TONY DAVIES COLUMN

FAIR practice

Antony N. Davies
SERC, Sustainable Environment Research Centre, Faculty of Computing, Engineering and Science, University of South Wales, UK

Following our articles on the FAIR initiative, our esteemed editor Ian Michael was very keen that we looked at some examples of the FAIRification of data handling, collection and archiving.

Fortunately, this request coincided with a webinar featuring some subject leaders in this field called "FAIR Data: From principles to best practices" and hosted by MyScienceWork on 30 March 2021.[1,2]

FAIR data: from principles to best practices

MyScienceWork describes themselves as having a mission of democratising science. Sally Ekanayaka who hosted the workshop stated that…

"MyScienceWork is a technology company that provides a suite of advanced data-driven solutions for research institutions, scientific publishers and private-sector R&D companies. Fostering digital discoverability and collaboration, securing long term accessibility and reusability of research outputs, and maximising research impact is at the heart of the company's repository solutions…"

and as such they clearly have a major interest in helping drive forward initiative in the FAIR field.

The planned key takeaways from the workshop were:

- An overview of FAIR data publishing standards
- Ways of operating the FAIR principles within research activities
- The core requirements for trustworthy digital repositories

So, it looked like a good bet to start gathering information on people and projects that have carried out FAIRification work.

FAIR for (data) publishers

Erik Schultes, who has been working full time for the last three years on FAIR implementation strategies at the GO FAIR Foundation and whose background is in the data intensive biological sciences, led off. His presentation focused on FAIR publishing and asked the question "what is your FAIR Implementation Profile". Erik's talk delivered on the first two "Takeaway" bullet points above. First, he emphasised that the GO FAIR Foundation is working with lots of organisations in different fields or domains and, as such, is completely "standards" agnostic. They deliberately leave the choice of which standards to adopt to the individual domain specialists. GO FAIR's support in their different interactions is to help these organisations reach the highest level of FAIR possible, whilst keeping an eye on what other initiatives are delivering in order to help drive convergence.

In January, Erik and Jan Velterop published a call-to-arms for a network of academic publishers to come together and make some joint decisions on FAIRification in the academic publishers' sector through a GO FAIR Implementation Network structure.[3] He reminded the listeners about the GO FAIR guide on "how to Go FAIR" which breaks down the FAIRification process into essentially three stages which starts off with decisions by communities of practice on domain relevant community standards. GO FAIR support this process by initiating Metadata 4 Machines activities to bring together domain experts with FAIR Metadata Experts to produce reusable, domain-specific FAIR metadata schema (Figure 1). Particular emphasis is to try, whenever possible, to use already agreed standards and avoid re-inventing the wheel.

From the perspective of available tools, Erik pointed out the CEDAR Workbench, an Open-Source tool for generating metadata that describe scientific experiments. Just be warned though, I had never heard of this tool and on Googling it I was presented with some very fine solid wood tables, but better links to a vast amount of information on the CEDAR tools can be found in Reference 4! CEDAR is the acronym for the *Center for Expanded Data Annotation and Retrieval*, a collaboration lead by Mark Musen with Stanford Co-PIs from Stanford, Oxford, Yale and Northrop Grumman. Their mission is to develop information technologies that make authoring complete metadata sets much more manageable, and that facilitate using the metadata in further research. The Open-Source CEDAR Workbench was specifically built to help deal with the extremely poor state of metadata generation, a critical activity in the FAIRification process as we have documented. It consists of a set of Web-based tools covering acquisition, storage, search and reuse of metadata templates including the simple construction of metadata acquisition forms. The metadata generated using CEDAR templates are FAIR compliant and interoperable with Linked Open Data and retrievable in JSON, JSON-LD and RDF formats. A nice short video explaining the nature of CEDAR is

