TONY DAVIES COLUMN

Finding data in today's information age: the Bayer COLID system

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

It is said we live in an information age. Some say we live in a mis-information age. We certainly enjoy easier access to a multitude of information sources to support our work than at any time in history. However, at what stage does the shear enormity of the information we are presented with mean that our ability to filter good from bad or even deliberately misleading is overwhelmed? Then we start making worse decisions than in earlier times when limited peer-reviewed information feeds were all we could rely on.

There are, of course, many Internet search engines available to you, and you will certainly have a favourite "go to" engine on which you rely for the majority of your information quests. However, companies can influence the results you are presented with, by paying the search engine providers or employing specialists whose job it is to understand how the search engines rank the information they have scraped from the Internet. Stand in front of a mirror and look yourself squarely in the eye and admit how many pages of search results you are prepared to scroll down through, and how often do you access multiple results pages to make an informed decision on the problem you are researching.

Finally, let's add a complication in that there may well be some superb work carried on within your own company which should certainly influence your decision making, but it was carried out in a division whose information is not available to common Internet search engines for obvious company confidentiality reasons. There is nothing worse than working on a problem for six-months only for someone at a corporate event to ask you "didn't you talk to Sheila in Formulations; she did some superb work on exactly these active ingredients working out of Brisbane and found a solution back in 2015 with all the confirmatory spectroscopic data you will ever need!".

Evolving solutions in a digital world

Even where we have access to well-curated and indexed archive solutions, the route into each archive is often very specific to each technique. Even where we have, for example, chemical structure information associated with specific data sets in the individual collections, they may well be in formats or encoded in a way which is fine for the specific data source but is missing content which makes comparison of information across sources problematical. This has already been partially covered in a recent column where we showed many of the various ways that NMR spectra can be found in different data sources—fine in their own environment but a real problem if you want to compare them against one another.[1] So, imagine you want to search for a specific item in a museum—clearly the museum's catalogue would be a great starting point—hopefully detailing not only what is on open display but what they have hidden away in the archives. Now expand that question to all museums around the world and all their archives… and you get some idea of the problem we are facing. Additionally, of course this isn't a static problem—the archives and information resources are also evolving and expanding even as we search.

Fortunately, there are digital solutions evolving to meet the expanded challenges of a digital world. Instead of a printed catalogue for a physical world meet the digital Finding Aid for the information age. In a subsequent article we will look at the exciting work which is currently ongoing within IUPAC on the use of Finding Aids to interrogate digital archives of supplementary information for spectroscopic data and associated chemical structures and metadata, but in this column, we will look at one which you can have access to for free to build your own Finding Aid, the Bayer COrporate Linked Data (COLID) system.

COLID not COVID!

Many years ago, I met Rolf Grigat who has been working on spectroscopic data systems for most of his active career, and we even worked in the same organisation at Creon•Lab•Control and Waters Corporation for four years. He is now working in Bayer in Leverkusen, Germany and has been the product owner for

DOI: 10.1255/sew.2021.a52

TONY DAVIES COLUMN

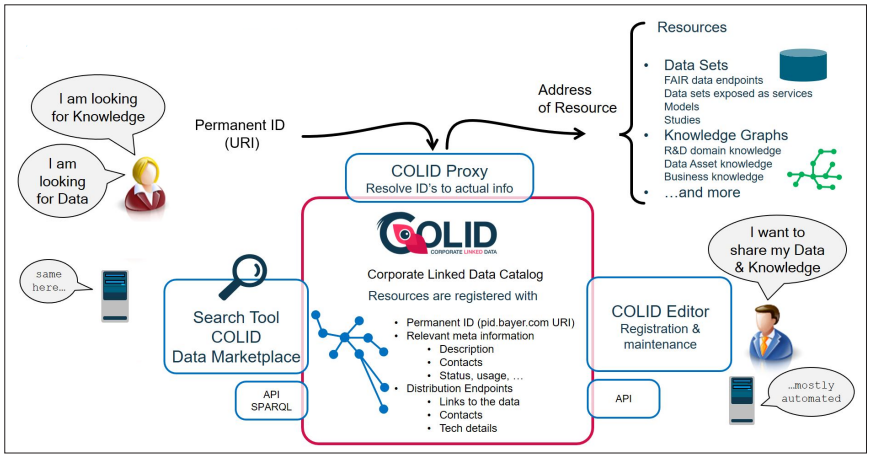

Figure 1. Highlights of the definition of data consumers user requirements for the COLID system.

