# Choosing the right format for open data

The 'format' of an open dataset refers to the way in which the data is structured and made available for humans and machines.

Choosing the right format helps ensure the data can by simply managed and reused. To maximise reuse of data, it may be necessary for a publisher to use a number of formats and structures available across different platforms to suit users' needs.

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

Why formats matter to open data

Choosing the correct structure

Accessing different open data formats

Keeping it simple with CSV

# Open data formats

ODI Trainer David Tarrant explains why formats matter in open data, how to choose the right structure and formats, and why CSV is a good starting point for most tabular open data.

# Usability, management and access

Open data must be available for anyone to access, use and share. In this section we consider what formats maximise the usability of data, ease of access for users and simple management for publishers.

## Usability

The most usable format for data is likely to be one in which the dataset was first created. In many cases, this format may be a proprietary software programme like Microsoft Excel. While such programmes often have the ability to export data in other formats, their popularity can make publishing in a different format a barrier to reuse by others.

Software programmes like Excel also come with richer content, such as styling on tables and graphs of data. These can all help to give context to a human who is trying to understand the data.

## Management

For some publishers, open data is now the main source for their data, even for staff in the publishing organisation. Using open data as a way for staff to access information inside an organisation has many benefits:

It increases knowledge of open data within the team.

It increases sustainability and trust in the data, by making it a key component of the organisation’s operations.

It creates clear opportunities for innovation.

## Access

For open data to be as accessible as possible, it must be:

In a format that the user can understand.

A format that machines can read.

In a format that supports the data's easy reuse.

In a format that doesn’t require expensive tools to access.

The format appropriate for each dataset may be different. A document (like a PDF) may make the data easy for a human to understand but not the most readable for a machine. For this reason, accessibility is often achieved by making the same data available in a range of different formats.

# Common data structures

Not all data can be appropriately expressed as a spreadsheet. Different file formats may have to be considered. There are three key structures to be aware of: tabular, hierarchical and network.

## Tabular

The most common structure for data is tabular. Data is organised into rows and columns listing sequential values, such as expenditure.

If the data is based on separate entries that are not linked to each other then a tabular file structure in a format such as CSV is ideal. An example of appropriate tabular data is the Italian data portal’s museum attendance figures.

[Take me to Italian Data Portal](http://www.datiopen.it/opendata/Visitatori_musei_pubblici_e_similari_titolo_d_accesso)

## Hierarchical

Hierarchical data shows the relationships between data points, such as a family tree or municipalities in each country. If the dataset depends on the relationship between data points and follows a structure in which data points are linked in vertical ‘trees’, a hierarchical data structure in a format such as JSON is ideal. An example of appropriate hierarchical data is this JSON file from json.org [Take me to the json.org example](http://json.org/example.html)

## Network

Network structured data allows relationships to exist between any combination of elements in any direction.

A good example of a network data structure is a social network. Think of your network of friends and their friends on Facebook; consider first, second and third degree contacts on LinkedIn.

The Web is another example of a network data structure, where webpages link to any number of other pages in any direction.

# Finding open data in the right formats

As a user of open data, it is important to understand how publishers are delivering different formats to you for reuse.

Not all data is suitable for download either because it is too large, too regularly updated or too complex to expose as static files. Some data may need to be split into smaller parts. Other data may need to be published as live feeds to make sure it stays up-to-date.

## Downloadable open data

Tabular data is the best suited for download. This is why most government open data portals are predominantly tabular data. However, managing millions of rows of data may need other considerations:

Should the data be split into smaller datasets?

How often should the data be updated?

How will changes in the way you publish affect the previous releases?

## Live data and data feeds

Some data is not suitable to be made available as a downloadable file.

Much of this data is updated so regularly that the file downloads would be too large for most users. This type of open data structure may be made available by a machine interface, also know as an application programming interface (API).

There are many services that make machine interfaces available over the Web. These services can be directly integrated into other Web applications. For an example of an open data API, see at the Belgian railway company.

[Take me to iRail API](https://hello.irail.be/api/1-0/)

# Choosing the right format for open data

When it comes to open data formats, start with CSV.

A comma separated values (CSV) file is simply lines of data, with each data point separated from the next by a comma. CSV is perfect for tabular data and can be easily loaded into and saved from applications like Excel, making it accessible to users.

