TONY DAVIES COLUMN

Are your spectroscopic data FAIR?

Leah McEwen,[a] David Martinsen,[b] Robert Lancashire,[c] Peter Lampen[d] and Antony N. Davies[e,f]
[a]Physical Sciences Library, Cornell University, Ithaca NY 14853, USA. ORCID: https://orcid.org/0000-0003-2968-1674
[b]David Martinsen Consulting, Rockville, MD 20850, USA. ORCID: https://orcid.org/0000-0002-8667-5855
[c]The Department of Chemistry, The University of the West Indies, Mona, Kgn 7, Jamaica.
ORCID: https://orcid.org/0000-0002-6780-3903
[d]Leibniz-Institut für Analytische Wissenschaften – ISAS, Dortmund, Germany.
ORCID: https://orcid.org/0000-0003-3004-1463
[e]Expert Capability Group – Measurement and Analytical Science, Akzo Nobel Chemicals b.V., Deventer, the Netherlands
[f]SERC, Sustainable Environment Research Centre, Faculty of Computing, Engineering and Science,
University of South Wales, UK. ORCID: https://orcid.org/0000-0002-3119-4202

Let us start with a definition: FAIR stands for Findable, Accessible, Interoperable, Reusable. Surely a clear target for anyone responsible for operating academic or industrial laboratories. Let me (AND) follow up with a confession… I had previously just thought of FAIR as another offering originating from the bioinformaticians around Open Science publishing. However, I learnt very quickly that what started as a movement to improve intelligent access to Open Science and supporting data contains all the tools and methods of working to have the potential to be extremely important in all our daily work. It is equally applicable as a time-saving strategy for confidential information located and retrieved inside a company. It is perhaps worthwhile to note that FAIR does not necessarily imply free. This column has mentioned FAIR once before, in relation to Henry Rzepa's NMR data repository,[1] but had not really gone into any depth.

Why explore this topic now? Well Leah McEwen, Chemistry Librarian in the Clark Physical Sciences Library at Cornell University in the USA, with assistance from David Martinsen (30 years' experience with the American Chemical Society publishing arm) organised and ran a very successful workshop under the auspices of the International Union of Pure and Applied Chemistry (IUPAC) and the Committee on Data of the International Council for Science (CODATA). "Supporting FAIR Exchange of Chemical Data through Standards Development" was held on 16–17 July 2018, hosted by the University of Amsterdam.[2] The workshop was co-sponsored by the IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS), their Subcommittee on Cheminformatics Data Standards (SCDS), and CODATA, and was attended by some very influential people. Richard Hartshorn, the current Secretary General of IUPAC, flew in from New Zealand and had some strong words to say about the essential need to understand how our next generation of scientists will expect us to have kept up with ensuring the provision of well-curated, reliable scientific data available at a single click.

If we can regularly find out what the President of the USA is thinking in his bathroom at 7:30 am in the morning, why do I get four pages of text hits when searching for a simple fact like the name of element 113? I spoke to my mobile phone and it told me immediately the correct answer and the reason behind the naming—but sourced from Wikipedia, not IUPAC. In an age of deliberate falsehoods and alternative truths being published and widely distributed in the service of some ideology or other, it is ever more important that reputable international bodies keep abreast of the current technological advances for information distribution. Having systems which exhibit the FAIR principles promises to make it much simpler to locate peer-reviewed, real scientific data in a form that we (and our IT support systems) need.

For some thoughts from Leah on libraries in transition, see an interview recorded at the Beilstein Open Science Symposium (22–24 May 2017).[3]

Can you call yourself FAIR?

In March 2016, Mark D. Wilkinson and a host of co-authors brought together current thinking on how we should all make sure that we improve accessibility to the data we generate.[4] An underlying assumption of FAIR is that it applies equally to human and machine interaction with scientific data. So, there is enormous emphasis on standardisation of metadata so that machines (such as my mobile phone) have far more information available to support access without the need for human interaction (like sifting through four pages of hits in text format). This approach is quite ground-breaking, as previous initiatives have almost singly focussed on improving retrieval systems for direct human consumption. Clear "false positives" are often ignored by us almost without thinking, but computer systems find that much more difficult and so need to be "fed" with much better accompanying information to provide appropriate

context. What I like about the approach in this seminal work is the fact that it is easily understandable, which is also unusual for such documents. It adds some nice detail to what is meant under the terms Findable, Accessible, Interoperable and Reusable (see below). Barend Mons and co-workers have also recently published a useful, easy-to-read paper setting FAIR in context.[5]

FAIR principles

The seminal publication[4] proposed technical definitions of what the terms making up FAIR mean for scientific data in repositories.

