# Finding hidden data on the Web

‘Open data’ does not only mean datasets available to download. Downloadable open data represents only a small fraction of the available data on the Web.

The majority of data available on the Web is hidden from the human eye. However, machines can find and read this data.

**In this module we’ll explore the following:**

How to locate hidden data

What benefits hidden data can provide

How to obtain hidden data

# Finding hidden data - in 2 minutes

ODI Trainer David Tarrant explains what hidden data is, what benefits it can provide and how you can obtain it.

# The data people can't see

Data is not always easy to find on the Web, because most of the available resources are visible only to machines. People navigate the Web visually, viewing content, clicking on links and downloading files.

Machines, by contrast, only navigate the Web through logical rules set out in the code. While humans look for engaging, interactive content, machines require structure, logic and clarity.

This difference creates ‘blind spots’ for how people and machines read the Web.

For example, machines cannot always instantly find a download link in a page; humans find it difficult to see structured, non-visual data contained in website code.

Downloadable open data files are just one type of data on the Web. Much more data remains hidden in website code.

# Finding hidden data

Not all data is available for download. Often it is hidden in the different websites that we visit every day.

**Why hide data?**

The shift towards the data-driven, mobile-enabled Web has led to a change in the way websites are built. To meet user expectations of interactive, visual websites, developers increasingly place data in a separate ‘layer’ or even in the code itself. This data layer is often accessible to people using the right tools.

## Data is hiding everywhere

From Wikipedia pages to holiday planners, fashion websites to energy providers, the Web is full of data sources.

The licences for this hidden data are not always clear; just because the data is machine-readable on a public website does not make it open data. If you are uncertain about whether hidden data is openly licensed, check with the website publisher.

**What type of data is hiding?**

Hidden data looks entirely different from an open dataset you download from a website.

Machine-readable data on the Web will be in formats like JSON – the most common and simple to use – embedded within HTML, in XML and sometimes in RDF.

Try the exercises later in this module to see if you can extract data in these formats.

# Benefits of hidden data

## Quality

In many cases hidden data will be data used by a publisher to present the webpage you are looking at on

their site.

The hidden data is therefore an important part of their online operations and it is likely that the publisher takes care over its quality.

Data extracted from the Web tends to be more reliable as it is likely to be data at the core of an organisation's operations. Therefore, it is more likely to be a dependable data source to work with.

## Relevance

Organisations do not always publish downloadable data. However, many of them run data-driven websites. By accessing the data directly through existing data feeds (or from within the website’s code) there is a clear use for the data already. This will help you to quickly identify relevant data from this existing use, rather than having to establish the use yourself. For example, the BBC News pages, sections and related data feeds

[Take me to BBC news feeds](http://www.bbc.co.uk/news/10628494)

## Context

When you access data directly from the Web you are already viewing it in a context it is used in by the publisher. As we have covered in previous modules, the context of the data is an important component of its use.

## Breadth

Often an organisation is limited in the amount of data it can publish as downloadable data. Finding hidden data opens up a broader range of resources.

# How to find hidden data

From data we gain information, from information we gain knowledge and insight.

Please note that the fact you can extract hidden data does not make the data open. You should always check the website's terms of use and licences before gathering or using data for any purpose. For more information on licensing, please refer back to module 4 of this course.

## Extension downloads

Some websites have been built to offer a way to extract data by adding an extension to the URL of the webpage you are viewing. For such websites, usually maintained by organisations who also publish downloadable open data, adding the correct extension will trigger a download of that page in a data format, as opposed to a browser format.

A good example of this is the UK Government website (gov.uk) which provides any page in a data format simply by adding the relevant extension like “.json”, for example www.gov.uk/browse/business.json.

Try it with the UK Trade data by viewing the following page and then adding “.json” to get the data.

[Take me to the UK gov trade tariff](https://www.gov.uk/trade-tariff/sections)

To view the data in a more human readable form, copy it into jsonlint.com.

## Data feeds

Many websites provide links to data syndication feeds that can be subscribed to.

These feeds may be RSS feeds, Twitter alert feeds or other less common types. Feeds can be used to get raw data about a broad selection of content.

For an example of a raw data feed, look at the BBC news feed in XML.

[Why not try out the news feeds from the BBC in XML?](http://feeds.bbci.co.uk/news/rss.xml)

Now you can use a free online tool to make that data easier to use [Go to codebeautify and enter the BBC feed URL under 'Load URL'](http://codebeautify.org/xmlviewer)

or get alerts for flooding from your nearest river (UK only).

[Take me to river level guages](http://www.gaugemap.co.uk/)

## Source code

Going beyond providing a simple route to data, many sites have data embedded in the webpage's source code.

Why not check out the ODI's experimental 'Hidden Data Extractor' tool to find source code?

