

FAIRSpectra

Metadata and Data Encoding, Promoting Open Research

Alex Henderson, University of Manchester, UK



A community-led project to develop metadata and file format standards for spectroscopy and spectrometry

Open Research

There is now a growing requirement for data acquired on projects supported by national funding programmes to be shared openly. This includes organisations such as UKRI, ERC, Horizon, NIH, NSF, Wellcome, DFG etc.

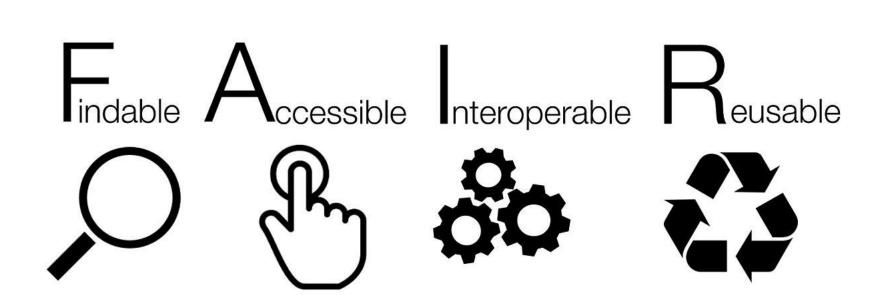
Sharing data requires that the files must be readable by others and can be understood in terms of their content.

What is FAIR?

The **FAIR Guiding Principles** describe things to consider when making scholarly data available: https://www.go-fair.org/. These are broken down into four interrelated topics: Data and metadata should be **Findable**, **Accessible**, **Interoperable**, and **Reusable**.

Findability and Accessibility relate to identifiers such as a DOI, and appropriate licencing terms. Interoperability and Reusability recommend that data are stored in a manner that is readable by all, and that information about the data (the metadata) is described in enough detail to **be understood by both humans and machines**.

FAIRSpectra aim to tackle issues with interoperability and reusability. This has been broken down into two areas of activity: open file formats and metadata vocabulary.



Open file formats

Open and de-facto file formats for SIMS include:

- VAMAS (ISO 14976:1998, ISO 22048:2004),
- JCAMP-DX,
- Grams SPC, and
- imzML.

However, few data analysis packages can read these formats and not all instrument vendors support them. Other open formats should be considered, for example **HDF5**, **Apache Arrow**/Parquet/Feather, and **Zarr** which have wide adoption, but limited formal semantics for our data.

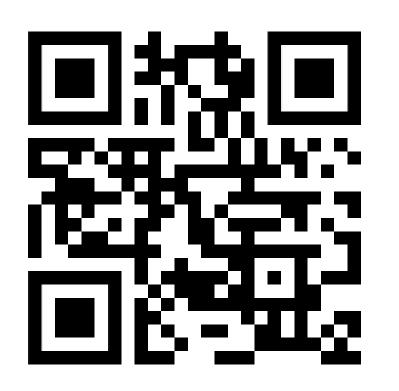
Metadata vocabulary

We will work to identify a common vocabulary for various spectrometric and spectroscopic domains. This vocabulary will consist of currently available terms for the parameters we need to describe and will propose new terms that are **agreed upon by the SIMS community** where necessary.

Other communities have brought semantics and modern file formats together. The climate community use NetCDF, (HDF5 with well-structured metadata layout). The microscopy community are looking to Zarr to underpin their Next Generation File Format (NGFF). Mass spectrometry use mzML, astronomy use FITS, and magnetic resonance imaging use DICOM. These are accepted standards in those communities and as a result many software packages are available to read, write, search, and archive those data.

Get involved!

This is a community project, and we need **your input** to ensure **your needs** are met. **No experience is necessary, only enthusiasm** for helping to develop a better future for our research.



Visit http://fairspectra.net