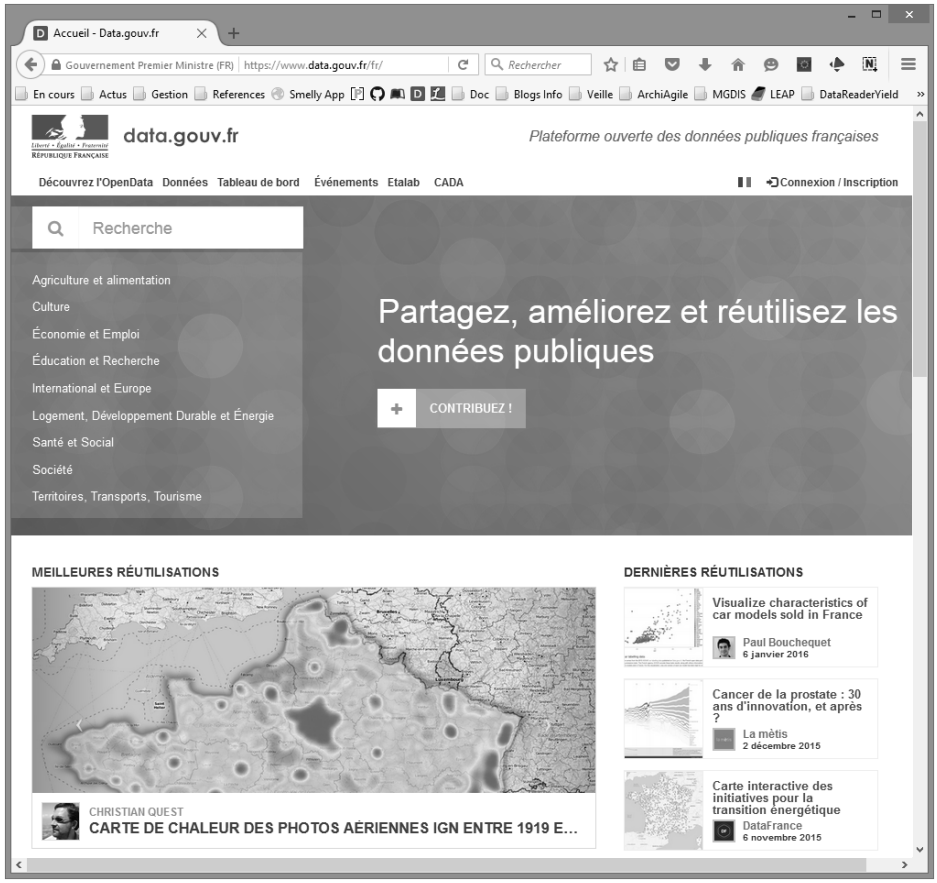
Chapitre 2
Consommer des flux Open Data

1. Trouver des flux

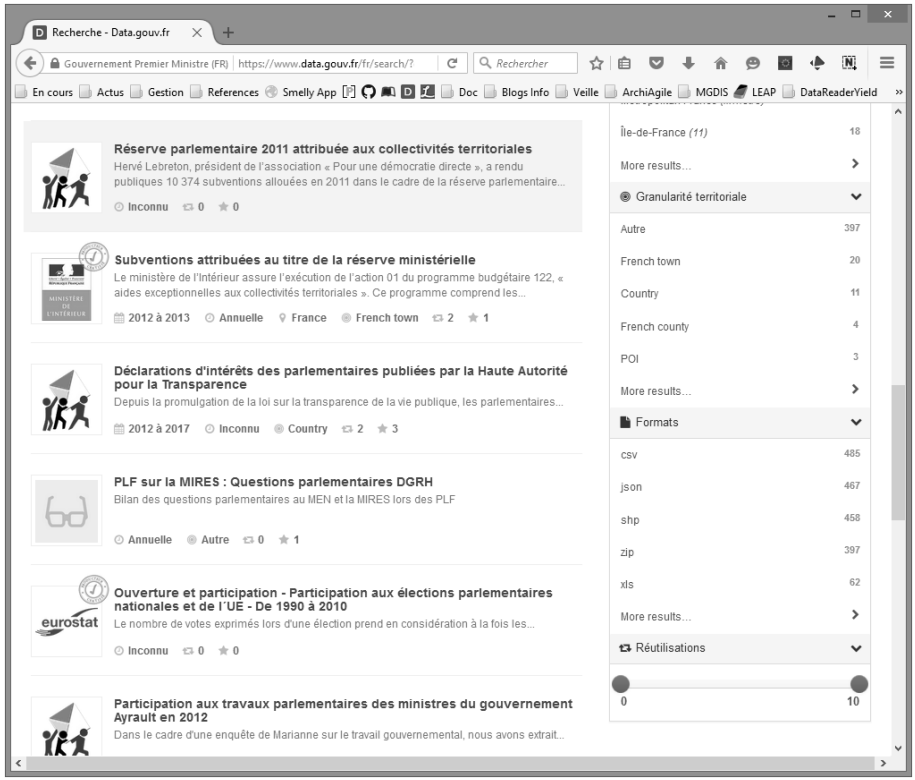
1.1 Data.gouv.fr

En France, la destination numéro un pour trouver de la donnée publique est le portail Open Data de l'État, fourni par ETALAB. ETALAB fait partie du Secrétariat Général à la Modernisation de l'Action Publique, et est donc directement rattaché au cabinet du Premier ministre, ce qui souligne l'importance de sa mission. Le site data.gouv.fr non seulement propose de la donnée, mais également agrège et centralise des données de nombreux autres fournisseurs au niveau national, comme l'INSEE, l'IGN, etc.

La page d'accueil du site propose un système de recherche par mot-clé ou par catégorie, puis affiche une sélection des jeux de données souvent ou dernièrement utilisés, ainsi que des mises en forme particulières. Le site supporte en effet le dépôt par tout un chacun d'une carte, d'un graphique ou d'un site apportant de la richesse à un jeu de données publié.



Lors d'une recherche, les résultats sont fournis sous forme de vignettes avec quelques métadonnées de base. Une bande de recherche sur la droite fournit des facettes qui permettent de filtrer les résultats plus avant. Ces facettes permettent de restreindre la donnée par la granularité territoriale, les formats de fichiers disponibles, le nombre de réutilisations publiées, etc.



1.2 Marchés de données