the COLID data and information catalogue for more than three years. This is a solution which is designed to meet the challenge of making data and information assets FAIR[2] and linked for all data consumers (human and computerised) from both internal and external information sources. Rolf collected concrete requirements from the various business units and started the implementation as a cloud native application (built with cloud tools rather than taking a classical client–server application which is hosted in the cloud, which is more efficient and easier to maintain!). Figure 1 shows graphically the results of the data consumers' wishes.

COLID has been fully operational for two and a half years and is used cross-divisional and remarkably has been published (without company confidential content of course!) as a well-documented open source GITHUB project so that anyone can host their own COLID service.

How does COLID work?

Well, COLID is essentially a "catalogue of catalogues" collating any data source with which it is connected. Metadata is harvested for all the content it finds within the data source—all of which are provided with permanent Uniform Resource Identifiers (URIs). These information sources can be both internal and external unifying information retrieval in one uniform access application. Such a Finding Aid is ideal for collecting and providing metadata about basically any resource that you want to incorporate and a) link endpoints, such as spectra, in a repository or details in a chemicals database and b) link it semantically with any other related resource. So COLID could be simply be spelt as "F" in FAIR!

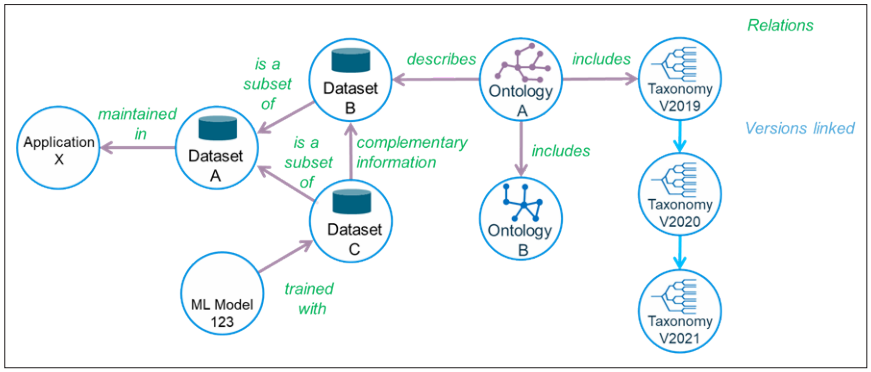

Figure 2. Simplified resources and relationships model for COLID showing the maintenance of linked versioning, transparent to the end user through the COLID applications.

A modified version of one of Rolf's simpler published explanatory graphics is shown in Figure 2.

So, if we were to take the available functionality and apply it to our use case outlined in the third paragraph, we might build something that looks like the outline system in Figure 3 with the taxonomies providing standardised classifiers for better comparison of the information entries.

My hope for the future is, of course, that Sheila in Formulations, who has such wonderful spectra, successfully registered her project, the spectroscopic data and the chemical entities involved in her work into a repository which is linked to my COLID Finding Aid. If so the chances that I would miss her work would hopefully be pretty much reduced to zero as my access application will find the corporate knowledge through spectral searches, chemical entity searches for any of the APIs not to mention reported intermediates. Hopefully, I will find all this out on the first morning of my new project and be enjoying reading about all the experimental results and associated project information, made available to me through the permanent URIs, in the same afternoon. Much better than waiting to find out six months later that my team have been wasting their time and Sheila in Brisbane beat us to it, on the other side of the planet, six years ago!

