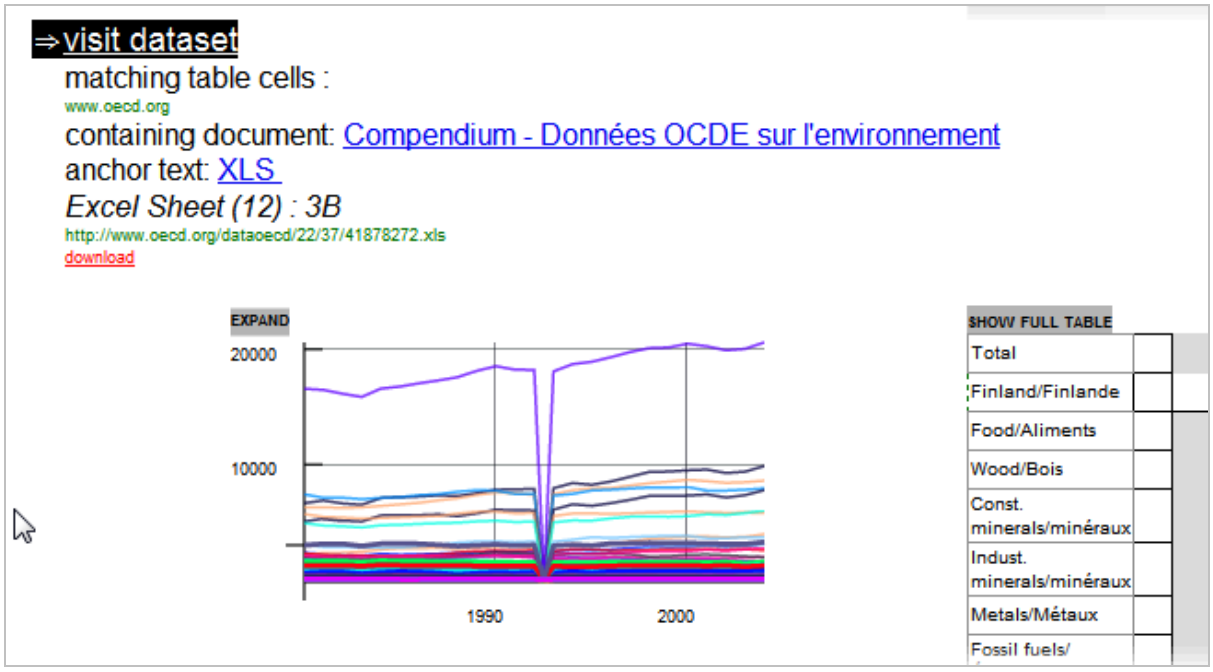
Moteurs de recherche :

Il s'agit ici de moteurs qui se sont spécialisés dans la recherche et l'indexation de jeux de données et qui proposent pour certains une interface de consultation directement accessible *via* la page de résultats.

Zanran (http://www.zanran.com/) : ce moteur se focalise sur les jeux de données présents dans les documents bureautiques et a la bonne idée d'afficher les contenus pertinents d'un simple passage de la souris sur un résultat. Très pratique.



D8taplex (http://d8taplex.com/) : moteur qui crawle le web pour indexer des jeux de données (feuilles de calcul, tables html, ...) et les interprète afin que l'utilisateur puisse les afficher dans un graphique et les télécharger.



On le voit, les services proposant des jeux de données en ligne sont extrêmement nombreux et la quantité de données structurées accessibles ne fera que croître à l'avenir. Ce mouvement est le résultat d'une dynamique qui intègre :

- Le citoyen/consommateur et son besoin d'être toujours mieux informé pour "piloter" sa vie et faire des choix qu'il veut rationnel.

- Les pouvoirs publics et gouvernement des pays occidentaux, de plus en plus sensibles à l'idée de transparence qui ne peut qu'être un gage de démocratie. Autrement dit, ne pas aller vers l'open data pourrait laisser entendre que vous n'êtes qu'une demi-démocratie ...

- La multiplication des capteurs de type puces RFID, qui diffusent des données en permanence et permettent de monitorer tous types d'activités.

- La multiplication d'objets connectés (internet des objets) qui communiquent entre eux et produisent également des données.

- L'augmentation de la mobilité individuelle rendue possible par les smartphones et autres terminaux nous permettant la réception et la "manipulation" de ces données dans n'importe quel contexte.

Attention tout de même à éviter quelques écueils évidents : utiliser des données statistiques nécessite de comprendre précisément le traitement qu'elles ont subi afin d'éviter l'effet "boîte noire" de certains logiciels permettant l'analyse de données pour tous. Par ailleurs il est toujours bon de rappeler que la quantité des informations disponibles et leur facilité d'accès ne présume en rien de leur qualité. Le "*garbage in, garbage out*" des informaticiens est donc plus que jamais de mise...

Christophe Deschamps
Consultant et formateur en gestion de l'information.
Responsable du blog Outils Froids (http://www.outilsfroids.net/)

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://blog-abonnes.abondance.com/2011/06/lopen-data-enjeux-et-outils.html