Findable:

F1. (Meta)data are assigned a globally unique and persistent identifier
F2. Data are described with rich metadata
F3. Metadata clearly and explicitly include the identifier of the data they describe
F4. (Meta)data are registered or indexed in a searchable resource

Accessible:

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  A1.1. The protocol is open, free and universally implementable
  A1.2. The protocol allows for an authentication and authorisation procedure, where necessary
A2. Metadata are accessible, even when the data are no longer available

Interoperable:

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

Reusable:

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

So, scanning through the FAIR principles it becomes clearer why the international scientific unions are becoming involved. They "own" official "domain-relevant community standards" and already have the processes in place to deliver updates etc.

After the introductory discussions, the workshop was split into two parallel streams. One stream dealt with the **GO FAIR Implementation Network for Chemistry** which is currently being created.[6,7] The network will include building a repository of FAIR resources useful to chemists. This workshop stream was tasked with the following:

Address the following themes in supporting FAIR data:

- Use cases and interoperability needs for chemical data and information across the enterprise and related disciplines
- Development of tools for researchers and other expert users to support application and use of standards for chemical data
- Mechanisms for validation and curation of standard representation of chemical data

The second stream was much closer to spectroscopists' hearts and ran under the title of **Interoperability Criteria for Spectroscopic Data Exchange**. The information distributed prior to the workshops explained the relevance of the IUPAC JCAMP-DX suite of recommendations in this context as:

"The IUPAC JCAMP-DX data standard has become a critical piece of this FAIR data exchange for spectroscopic data. It satisfies a number of critical criteria in that JCAMP-DX file export is available in nearly all software packages for spectroscopic instruments, it is ASCII not Binary, it is non-proprietary and there has been a large amount of data already generated."

Since the much-documented merger of the IUPAC XML in Chemistry initiative with the ASTM AnIML standardisation effort, no maintenance work on the IUPAC standards has taken place in the hope that the AnIML initiative would take up this challenge. However, work carried out by Greg Banik (Bio-Rad) surveying the use of JCAMP-DX and others has shown there is a clear and urgent need to make a decision on the future of these standards.

Again, the briefing notes clearly set the scene for some decision making…

"IUPAC is reviewing the current status of the JCAMP-DX format, including the extent to which it is being used, what enhancements users would like to see, and the extent to which the files that are generated in 'JCAMP-DX format' adhere to the JCAMP-DX standards. Another key step towards FAIR spectroscopic data is the development of standard criteria for publishing spectroscopic data that will optimize data use, reuse, and interoperability across domain repositories."

Interoperability criteria for spectroscopic data axchange stream

So, with a clear remit to decide on the future of the IUPAC JCAMP-DX standards, the workstream group sat down to review the current position and make clear proposals on the requirements going forward. The first half of the workshop was set the following tasks:

JCAMP-DX review and future requirements

- Benefits of JCAMP/agnostic data exchange
- Deficiencies of JCAMP format
- Requirements for evolving JCAMP (extensions, XML etc.)
- Community engagement
- Validation requirements

There was some very plain talking amongst the participants—contrasting the original requirements which had established the JCAMP-DX series of standards with those required for a fully FAIR compliant system. In the original standards, the aim had been to facilitate the creation of reference spectroscopic databases by providing a common

format that all instrument vendors could sign up to which had the minimum amount of metadata required to correctly identify and interpret (plot) the measured data accurately. Additional comments and structured metadata were allowed, including the introduction of "private" labels for information which was not internationally standardised but crucial for internal uses of the format within specific communities. Peter Lampen highlighted the fact that the standard also included complicated and potentially loss-less data compression schemes, which were critical to meeting the historical data file size challenges. However, these had caused enormous headaches for programmers not familiar with their unique concepts. These schemes are now less relevant with enormous improvements in network speeds and the availability of huge data storage capacity.

A brainstorm amongst the participants highlighted the following points which needed addressing:

- Clarity on IUPACs position and funding—is IUPAC committed to support maintenance of the standard?
- Standards that can be used to set up repositories
- JCAMP-DX—Yes or No?
- JCAMP-DX—Minimal vs Comprehensive. As discussed above, with a push toward data publication and with a greater demand for more detailed metadata, a more comprehensive approach might be needed.
- Practical implementation for Open Science
- Data + Publications—what are the requirements for primary research data that supports journal publications?
- Original + Processed—Should only the processed human-readable spectrum need to be defined, or does the original data also need to be included?
- Community direction, and appearance of somewhat fragmented communities—IUPAC/JCAMP, Allotrope, NMRedata, IRUG and others are creating somewhat independent solutions.
- Where is the one button? One click to go from lab to publication, with appropriate standards, identifiers etc.
- Granularity—should a data package contain multiple spectra, or multiple substances? Or should it be restricted to data for a specific molecule? Or should it be single spectra for a single molecule? What should be registered as a DOI?
- There is an urgent to update NMR, NIR, Raman
- MS is less urgent, since most vendors support netCDF rather than JCAMP-DX

Based on this assessment the focus for immediate action shifted to the NMR community needs. There had been significant developments in this field since the first IUPAC NMR standard recommendations were published. Also, at the time of the merger of efforts with the AnIML group, a multi-dimensional NMR standard, JCAMP-DX 6.0, had been almost ready for publication. It seems the vendors have adopted this in its draft form across the board, but it still needs some work to cover the majority of use cases in NMR.