[Take me to ODI Hidden Data Extractor](http://odinprac.theodi.org/hidden-data-extractor/)

## APIs or machine-readable interfaces

Some websites or Web services will expose a machine-readable interface, or API, for querying and accessing their data.

Examples of services that have APIs include: [Open Corporates](https://api.opencorporates.com/) (open), [OpenSteetMap](http://wiki.openstreetmap.org/wiki/API) (open), [Twitter](https://dev.twitter.com/rest/public) (not completely open), [Flickr](https://www.flickr.com/services/api/) (some content open), [LinkedIn](https://developer.linkedin.com/docs/rest-api) (not open).

Most of these services use a Web-based API and allow addition of extensions (try “.json” on

OpenCorporates). Although this API looks the same as the earlier example in this section, its formal status as an API provides a guarantee that the service will be available.

## Scraping the Web

If none of these techniques work, you may need to scrape the human-readable webpages. While this can be reliable, there is a risk of the layout changing as well as the content.

Why not try out [magic.import.io](http://magic.import.io/) to see how easy it can be?

**Are you ready to find hidden data?**

Can you recall the key things about hidden data?

**What is hidden data?**

**Hidden data is...**

part of a treasure hunt data visible to machines but invisible to human eyes files that are difficult to download

**That’s right!**

Hidden data is data that is invisible to people as they navigate the Web but that a machine can easily find and understand **Are you sure?**

Hidden data is data that is invisible to people as they navigate the Web but that a machine can easily find and understand

**Why is data hidden?**

**People hide data because...**

they are ashamed of it it is messy most of the time they intentionally haven't exposed it

**That’s right!**

Most hidden data is not deliberately hidden. Often the organisation building the website is concentrating on how humans navigate the pages and therefore doesn't think about how the data is provided.

**Are you sure?**

Most hidden data is not deliberately hidden. Often the organisation building the website is concentrating on how humans navigate the pages and therefore doesn't think about how the data is provided.

# Benefits of hidden data

**Which of the following is not a benefit of hidden data?**

Breadth

Quality

Readability

Relevance

Context

**That’s right!**

Hidden data can increase the relevance and quality of the data, the breadth available and highlight the context in which the data is used.

**Are you sure?**

Hidden data can increase the relevance and quality of the data, the breadth available and highlight the context in which the data is used.

**How might you extract hidden data from a website?**

**Which of the following is a useful way to extract hidden data from a website?**

Caching the website link

Reloading the page

Inspecting the source code

**That’s right!**

Inspecting the source code of a website often reveals hidden data embedded in the page.

**Are you sure?**

Inspecting the source code of a website often reveals hidden data embedded in the page.

# Finding hidden data on the Web

Data available for download is only a small fraction of data available on the Web.

While the human eye cannot always see data, it is possible to identify and grab a wealth of new data using machines.

The data accessed from these websites is often much broader and more relevant to your work than downloadable files and also draws directly from the context in which the publisher is already using it.

By raising your awareness of hidden data and learning some of the simple approaches in this module, such as extension queries and APIs, you will be able to access a broad range of relevant data for your future work.

Next module

Main menu

# Module 12 – video

I’m Dave Tarrant, I’m the Trainer and Data Scientist here at the open data institute. Open data is not always easy to find. We as humans consume things visually, and so we looking at pretty webpages and images on these webpages that cue us to where we can get data from and download it. Computers don’t operate in this way they like structure, logic, clarity and mainly text. So the challenge for us is to understand how a machine likes to consume data and be able to get into its mind-set to be able to excavate this hidden data. The web is a data rich environment, and many websites already bring together data from various different sources in order to present you with a collated set of information. So think of planning a holiday. Many holiday websites already bring together a rich amount of location data pictures, reviews, prices and maybe even transport data that allows you to plan your route to get to a certain place. So what we want to be able to do is realise that we can exploit these same data connections that exist between the websites, so that we can get access to the data ourselves to present it in a whole new context to build a new set of information and thus a new set of knowledge. Some websites give you access to data simply by adding and extension to the end of the page you are on. So if you can see that there may be some sort of data on the page you might be able to add like a .csv to the end of the page and it might download you some data. One of the most common formats online is actually the .JSON format and you can get access to all of the data off of gov.uk by adding a .json to the end. Other websites might give you feeds of data, such as news websites and we’re used to consuming these within mobile applications. Beyond these it gets a little more tricky some data you may have to look for from the source code of the page where it’s been embedded, these websites are already using it so you may as well look for it within the page itself. Finally, many sources of data give you a specific way to access it a machine interface or an API, so you might have to learn how to use one of these. So by using data that’s already an integral part of an existing service you can guarantee a higher quality of relevance and usability. Additionally, you’re accessing it in the context in which it’s already used rather than through a separate publication platform.