## Trouver les données cachées sur le Web

‘Open data’ signifie bien plus que de simples séries de données disponibles à télécharger. Les données téléchargeables ne représentent qu’une partie infime des données disponibles sur le Web.

La plupart des données disponibles sur le Web ne sont pas visibles pour l’humain. Cependant les machines sont capables de trouver et de lire ces données. Lors de ce module, nous verrons quelles techniques utiliser pour déverrouiller les données cachées.

##### Dans ce module, nous allons explorer les points suivants :

* Comment trouver des données cachées
* Quels sont les avantages des données cachées
* Comment obtenir des données cachées

## Des données invisibles pour l’humain

Les données ne sont pas toujours faciles à trouver sur le Web parce que la plupart des ressources disponibles ne sont visibles que par des machines. En général, les internautes surfent le Web de manière visuelle : ils visualisent le contenu, cliquent sur des liens et téléchargent des fichiers.

Les machines naviguent le Web d’une façon bien différente, en utilisant uniquement les règles logiques contenues dans le code. Alors que les humains recherchent un contenu engageant et interactif, les machines demandent de la structure, de la logique et de la clarté.

Cette différence crée un ‘angle mort’ pour les humains ou les machines, selon leur manière de lire le Web.

Ainsi, il arrive que les machines ne trouvent pas immédiatement le lien de téléchargement sur une page ou que les humains aient des difficultés à voir les données structurées et non-visuelles qui se trouvent dans le code.

Les fichiers téléchargeables d’open data ne sont qu’un type de données sur le Web parmi d’autres. La majorité des données sont cachées dans le code d’un site Web.

## Trouver les données cachées

Certaines données ne sont pas disponibles au téléchargement. Elles sont souvent cachées au sein des différents sites Web que l’on visite quotidiennement.

### Pourquoi cacher les données ?

La nouvelle tendance du Web, aujourd’hui centré presque exclusivement sur les données et des fonctions compatibles avec les Smartphones et autres gadgets mobiles, a révolutionné la manière de concevoir les sites web. Afin de satisfaire les exigences du consommateur quant à l’interactivité et l’aspect visuel des sites, les développeurs ont de plus en plus tendance à mettre les données dans une ‘couche’ distincte ou encore dans le code lui-même. Cette couche de données est souvent accessible à ceux qui disposent des bons outils.

### Les données se couchent partout

Des pages Wikipedia aux sites de vacances en passant par des sites de mode et de fournisseurs d’énergie, le Web foisonne de sources de données.

Les licences concernant ces données cachées ne sont pas toujours claires : ce n’est pas parce que des données sur un site Web public sont lisibles par une machine que celles-ci sont ouvertes. Si vous avez le moindre doute quant à la licence, vérifiez avec l’éditeur du site.

### Quels types de données se cachent ?

L’apparence des données cachées est totalement différente de celle d’un ensemble de données ouvertes que l’on aurait téléchargé sur un site Web.

Les données lisibles par des machines sur le Web se présentent souvent sous format JSON, le format le plus courant et le plus facile à utiliser, XML ou parfois RDF.

Complétez les exercices plus tard dans ce module et essayez d’extraire des données sous ces formats.

## Les avantages des données cachées

### Qualité

Dans la plupart des cas, les données cachées sont celles utilisées par l’éditeur du site pour présenter la page visible sur le site.

Les données cachées jouent donc un rôle particulièrement important dans leurs opérations en ligne et il est fort probable que l’éditeur prenne grand soin de leur qualité.

Les données extraites du Web ont tendance à être plus fiables car il s’agit probablement de données clés quant aux opérations d’une organisation. Par conséquent, il s’agit souvent de données avec lesquelles on peut travailler en toute confiance.

### Pertinence

Les organisations ne publient pas toujours des données téléchargeables. Cependant, beaucoup de leurs sites sont centrés sur les données. En accédant aux données grâce aux flux de données existants (ou à partir du code du site Web), il y a déjà une utilisation claire pour les données, ce qui peut vous aider à identifier rapidement les données pertinentes plutôt que de les trouver vous-même. Consultez par exemple les pages des informations de la BBC et les flux de données associés.