Urgent improvements, including IUPAC JCAMP-DX for NMR

Having decided the highest priority moving forward was NMR, discussions revolved around a number of technical and strategic considerations which would need to be addressed before the standard could be put forward for publication as an IUPAC Recommendation.

- XML vs traditional JCAMP (incremental extension of current format would be least disruptive; conversion to XML would be very disruptive)
- New metadata requirements (focus on JCAMP-DX as a canonical data model rather than a specific format)
- Metadata for FAIR implementation
- Newer experimental techniques, e.g. *n*-dimensional NMR, discontinuous data (some features may not be easily implemented using the current data model, for example, *n*-dimensional NMR may be simpler to implement in XML)

TONY DAVIES COLUMN

- Synchronisation of JCAMP across experimental techniques
- Private label terms that need/can be made standard (where appropriate, re-use tags, definitions, look at descriptions from other communities, e.g. Allotrope, NMRedata etc.)
- An application programming interface (API) is needed to assist implementation
- Raw (FID) + spectrum: the NMR community prefers the data to be stored in the FID form and reprocessed to spectra upon opening, but this is a problem for saearchable reference databases and publications where the figures are all of spectra rarher than FIDs.

It looks as if quite a lot of urgent work is required to catch up on so many years when the hopes were that AnIML would deliver the necessary steps forward. Therefore, several projects were defined to get this moving as quickly as possible.

Project Group 1: focus on JCAMP-DX extensions for NMR data

- The first phase will be to quickly survey the major NMR vendors to fully document their issues with the current JCAMP-DX, and their interest in supporting the effort to update JCAMP-DX.
- They are also tasked with assessing the level of effort involved in updating JCAMP-DX for NMR to include the 2D NMR specifications from the draft version 6.0 and the recommendations for standardising the private labels.
- A project proposal will be developed for the second phase, focusing on delivering the new NMR recommendation for JCAMP-DX implementing the specifications captured in the first phase.

Project Group 2: focus on metadata for data publication and the items that could be considered important to FAIRify the data

These include such things as:

- ORCID
- Organisation ID
- InCHI
- DOI of the data
- DOI of the associated article, if there is one
- Association of structures to spectral features, as NMRedata
- Funding information
- Instrument ID
- Owner
- License information

Project Group 3: focus on tools and workflows

- Develop a validator building on the experiences gained running the old JCAMPCHECK and DXCHECK programs. Validation should be carried out at different levels:
  - Validator level 0: check format—does it correspond to the standard?
  - Validator level 1: is the minimum required data present?
  - Validator level 2: is the content reasonable science?
- Visualisation
- Export from lab (instrument or ELN) to repository or to publisher

Project Group 3 will also need to consult on and provide recommendation on whether IUPAC develop these tools or does IUPAC give seal of approval to third party tools?

Summary

So, we have a green light to proceed after years of stalled developments… the big challenge for us all is to deliver into this rapidly changing environment!

References

1. A.N. Davies, D. Martinsen, H.S. Rzepa, C. Romain, A. Barba, F. Seoane, S. Dominguez and C. Cobas, "Simplifying spectroscopic supplementary data collection", *Spectrosc. Europe* **29(4),** 6–8 (2017). http://bit.ly/2v6JeVy
2. https://iupac.org/event/supporting-fair-exchange-chemical-data-standards-development/
3. Beilstein TV, *Libraries in Transformation*. http://www.beilstein.tv/video/libraries-in-transformation/
4. M.D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship", *Sci. Data* **3,** 160018 (2016). doi: https://doi.org/10.1038/sdata.2016.18
5. B. Mons, C. Neylon, J. Velterop, M. Dumontier, Michelf, L.O.B. da Silva Santos and M.D. Wilkinson, "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud", *Inform. Serv. Use* **37,** 49–56 (2017). doi: https://doi.org/10.3233/ISU-170824
6. Go FAIR, *Implementation Networks*. https://www.go-fair.org/implementation-networks/
7. Go FAIR, *FAIR Principles*. https://www.go-fair.org/fair-principles/