And stay informed!

Now we are happily informed about what Sheila achieved through our new Finding Aid, our work is supported further by keeping us up-to-date on any changes to the entries we have subscribed to. What we quickly discovered, for example, is that Sven's Applications Team in Västerås, Sweden is currently working with the same active ingredients, but different excipients, on a project very similar to ours but for a totally different client grouping. So there seems to be enormous potential to support each other and get both our projects over the line faster and cheaper than was originally budgeted for. Oddly enough, a small research group from the University of Mälardalen has also published some peer-reviewed papers from a local

TONY DAVIES COLUMN

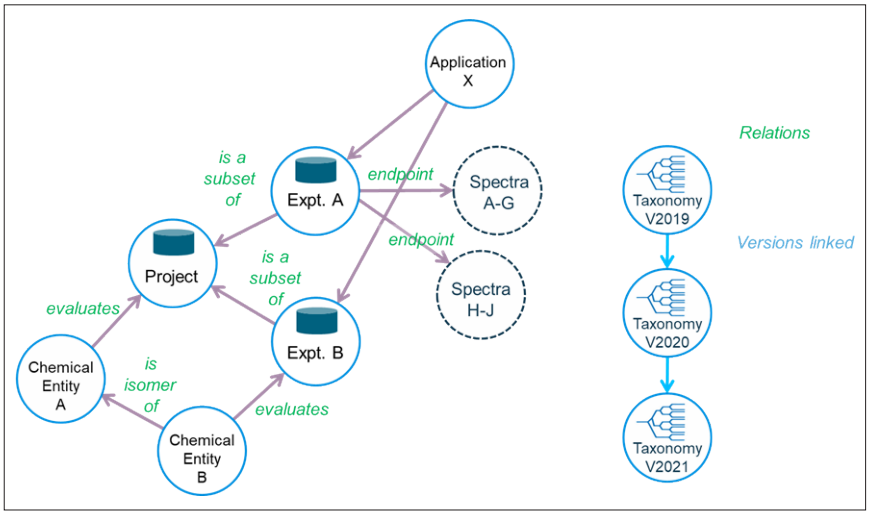

Figure 3. Potential COLID deployment capturing project, experiment, chemical and spectroscopic data with all associated metadata and relationship information.

conference in Swedish, including some of the chemistries we plan to explore—I would never have spotted these papers without the help of the Finding Aid. We will be keen to see how much freedom to operate we have in both our projects!

Conclusions

In the beginning of the article, I painted a deliberately grim picture of, to be quite honest, how I really feel about the way information is being made available to us in this information age. There are far too few fast, helpful tools to support us in our decision making and the sheer volume of information at our fingertips is often swamped by data of uncertain pedigree. Reading about functionality of these Finding Aids like COLID and following the excellent work being done within IUPAC on a Finding Aid aimed at supplementary spectroscopic information from peer-reviewed publications gives me hope that we have a brighter, more informed future ahead. Of course, as Rolf pointed out to me recently, the actual data sources are often highly confidential in nature so "F"inding out that information exists somewhere is not the same as actually having the permissions allowed to access it—which makes our Sheila and Sven example somewhat idealistic!

If you think you may have a requirement that may be met by a COLID deployment have a look at the documentation in the Github site[3] or even look to some information published by the Pistoia alliance.[4]

Figure 4. COLID logo.

References

1. R.M. Hanson, D. Jeannerat, M. Archibald, I. Bruno, S. Chalk, A.N. Davies, R.J. Lancashire, J. Lang and H.S. Rzepa, "FAIR enough?", *Spectrosc. Europe* **33(2),** 25–31 (2021). https://doi.org/10.1255/sew.2021.a9
2. https://www.go-fair.org
3. https://github.com/Bayer-Group/COLID-Documentation
4. https://fairtoolkit.pistoiaalliance.org/methods/fairification-workflow/

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