[Lien vers données pertinents du BBC](http://www.bbc.co.uk/news/10628494)

### Contexte

Le fait d’accéder aux données directement sur le Web implique que celles-ci sont contextualisées. Nous avons déjà expliqué l’importance du contexte des données pour leur utilisation lors de précédents modules.

### Ampleur

Une organisation est souvent limitée quant à la quantité de données téléchargeables qu’elle peut publier. Le fait de trouver des données cachées offre une plus large gamme de ressources.

## Comment trouver les données cachées ?

Grâce aux données, on obtient de l’information ; grâce à l’information, on accède à la connaissance et à la compréhension.

Veuillez noter que le simple fait qu’il soit possible d’extraire des données cachées ne les rend pas pour autant ouvertes. Il faut toujours vérifier les conditions générales de la licence du site Web avant de recueillir ou d’utiliser des données, quelle que soit l’utilisation prévue. Pour plus de détails sur les licences, veuillez vous référer au module 4.

### Le téléchargement des extensions

Certains sites Web sont conçus avec la possibilité d’en extraire les données en ajoutant une simple extension à l’URL. Pour ces sites, souvent maintenus par des organisations qui publient aussi de l’open data téléchargeable, le fait d’ajouter la bonne extension déclenche un téléchargement de la page sous forme de données plutôt que sous le format de navigation.

Le site Web du gouvernement britannique en est un bon exemple : chaque page peut être accessible sous forme de données tout simplement en ajoutant l’extension nécessaire comme ‘.json’, par exemple https://www.gov.uk/browse/business.json.

Essayez avec les données de commerce du Royaume-Uni. Allez au site Web et ajoutez ‘.json’, à l’URL pour obtenir les données.

[Lien vers les données de commerce du gouvernement britannique](https://www.gov.uk/trade-tariff/sections)

Afin de voir les données sous un format plus lisible par l’humain, copiez-les sous jsonlint.com

### Les flux des données

Beaucoup de sites Web proposent des liens vers des flux de données auxquels on peut s’abonner. Ces flux peuvent être des flux RSS, des flux d’alertes Twitter ou encore d’autres types de flux moins courants. On utilise les flux pour obtenir des données brutes concernant une vaste gamme de contenu.

Pour un exemple de flux de données brutes, voyez le flux de BBC news sous format XML.

[Lien vers données pertinents du BBC en XML](http://feeds.bbci.co.uk/news/rss.xml)

Vous pouvez utiliser un outil gratuit en ligne qui rend ces données plus faciles à utiliser.

[Allez vers codebeautify et entrez l’URL du flux BBC dans la section ‘Load URL’](http://codebeautify.org/xmlviewer)

### Le code source

Au delà des sites Web qui offrent un accès simple aux données, beaucoup de sites intègrent leurs données dans le code source de la page Web.

Pourquoi ne jetteriez-vous pas un œil à l’outil expérimental de l’ODI ‘Hidden Data Extractor’ (Extracteur de Données Cachées) pour trouver les données cachées dans le code source d’un site Web ?

[Lien vers le ‘Hidden Data Extractor’ de l’ODI](http://odinprac.theodi.org/hidden-data-extractor/)

### API ou interface lisible par une machine

Certains sites ou services Web prévoient une interface lisible par machine, ou API, pour interroger ou accéder à leurs données.

Voici quelques exemples de services utilisant une API :[Open Corporates](https://api.opencorporates.com/) (ouvert), [OpenSteetMap](http://wiki.openstreetmap.org/wiki/API) (ouvert), [Twitter](https://dev.twitter.com/rest/public) (pas ouvert ), [Flickr](https://www.flickr.com/services/api/) (quelque contenu ouvert), [LinkedIn](https://developer.linkedin.com/docs/rest-api) (pas ouvert).

