Open data vs FAIR data

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 741527 and runs from May 2017 to April 2021.

What is Open Data?

FAIR

Now that we have explored what FAIR data looks like, let's take a closer look at the concept of **open data**.

Open Knowledge Foundation definition:
"Data is open if it can be freely accessed, used, modified and shared by anyone for any purpose - subject only, at most, to requirements to provide attribution and/or share-alike. Specifically, open data is defined by the Open Definition and requires that the data be:

Open data is typically the product of *open science* or *open research*, and for the purposes of this module we'll assume this to be the case, while in some cases it may retrospectively be made open.

A. Legally open: that is, available under an open (data) license that permits anyone freely to access, reuse and redistribute

B. Technically open: that is, that the data be available for no more than the cost of reproduction and in machine-readable and bulk form.

ORION open science

Icons made by https://www.flaticon.com/authors/eucalyp

Is Open Data Equal to FAIR Data?

A common misconception is that data that is FAIR compliant has to be open at the same time.
This is definitely not the case and a common phrase used in the community is *"open as possible, closed as necessary".*

There are numerous examples where data cannot be made open by default, at least in the first instance, including data that is subject to IPR issues and data that is sensitive in nature.
In the context of this MOOC, we will concentrate on the latter since patient data is sensitive.

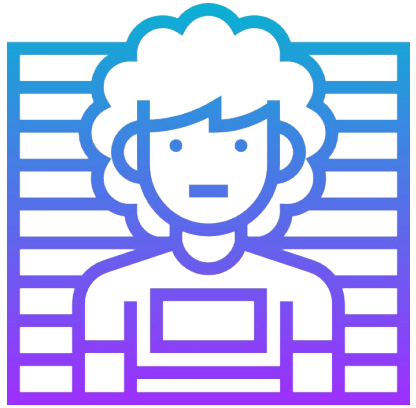
While there may be many hurdles that hinder the ability to make a dataset open (and/or FAIR) but established methods in health research are:

- ❑ **anonymisation** and/or
- ❑ **access control**
- ❑ making sure that there are also **consent forms** in use where necessary

Icons made by https://www.flaticon.com/authors/eucalyp

Open Data Methods

Anonymisation:
The information that could identify an individual is randomly altered or stripped out.



Direct identifiers: such as a person's name which need to be anonymised

Indirect identifiers: These may be used to infer the identity of an individual from multiple sources that provide partial information such as a person's age or occupation

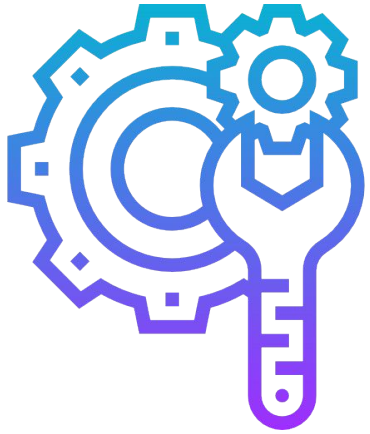
With indirect identifiers, extra care needs to be taken that these data are anonymised sufficiently and will likely require multiple methods including blurring, removal and redaction.

ORION open science

Icons made by https://www.flaticon.com/authors/eucalyp

Open Data Methods

Anonymisation:
The information that could identify an individual is randomly altered or stripped out.



There are many tools available for anonymisation, but one that has been developed through the OpenAIRE is Amnesia.

This tool is still in a beta version but provides a centralised method for anonymisation through a web interface by simply uploading your data files using *drag and drop* and it will then produce your desired anonymised data.

It is part of a suite of tools that are part of Open AIRE and the European Open Science Cloud (EOSC) which provide the tools for the workflow through the data lifecycle. Many commercial options exist too but anonymisation should also be possible through rigorous application of a workflow that applies established checks and balances.

ORION open science

OpenAIRE http://catalogue.openaire.eu/search

EOSC-hub https://www.eosc-hub.eu/services

Icons made by https://www.flaticon.com/authors/eucalyp

Open Data Methods

Access control is also a method that can be applied to controlling patient data.

We can typically think of three tiers (outlined by the UK Data Service) of access control and these can be employed to restrict who can use these data:

- Open
  - Data licensed for use with an ‘open licence’ are data which are not personal and have relatively few restrictions to use.
- Safeguarded
  - Data licensed for use in the 'safeguarded' category are not personal, but the data owner considers there to be a risk of disclosure resulting from linkage to other data, such as private databases.
- Controlled
  - Controlled data are data which may be identifiable and thus potentially disclosive. These data are only available to users who have been accredited and their data usage has been approved by the relevant Data Access Committee. Controlled data require registration/authentication.

ORION open science

Is Open Data Equal to Free Data?

This question can be interpreted in a number of ways are open data free ***to use***, are they free to *acquire* and is there a cost to their proper *management*?