Il existe des sites spécialisés dans le recensement des sources de données en provenance d'origines multiples, et appelés marchés de données. L'un des plus connus est Azure Data Market, qui recense de nombreux jeux de données internationaux.

Même si cela paraît contre-intuitif, les portails internationaux peuvent se révéler d'excellentes sources pour de la donnée nationale. Le fait de disposer de données d'autres pays en plus de celui ciblé permet d'avoir un regard plus critique sur les valeurs de ce dernier, en les comparant, mais aussi en utilisant des critères qui ne sont pas nécessairement familiers alors qu'ils sont souvent utilisés ailleurs.

Pour ne donner qu'un exemple, les données sur la population issues des Nations Unies sont de très grande qualité. Elles compensent une granularité moins fine que celles de l'INSEE par la fourniture de nombreux indices permettant des comparaisons riches d'enseignements avec les autres pays dans le monde.

Les marchés de données recensent toutes les données, qu'elles soient ouvertes ou pas, gratuites ou pas, etc. Il convient donc de faire particulièrement attention aux licences.

1.3 Sites des collectivités

Comme expliqué dans le premier chapitre, les collectivités sont de plus en plus nombreuses à produire de la donnée sous forme publique. En réalisant une recherche sur le nom de la grande ville la plus proche, ou bien sûr celui du département ou de la région de rattachement, suivi des mots-clés "Open Data", il est courant de trouver un site dédié sur lequel une recherche pourra ensuite être réalisée plus en détail.