DOI: 10.1255/sew.2021.a14

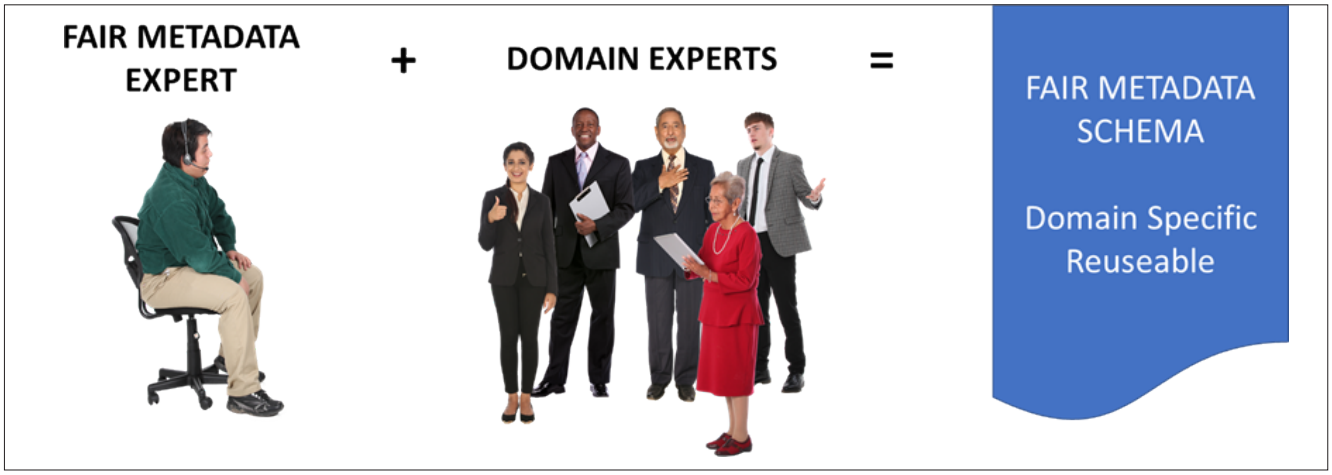

Figure 1. GO FAIR Metadata 4 Machines workshops are designed to bring domain experts and FAIR metadata experts together to generate domain-specific FAIR metadata schema.

available at https://more.metadatacenter.org/video/introductory-video-cedar-all-basics.

The creation of FAIR metadata is of course only part, even if a very important part, of the FAIRification process. The GO DATA FAIR Implementation Profile (FIP) concept which is the next step after the creation of the domain-specific metadata schema. At the time of writing there have been 25 FIPs created out of the FIP workshops and an article by Barabara Magangna from the Vienna Umweltbundesamt GmbH in Austria and co-workers from Prague and the Netherlands about the use of FIPs as accelerators for FAIR convergence is available currently on the OSF Preprints.[5]

Access to digital content—repositories

So, with the principles sorted the workshop moved on to the role of repositories and the core requirements of software to be capable of fulfilling this challenging role. Yann Mahe from the meeting hosts MyScienceWork took up the challenge of presenting concepts for data science solutions for research which can cope with the complexity and multidisciplinary nature of the content whilst satisfying the needs of diverse stakeholders. We have all observed the explosion of scientific data sources in the last 20 years, most of which has been outside of the classical peer-reviewed publishing model. In fact, it is probably fair to say that currently the majority of scientific data is "published" outside of the classical peer review quality control mechanism of the past. The COVID-19 pandemic has highlighted a real modern-day challenge facing data consumers in identifying and filtering out information that is either false or contextually extremely misleading in its presentation. Yann addressed the challenges to link, store and subsequently visualise this diverse "ocean" of data to which we all have access in a meaningful way.

The concept of Openness and Open Access in science gives huge potential opportunities for the scientific community. However, there can be gaps between the expectations of scientists regarding "Big Data" and the available tools for use and reuse. Such issues can inhibit obtaining the full benefit of the fundamental mould-breaking approach to science publication. Yann emphasised that the implementation of the FAIR principles is one solution to brining the expectations of researchers closer to fulfilment, but highlighted the different levels of progress between the early innovator countries and the rest of the world. This variation in support is not only reflected across the globe but also between individual scientific domains and communities who can be seen to have quite different concrete goals set by their stakeholders under the FAIR banner.

The essential role of repositories in collation, curation and presentation of data in a machine and human readable form was discussed. The fact that, in real-world operation, specific tooling meets multiple requirements around Findability, Accessibility, Interoperability and Reusability—a concept that the IUPAC FAIRSpec project is also confronting as it continues its domain-specific work.[6] Yann reviewed the concept of a repository especially around the role of repositories in lowering the barriers to the adoption of FAIR data principles whilst also generating additional connectivity to other related content both within the repository itself and between different repositories. This will greatly help mitigate the challenges around addressing different stakeholder demands.

For commercial entities, a well thought out repository is also essential in meeting underlying compliance and business critical drivers, such as around managing rights to finding and accessing specific content. There is also a need to deploy data retention policies, where data may be required to be deleted after a specific lifetime but also held should a legal hold be in place on the information content. Keeping systems future-safe also requires a well thought out end-of-life plan for any infrastructure and a well-managed FAIR repository is also far simpler to migrate to the next generation of software and hardware. Longevity is also ensured through implementing well-documented, controlled vocabularies within the repositories, which is where we come back to the work described by Erik above!

TONY DAVIES COLUMN

Deployment examples—from the Pistoia Alliance work in life science

The third workshop presentation was by Ian Harrow, Project Manager with the Pistoia Alliance based on the importance of the FAIR data initiative in the life sciences. Ian provided insights they have gained by building a toolkit for FAIR. Ian started off by reviewing the enormous growth in available data from the individual measurement through the move towards high-throughput methodologies, "omics" and now Big Data and the internet of things where data mining becomes essential and the FAIR guiding principles and data tools a key enabler. These tools and services also serve as a value multiplier for the available data, something worth remembering when you are trying to obtain funding your own projects! The FAIR principles are not a standard, but Ian pointed to the Research Data Alliance (RDA) FAIR Data Maturity Model published last year as a major step forward in "standardising" approaches to FAIRness.[7] There is a wealth of information on the RAD website at https://www.rd-alliance.org and if you are interested in learning more or contributing to their work there is the opportunity to join this organisation as a member or an organisation.