La plupart des services utilisent une API Web et permettent l’ajout d’extensions (essayez ‘.json’ sur OpenCorporates). Bien que cette API ressemble aux exemples de sites cités précédemment, où il suffit d’ajouter une extension pour obtenir les données, le statut formel d’OpenCorporates en tant qu’API donne plus de certitude quant à la disponibilité du service.

### Le ‘scraping’

Si aucune des techniques mentionnées jusqu’à présent ne fonctionne, vous pouvez avoir recours à la technique dite de ‘scraping’ (utilisation d’un programme qui extrait l’information d’un site) sur les sites lisibles par l’humain. Bien que cette technique soit relativement fiable, vous n’êtes pas à l’abri de changements dans la disposition ou le contenu du site.

N’hésitez pas à essayez magic.import pour constater par vous-même à quel point cette technique est facile.

[magic.import.io](http://magic.import.io)

## Êtes vous prêt(e) à trouver les données cachées sur le Web?

Il y a 3 choses utiles à prendre en trouver les données cachées sur le Web. Vous vous en souvenez ?

## Qu’est-ce que les données cachées ?

Les données cachées...

* Les indices d’une chasse au trésor
* Des données visibles pour des machines mais invisibles pour l’humain
* Des fichiers difficiles à télécharger

##### Vous avez raison !

Les donnés cachées sont ne sont pas visibles à l’œil nu lorsqu’on surfe le Web mais une machine saura les trouver et les comprendre facilement.

##### Êtes vous sûr(e) ?

Les donnés cachées sont ne sont pas visibles à l’œil nu lorsqu’on surfe le Web mais une machine saura les trouver et les comprendre facilement.

## Pourquoi les données sont-elles cachées ?

Les gens cachent les données parce que...

* Ils en ont honte
* Ça fait désordre
* La plupart du temps, ce n’est pas volontaire

##### Vous avez raison !

La plupart des données ne sont pas cachées délibérément. Souvent, la société qui construit le site Web se concentre sur la façon dont les internautes naviguent la page Web et ne considère pas forcément la façon dont les données sont fournies.

##### Êtes vous sûr(e) ?

La plupart des données ne sont pas cachées délibérément. Souvent, la société qui construit le site Web se concentre sur la façon dont les internautes naviguent la page Web et ne considère pas forcément la façon dont les données sont fournies.

## Les avantages des données cachées

Lequel des points suivants n’est pas un avantage des données cachées ?

* Ampleur
* Qualité
* Lisibilité
* Pertinence
* Contexte

##### Vous avez raison !

Trouver des données cachées peut augmenter la pertinence et la qualité des données, l’ampleur disponible et souligner le contexte dans lequel on utilise les données.

##### Êtes vous sûr(e) ?

Trouver les données cachées peut augmenter la pertinence et la qualité des données, l’ampleur disponible et souligner le contexte dans lequel on utilise les données.

## Comment extraire les données cachées d’un site Web ?

Lequel des points suivants décrit une façon utile d’extraire les données cachées sur un site Web ?

* Mettre le lien du site Web en cache
* Recharger la page
* Inspecter le code source

##### Vous avez raison !

Inspecter le code source d’un site Web révèle les données cachées intégrées dans la page.

##### Êtes vous sûr(e) ?

Inspecter le code source d’un site Web révéle les données cachées intégrées dans la page.

## Trouver les données cachées

Les données téléchargeables ne représentent qu’une quantité infime des données disponibles sur le Web.

Alors que l’œil humain ne perçoit pas toujours ces données, on peut identifier et récolter une vaste quantité de nouvelles données en utilisant des machines.

Les données obtenues sur ces sites Web sont souvent plus conséquentes et pertinentes que des fichiers téléchargeables et puisent directement du contexte déjà utilisé par l’éditeur.

En augmentant vos connaissances en données cachées et en apprenant quelques approches simples, telles que l’utilisation d’extension et les API, vous aurez accès à un vaste éventail de données pertinentes pour votre travail à venir.