Ces sites dédiés sont aujourd'hui une cinquantaine en France, ce qui permet d'obtenir de la donnée assez finement localisée. Leur proximité par rapport au consommateur permet de garantir une grande pertinence en général, car les évènements et critères considérés comme les plus importants par les usagers principaux sont mis en avant.

1.4 Demander des flux

De par leur taille réduite par rapport à de grands ensembles institutionnels, les collectivités locales sont assez accessibles aux usagers. Comme la responsabilité des portails Open Data est en général concentrée sur quelques personnes seulement, il est relativement facile de demander des données supplémentaires, de proposer des améliorations, sachant que les équipes en charge sont souvent justement à la recherche de telles demandes. Leur provenance en direct des utilisateurs leur permet d'être plus sûres de leur alignement sur les besoins réels.

Si une recherche sur Internet ou sur les portails Open Data locaux ou nationaux ne donne pas satisfaction, il ne faut donc pas hésiter à solliciter les personnes compétentes. Tous les sites disposent d'une rubrique "contact" à cet effet.

2. Principes de consommation

2.1 Les questions à se poser

La question de la recherche de données a été traitée ci-dessus assez rapidement car les exemples dans les trois chapitres qui suivent vont donner de nombreuses pistes complémentaires. De la même manière, les façons de consommer ces données, de les nettoyer ou les analyser, vont être traitées en profondeur par la suite, donc cette section a seulement pour objectif de donner les grands principes de consommation de la donnée.

Le lien étant fort entre Open Data et open source, l'auteur a cherché dans cet ouvrage à équilibrer au maximum les usages d'outils propriétaires avec ceux d'outils libres, ou à défaut disposant d'une version communautaire gratuite. Tous les outils sont accessibles financièrement, y compris à des particuliers, et donc également à des PME ou des administrations consommatrices de taille réduite.

Outre les outils, quelques questions sont à se poser avant de consommer la donnée. Elles peuvent paraître des évidences une fois énoncées, mais il n'empêche qu'elles constituent une première étape de sélection permettant de transformer une demande fonctionnelle (le souhait du consommateur d'obtenir de l'information) en un ensemble d'exigences techniques (toutes les questions détaillées auxquelles il faudra répondre pour, partant de rien, dénicher la bonne donnée et la transformer en information).

Ces questions sont nombreuses mais les principaux critères sont les suivants :

– Consommation d'un fichier ou d'une API ? Si le format n'importe guère car l'analyse sera faite une seule fois, l'approche n'aura rien à voir avec une analyse présentée comme la plus automatique possible car elle devra être souvent mise à jour, et si possible sans aucune intervention humaine. Dans ce second cas, la disponibilité d'une API pour consommer la donnée sera essentielle.
– Données propres ou à nettoyer soi-même ? Pour une même donnée, il existe parfois des dizaines de sources différentes. Certaines contiennent des informations pour toutes les périodes dans le temps, mais pas de valeur pour tous les attributs. D'autres présentent les propriétés inverses. D'autres encore posséderont toutes les colonnes souhaitées sur une chronologie large, mais contiendront des blancs ou des erreurs. Il est important de réfléchir à ce qui est le plus important pour ne pas se retrouver à passer énormément de temps à créer la source ultime de grande qualité si le résultat n'en vaut pas le temps passé.
– La donnée est-elle statique ou dynamique ? Parfois, seules les périodes correspondant au passé sont statiques, et il convient de récupérer régulièrement la donnée pour l'année en cours de façon à être en mesure d'analyser une évolution fine. Dans ces cas, les méthodes de consommation (et d'agrégation du passé et du présent) ne seront pas les mêmes.

2.2 Le choix du bon outil

Après toutes ces questions sur la façon de trouver le bon jeu de données et de le consommer correctement vient la question de l'outil d'analyse lui-même. Il est essentiel de bien connaître la palette de logiciels disponibles (et c'est un des objectifs de ce livre), de façon à choisir le bon outil pour la bonne manipulation. Tous possèdent à peu de choses près les mêmes fonctionnalités de base, mais certains se distinguent par leur approche ergonomique, d'autres par leurs fonctionnalités avancées.