What is nice about the RDA approach is that it has looked at all the FAIR principles, which in their own right look quite daunting when starting out on your own FAIRification project and have assigned each one of three levels of importance.

The three levels of importance are defined as:

1) **Essential:** such an indicator addresses an aspect that is of the utmost importance to achieve FAIRness under most circumstances, or, conversely, FAIRness would be practically impossible to achieve if the indicator were not satisfied.
2) **Important:** such an indicator addresses an aspect that might not be of the utmost importance under specific circumstances, but its satisfaction, if at all possible, would substantially increase FAIRness.
3) **Useful:** such an indicator addresses an aspect that is nice-to-have but is not necessarily indispensable.

I see this as an excellent supporting document to help organisations develop and prioritise what Erik discussed under the GO FAIR FAIR Implementation Plan. Ian pointed out that the Pistoia Alliance had held three workshops in 2018 and 2019 in Europe and the USA which had clearly identified one of the major hurdles to FAIRification being "*just where do you start*"! It is clear that projects need to discuss and agree what are the underlying business goals and unambiguously document what does "FAIR Enough" mean in their worlds.

The Pistoia FAIR Toolkit arose to meet these needs recognising this needs to be a bottom-up approach and conceptualised their work around documenting use cases, identifying good tools, recognising the importance of training and the requirement across the whole work of good change management activities.[8,9]

With the space we have available it's impossible to go into the details of each individual project from birth to delivery that the alliance has run but if you follow the link in Reference 10 you will find five really interesting examples from Roche, Bayer, AstraZeneca, The Hyve and SciBite to go through.

Oddly enough, an old colleague Rolf Grigat who is now working at Bayer as a FAIR and Linked Data Enabler, had recently pointed me to the Bayer COLID implementation which is freely available on GitHub (https://github.com/Bayer-Group). He is particularly proud of their logo (Figure 2) and with any luck we will be able to feature in more detail the development journey that Bayer and the Pistoia Alliance took in a future column.

Figure 2

Conclusions

The timing workshop could not have been better. Covering concepts for FAIR deployments through systems to real-world examples of completed projects it really took the FAIRification idea forward from theory to real-world deployments. If you are still finding lockdown stressful why not take an hour out to listen to the whole workshop? You can use the link in Reference 1.

Everyone please, stay safe!

References

1. *FAIR Data: From Principles to Best Practices*, Workshop 31 March 2021. Hosted by MyScienceWork, https://youtu.be/FP7H_VmMl_U
2. S. Ekanayaka, I. Harrow, Y. Mahé and E. Schultes, *FAIR Data: From Principles to Best Practices* (webinar). MyScienceWork (2021). https://youtu.be/FP7H_VmMl_U
3. J. Velterop and E. Schultes, "An Academic Publishers' GO FAIR Implementation Network (APIN)", *Information Services & Use* **40(1–2),** 1–9 (2021). https://doi.org/10.3233/ISU-200102
4. https://more.metadatacenter.org/
5. B. Magagna, E.A. Schultes, R. Pergl, K.M. Hettne, T. Kuhn and M. Suchánek, "Reusable FAIR Implementation Profiles as accelerators of FAIR convergence", *OSF Preprints* (2020). https://doi.org/10.31219/osf.io/2p85g
6. A.N. Davies, R.M. Hanson, D. Jeannerat, M. Archibald, I. Bruno, S. Chalk, J. Lang, H.S. Rzepa and R.J. Lancashire, "FAIR enough?", *Spectrosc. Europe* **33(2),** 21–23 (2021). https://doi.org/10.1255/sew.2021.a9
7. RDA FAIR Data Maturity Model Working Group, *FAIR Data Maturity Model: Specification and Guidelines*. Research Data Alliance (2020). https://doi.org/10.15497/RDA00050
8. J. Wise *et al.*, "Implementation and relevance of FAIR data principles in biopharmaceutical R&D", *Drug Discovery Today* **24(4),** 933–938 (2019). https://doi.org/10.1016/j.drudis.2019.01.008
9. https://fairtoolkit.pistoiaalliance.org
10. https://www.pistoiaalliance.org/projects/current-projects/fair-implementation/

TONY DAVIES COLUMN

Tony Davies is a long-standing *Spectroscopy Europe* column editor and recognised thought leader on standardisation and regulatory compliance with a foot in both industrial and academic camps. He spent most of his working life in Germany and the Netherlands, most recently as Lead Scientist, Strategic Research Group – Measurement and Analytical Science at AkzoNobel/Nouryon Chemicals BV in the Netherlands. A strong advocate of the correct use of Open Innovation.

https://orcid.org/0000-0002-3119-4202

antony.n.davies@gmail.com