FORgreensoft
SOFT MATTER MORE GREEN

# THE FAIR DATA PRINCIPLE AND METADATA FOR EXPERIMENT AND SYNTHESIS

Benoit Loppinet, FORTH - IESL, April 2024

The FAIR Data Principles are:

**Findable:** Findable data should be easy for both humans and machines to find.

Findable data requires that:

F1. (Meta)data are assigned a globally unique and persistent identifier.

F2. Data are described with rich metadata (defined by R1 below).

F3. Metadata clearly and explicitly include the identifier of the data they describe.

F4. (Meta)data are registered or indexed in a searchable resource.

The best way to achieve Findable data is by:

- Depositing your dataset into a recognized data repository which assigns globally unique persistent identifiers (such as DOIs).
- Add as much contextual information (metadata) as possible when depositing your dataset into the repository.

**Accessible:** Accessible data refers to data that can be accessed once found; this may involve authentication of the user and authorization of access. Accessible data requires that:

A1. (Meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorization procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

The best way to achieve Accessible data is by:

- Depositing your dataset into a recognized data repository which uses standard communications protocols like HTTP.
- Ensuring that the data repository you choose gives continued access to metadata even when datasets are removed. Peer-reviewed Manuscript Accepted for Publication) arising from this submission, in accordance with the grant’s open access conditions.

**Interoperable:** Interoperable data refers to data that can be compared and combined with data from different sources, by both humans and machines. Interoperable data requires that:

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

The best way to achieve Interoperable data is by:

- Checking FAIRsharing.org for the standards that apply to your data type and using them.
- Ensuring that the data repository you choose allows you to include links or references to other related data.
- Using open, non-proprietary file formats for your data.

**Reusable:** Sharing data which can be reused by others is the main goal of the FAIR Principles.

Reusable data requires that:

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

The best way to achieve Reusable data is by:

- Adding as much contextual information (metadata) as possible when depositing your dataset into a repository.
- Applying an open license to your data, preferably CC0 or CC-BY 4.0.
- Checking FAIRsharing.org for the standards that apply to your data type and using them. For all subscription articles for which APS holds copyright, authors have the right to post and update the article on free-access e-print servers (repositories, including institutional and subject specific such as the ArXiv) as long as the version of record prepared and/or formatted by APS or its vendors are not used for that purpose, and the APS copyright is clear. All article depositions should include a link/DOI to the published manuscript.

From https://open-research-europe.ec.europa.eu/for-authors/data-guidelines

A general definition of metadata reads as: Metadata refers to the information needed to discover, use, and understand data. It provides context and details about the data, answering questions such as who, what, when, where, why, and how. In the context of research, metadata ensures that users understand the content and limitations of the data. Experimental Metadata describes experimental conditions, the experimental protocol, and the equipment used to generate the data.

In view of the FAIR data management policy, we need to provide the relevant metadata in the clearest possible form.

There is so far no accepted "universal" standards in the field of material sciences. Below is an example of what *Angewandte Chemie* requires for published data. This should be relevant and provide good guide lines for synthesis and experiments activities within FORGreenSoft.

**Instruction for synthesis and characterization data reporting / metadata (from Angewandte Chemie Notice to Authors)** MAY NEED TO BE EDITED

Angewandte Chemie International Edition - Notice to Authors - Wiley Online Library

A. Synthetic Procedures

"Literature references to known but non-commercial compounds should be given, and hazardous chemicals, equipment and techniques must be emphasized. If practical, authors should use a systematic name (IUPAC or Chemical Abstracts) for each title compound in the Experimental Section. "

"New synthetic procedures should contain reactant quantities in weight or volume and molar units. Equipment details, such as reaction vessel, type of heating (conventional, microwave, or photoirradiation), irradiation wavelength, optical irradiance, cut-off filters, and details of purification techniques and solvents, among others, should be included. Yields of purified products should be in weight and percentage (e.g., 109 mg, 95%). Physical data should be quoted with decimal points and negative exponents (e.g., 25.8 $JK^{-1}mol^{-1}$). If chromatographic methods are used for purification, the nature of the stationary phase and thin-layer chromatography data/retention factor (e.g., Rf=0.38 ($CHCl_3$/MeOH 9:1)) should be provided. Please note that reactions following a new synthetic procedure can be conducted on a small scale, but at least one example should be at a scale of 1mmol."

B Characterization Techniques

*Please note: manipulation of spectra to misrepresent data is unethical and will not be tolerated.*

**Nuclear magnetic resonance spectroscopy:** Proton and carbon resonances must be provided for each new compound in the Experimental Section; solvent and instrument frequencies are required. Depending on the compound, other resonances, such as $^{19}$F, $^{29}$Si, or

$^{31}P$, should be added. NMR spectra should have sufficiently high signal-to-noise ratios so that all peaks can be adequately resolved.

$^{1}H$ NMR resonances should be reported to the nearest 0.01ppm. Multiplet abbreviations, number of atoms represented by each signal, and coupling constants should be provided; e.g., $^{1}H$ NMR (200 MHz, [$D_8$]THF, 25°C, TMS, ppm): δ=7.64–7.48 (m, 6H; Ar–H), 1.33 (q, *J*=8 Hz, 2H; $CH_2$). Mutually coupled protons in $^{1}H$ NMR spectra must be quoted with precisely matching *J* values to assist thorough interpretation. In instances where computer print-outs provide ambiguous readings, mean *J* values should be quoted that are rounded to the nearest decimal point.