Data acquired through open licences are generally free to use but the openness of the licence could be one of several, whether they are Creative Commons or using another brand of license.
Using Creative Commons as an example, CC-0 and CC-BY are clearly very open and free. However, going further through the spectrum of CC licences to CC-BY-SA you would be restricted to making sure that any derivatives produced from the original have the same licence, while those with NC cannot be used for commercial purposes.

	Can someone use it commercially?	Can someone create new versions of it?
Attribution	👍	👍
Share Alike	👍	Yup, AND they must license the new work under a Share Alike license.
No Derivatives	👍	👎
Non-Commercial	👎	Yup, AND the new work must be non-commercial, but it can be under any non-commercial license.
Non-Commercial Share Alike	👎	Yup, AND they must license the new work under a Non-Commercial Share Alike license.
Non-Commercial No Derivatives	👎	👎

http://tiny.cc/researchgatecc

Icons made by https://www.flaticon.com/authors/eucalyp

Is Open Data Equal to Free Data?

This question can be interpreted in a number of ways are open data free *to use*, are they free to ***acquire*** and is there a cost to their proper *management*?

On the other hand, when taking financial considerations into account, there can be times when you can acquire data for a fee that you must then be careful to read their terms and conditions carefully.

In many cases, this method of acquiring data will be licensed using bespoke licences which may at first glance look very similar to licences such as CC but will have significant variations which ties back into their freeness to use.

An example here is the data that can be acquired from the British Geological Survey (BGS). In some cases, data may only be acquired for a fee and the BGS has its own license which you would need to adhere to. Another example can be found at the UK's Meteorological Office, while many organisations in the UK have also adopted the Open Government Licence (OGL).

British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

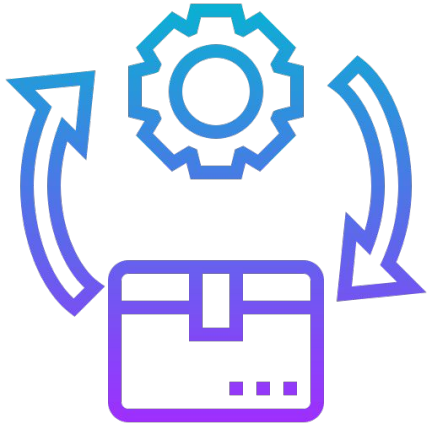
ORION open science

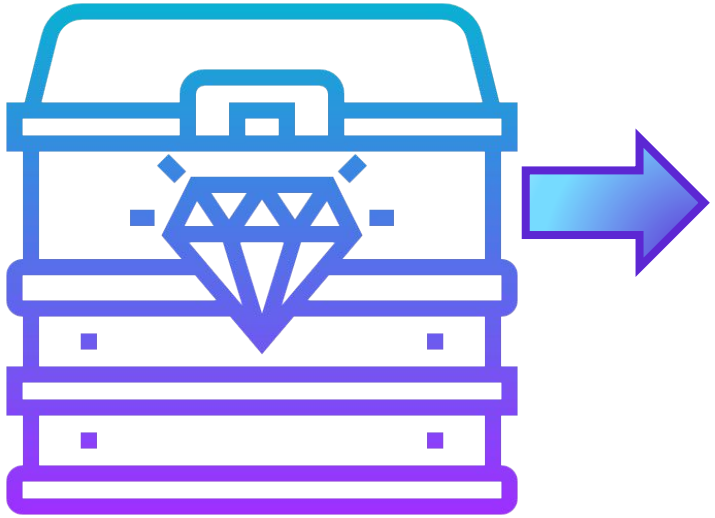
Icons made by https://www.flaticon.com/authors/eucalyp

Is Open Data Equal to Free Data?

This question can be interpreted in a number of ways are open data free *to use*, are they free to *acquire* and is there a cost to their proper ***management***?

When conducting proper research data management through the course of a project, data objects typically progress through a lifecycle from capture to analysis to eventual preservation. Each of the steps through this curation lifecycle will most likely entail a financial cost and in some cases this may be quite high.





Therefore, *open data do not necessarily mean that they are free data:*

- Costs associated with preservation in a repository and/or archiving of data for long term posterity may be very high depending on the size of the data and costs per MB.
- Unlikely that the end user will incur charges to access the data, the data depositor will typically absorb the burden of these costs.

ORION open science

Icons made by https://www.flaticon.com/authors/eucalyp

Is Open Data Equal to Free Data?

This question can be interpreted in a number of ways are open data free *to use*, are they free to *acquire* and is there a cost to their proper ***management***?

There are costing tools such as this one from OpenAIRE which can be used to estimate how much should be factored into a research grant proposal while we recommend the use of data management plans (DMPs) as a starting point to consider the costs in advance of a project starting.

Some projects may produce very little data, both in volume and size, while others will be the opposite so there can be no single answer to how much costs will be incurred. Indeed, other factors such as being able to use existing or free services will affect costs and will be entirely dependent in a case by case manner.