Although CSV doesn’t maintain formatting and graphs like Excel formats, it is an open, machine-readable format. CSV represents the simplest format that still supports broad reuse of open data. In other words, CSV is the ‘lowest common denominator’ for open data – open data should be made available in this format wherever possible?

# Geospatial open data formats

Geospatial data is often more complex than simple tabular datasets.

It can exist as a hierarchical dataset, detailing countries and counties/states, or as a network dataset, detailing roads.

When publishing this type of data, formats like [geoJSON](https://en.wikipedia.org/wiki/GeoJSON) (based upon JavaScript Object Notation - JSON) and [KML](https://en.wikipedia.org/wiki/Keyhole_Markup_Language) (based upon Extensible Markup Language - XML) should be considered.

These formats are specifically designed with usability in mind and can easily be imported and exported from specialist mapping tools like [Open Street Map](http://www.openstreetmap.org/) and [CartoDB](https://cartodb.com/).

**Are your formats in order?**

Test your knowledge on open data formats with our questions. Can you recall the important points?

**Why is it important to pickÂ the right format?**

**It is important to pick the right format to...**

Ensure that privacy is protected

Drive down costs

Ease, usability, management and access

**That’s right!**

Picking the correct formats will assist usability, make management easier and lower the barriers of access. Choices here will not affect privacy and may even increase cost **Are you sure?**

Picking the correct formats will assist usability, make management easier and lower the barriers of access.

Choices here will not affect privacy and may even increase cost.

**Which data structure would you use for the data in the picture?**

**Which data structure would you use for the data in the picture?**

Tabular

Hierarchical

Network

**That’s right!**

The picture shows that relationships between data points are important. These relationships are all 'parentchild' relationships, meaning that a hierarchical structure is ideal.

**Are you sure?**

The picture shows that relationships between data points are important. These relationships are all 'parentchild' relationships, meaning that a hierarchical structure is ideal.

**In which format should open data be made available wherever possible?**

**What format should open data be made available in wherever possible?**

CSV

PDF

DOC

**That’s right!**

The comma separated values (CSV) format is a simple to understand, highly reusable format. It is also open for anyone to write new tools to interpret it. The other formats are useful but not as widely adopted for data.

**Are you sure?**

The comma separated values (CSV) format is a simple to understand, highly reusable format. It is also open for anyone to write new tools to interpret it. The other formats are useful but not as widely adopted for data.

# Open data formats

The 'format' of an open dataset refers to the way in which the data is structured and made available for humans and machines.

Choosing the right format helps ensure the data can by simply managed and reused. To maximise reuse of the data, it may be necessary for a publisher to use a number of formats and structures available across different platforms to suit users' needs.

Wherever possible, the CSV file format should be used for sharing open data. CSV is a simple to understand, highly reusable and machine-readable format.

Next module

Main menu

# Module 9 – video

I’m Dave Tarrant I’m the Trainer and Data Scientist here at the open data institute. So the format of open data refers to how that data is made available and understandable for both humans and machines and as well as considering the format it’s also critical to consider the structure of the data and also how it is delivered to the users. So choosing the right format is critical to maintain the usability and the access and there are three main characteristics you choose the format, the structure and its delivery mechanism. So when choosing the format, the most reusable format is likely to be that in which it is already managed and used. This will be closest to that context in how it’s used currently but this format might not be the most widely available, widely understood so it might be necessary to publish in a format like .csv in addition to that other format. So as you can see, you should also consider publishing in a number of different formats and also be prepared for any particular format’s obsolescence. When it comes to the structure there are several different structures that data can exist in. So the most common one again is tabular, where you’ve got things in rows and columns this is used for things like expenses data. But there’s also hierarchical data which represents relationships between people so think of your family tree and there is also network data which is very hard to think of in terms of a file format, but this represents the links such as those between your friends on Facebook, they can go in any direction. And finally, you need to consider the delivery mechanism so do you make it available as a file download? And if you do, you should carefully consider the size so that it’s easy to download quickly and it’s not going to take a long time the frequency of which it’s updated so does it need to be updated regularly, monthly and also the terminology that you use, to make that understandable. You should also consider if the data is more live like transport data, that it might be more wise to provide a machine readable feed of that data and that would be more appropriate than a file based delivery. So we encourage the broadest range of formats structures and delivery mechanisms for data to be used. But core to all of these is the .csv format the most commonly used open data format online. Beyond .csv when deciding other format structures and delivery mechanisms it is critical to engage users and for users to engage providers of data. This saves both parties a lot of time.