$^{13}C$ NMR resonances should be reported to the nearest 0.1ppm. Provide resonances with high precision only in case of closely spaced signals. The number of attached hydrogen atoms can also be included; e.g., $^{13}C$ NMR (75 MHz, $CDCl_3$, 25°C, TMS, ppm): δ=72.5 (CCH), 26.8 ($CH_3$).

Copies of the $^{1}H$ and $^{13}C$ NMR spectra of all key intermediates and all final products must be included in the Supporting Information. The spectra must be those resulting from the specific reactions reported in the manuscript and not copies from previous syntheses of the same compounds. Each spectrum must be legible and should be labeled with the compound number used in the manuscript and an image of the structure. The minimum chemical shift range for $^{1}H$ NMR spectra should be −1 to 10ppm and for $^{13}C$ NMR spectra −10 to 200ppm. All peaks should be labeled and integrated.

**Infrared spectroscopy**: It is not necessary to provide a full list of IR stretches. Only those signals that are diagnostic of the compound's functional groups need to be listed. The band frequencies should be given to the nearest 1 $cm^{-1}$ and their intensity should be provided (very strong (vs), strong (s), medium (m), weak (w), broad (br)); e.g., IR (KBr, $cm^{-1}$): $\tilde{v}$=1780 (vs), 1790 (s) (C=O).

**Mass spectrometry (MS):** It is not necessary to provide a full list of MS peaks. Only those signals for which the molecular fragment can be identified need to be listed. Molecular ion peaks, and any other fragmentation peaks, should be reported in comparison to the calculated mass for the ion. Please use an ionization technique suitable for your compound. Relative intensities of the signals should be provided; e.g., MS (EI, 70 eV) m/z (%): 173 (32), 171 (100) $[M+H]^+$.

**Elemental analysis or high-resolution MS (HRMS):** Evidence for elemental constitution should be provided by satisfactory elemental analysis, in which case duplicate analysis should be obtained and an average presented (both sets of data used to calculate the average should be provided). Elemental analysis; e.g., Elemental analysis calcd for $C_{20}H_{32}N_2O_5$: C 63.14, H 8.48, N 7.36, found: C 62.88, H 8.41, N 7.44. High-resolution mass spectrum; e.g., HRMS (ESI) *m/z* calcd for $C_{32}H_{47}NO_5+Na^+$: 548.3352 $[M+Na]^+$; found: 548.3331. For compounds where elemental analysis data is not provided, the HRMS data should be accompanied by NMR spectra with sufficiently high signal-to-noise ratios. Note that elemental analysis data must be provided for papers detailing the isolation and structure elucidation of natural products.

**UV/Vis and fluorescence spectroscopy:** If UV/Vis or fluorescence spectroscopy data are discussed in the manuscript, the following data should be provided in the Experimental Section of the Supporting Information, along with any copies of the corresponding spectra. UV/Vis absorptions (peak, extinction coefficient); e.g., UV/Vis (n-hexane, nm (mol$^{-1}$dm$^{3}$cm$^{-1}$)): $\lambda_{max}$ ($\varepsilon$)=320 (5000), 270 (12000). Fluorescence excitation and emission; e.g., fluorescence (CH$_2$Cl$_2$, nm): $\lambda_{ex}$=435.5nm, $\lambda_{em}$=659, 726nm.

**X-ray absorption spectroscopy (XAS):** Collection and treatment of XAS data require specialized knowledge. Please consult a beamline scientist or other experienced person to ensure the appropriate steps are taken to collect, process, and interpret XAS data correctly. The Athena User's Guide provides general guidance for XAS data processing.

**Specific rotation:** Specific rotations following [ɑ]=(100·ɑ)/(*lc*) should be provided for isolated and synthesized natural products as well as for other enantioenriched compounds; e.g., [*α*]D20=−13.5 (c=0.2 in acetone). For new compounds that are chiral, HPLC or GC traces should be included in the Supporting Information.

**Physical state and melting point:** A description of the physical state and color of a new compound should be given; e.g., yellow needle-like crystals. Melting point ranges should only be provided for crystalline compounds; e.g., mp: 90.2–91.2°C.

**Isomeric mixtures:** Where isomeric mixtures are reported, such as diastereomeric or enantioenriched mixtures, please provide percentage compositions and information about how these values were obtained (e.g., NMR spectroscopy, HPLC, etc.). If certain spectroscopic signals (e.g., NMR signals) can be attributed to either of the isomers, these data should be reported in separate lists and not in combined lists.

**Microscopy** images should be captured at an appropriate magnification to show a representative sample. When high-magnification images of selected particles are used they must be supplemented by low-magnification images of the broader sample, and the use of histograms and statistics to describe size and shape distributions is encouraged.

**Powder X-ray diffraction** data: XRD (Cu Kα1, 0.15406 nm, 2θ (°) (d (nm)): 5.2 (1.68), 10.1 (0.83).