Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

The Physical Sciences Data-Science Service (PSDS)

Patterns

Failed it to Nailed it: Data Citations & Publishing

03/12/2020
Online Event

Dr Samantha Kanza & Dr Nicola Knight
University of Southampton

03/10/2022

AI3SD-Event-Series:Report-22

Failed it to Nailed it: Data Citations & Publishing
AI3SD-Event-Series:Report-22
03/10/2022
DOI: 10.5258/SOTON/P0035
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

An EPSRC National Research Facility to facilitate Data Science in the Physical Sciences: The Physical Sciences Data science Service (PSDS)
This Facility is EPSRC Funded under Grant No: EP/S020357/1

Principal Investigator: *Professor Simon Coles*
Co-Investigators: *Dr Brian Matthews, Dr Juan Bicarregui & Professor Jeremy Frey*

Contents

1 Event Details

Title	Failed it to Nailed it: Data Citations & Publishing
Organisers	AI^3 Science Discovery Network+, Patterns Journal & Physical Sciences Data-Science Service
Dates	03/12/2020
Programme	AI3SD Event Programme
No. Participants	23
Location	Online Event
Organisation / Local Chairs	Dr Samantha Kanza & Dr Nicola Knight

2 Event Summary and Format

This event was the fourth of the 'Failed it to Nailed it' online data seminar series. The event was hosted online through a zoom conference. The event ran for approximately 3 hours in an afternoon session.

There were three talks given on the topics of data publishing, data citations and supporting information for publications, both from a domain agnostic point of view and from specific domain experts. These talks were followed by an interactive panel session with the speakers discussing the different issues in these areas.

3 Event Background

This event is part of the 'Failed it to Nailed it' data seminar series. This event series, currently comprised of 4 online events is a collaboration between AI^3 Science Discovery Network+, Patterns & the Physical Sciences Data-Science Service (PSDS). This event series follows on from a data sharing survey that was undertaken earlier in 2020. Each event in the series handles a different aspect of dealing with data aiming to educate and inform researchers about how to work well with their data, as well as encouraging discussion along the way. Following on from these events the organisers hope to be able to organise more face-to-face events in 2021 which will expand this event series.

Understanding the different issues that need to be considered with respect to publishing data, and making it citeable is vital to ensure that researchers can both maximise the value of their research, and receive full credit for their work. In these events we want to encourage researchers to consider this as a fundamental aspect of managing, presenting and organising their data rather than an afterthought, or something they wish they did but never got round to. This event aimed to provide information and advice on different areas of data publishing and data citations to empower researchers to implement this advice in practice.

4 Talks

4.1 Data publication - a personal tale - Dr Sarah Callaghan (CellPress Patterns)

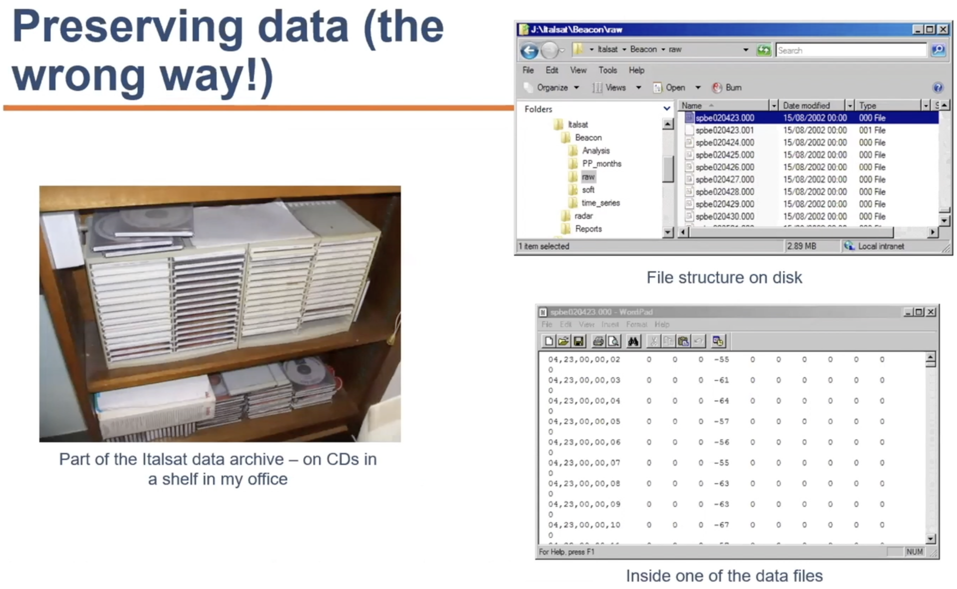
https://orcid.org/0000-0002-0517-1031

Figure 1: Sarah Callaghan

The full video of Sarah's talk can be viewed here: https://youtu.be/6SneQbYHwO0 [1]

Sarah Callaghan has extensive experience in creating, managing and analysing scientific data, and she currently works as the Editor in Chief for the Patterns Journal run by Cell Press. Her research started as a combination of radio propagation engineering and meteorological modeling, then moved into data citation and publication, visualization, metadata, and data management for the environmental sciences. She was editor-in-chief of Data Science Journal for 4 years and has more than 100 publications. Her personal experience means she understands the frustrations that researchers can have with data.

Sarah's talk takes us on a journey of her wide ranging experiences with data publications throughout her 20 year scientific career. She starts her story in 1999 right after she graduated from University with a degree in physics and music. Sarah began working on radio propagation measurements for a radio communications research unit at Weatherford Appleton Laboratory in Oxfordshire. Sarah's experiences in this position demonstrated how different data management used to be 20 years ago, although some of the issues with data management strategies from the 90's unfortunately still exist today, but thankfully there have also been significant improvements. The main aim of Sarah's job was to improve bandwidth by improving the congestion issues of the radio spectrum. Sarah's team collected signal frequency data from frequencies that were badly impacted by the weather, and to process and analyse them and turn them into statistics. This process was quite complicated and involved a large degree of data, and some intensive computation with a range of different software.



Preserving data (the wrong way!)

J:\Italsat\Beacon\raw

Name	Date modified	Type
spbe020423.000	15/08/2002 00:00	000 File
spbe020423.001	15/08/2002 00:00	001 File
spbe020424.000	15/08/2002 00:00	000 File
spbe020425.000	15/08/2002 00:00	000 File
spbe020426.000	15/08/2002 00:00	000 File
spbe020427.000	15/08/2002 00:00	000 File
spbe020428.000	15/08/2002 00:00	000 File
spbe020429.000	15/08/2002 00:00	000 File
spbe020430.000	15/08/2002 00:00	000 File

1 item selected 2.89 MB Local intranet

File structure on disk

Part of the Italsat data archive – on CDs in a shelf in my office

spbe020423.000 - WordPad

```
04,23,00,00,02     0    0    0  -55    0    0    0    0    0    0
0
04,23,00,00,03     0    0    0  -61    0    0    0    0    0    0
0
04,23,00,00,04     0    0    0  -64    0    0    0    0    0    0
0
04,23,00,00,05     0    0    0  -57    0    0    0    0    0    0
0
04,23,00,00,06     0    0    0  -56    0    0    0    0    0    0
0
04,23,00,00,07     0    0    0  -55    0    0    0    0    0    0
0
04,23,00,00,08     0    0    0  -63    0    0    0    0    0    0
0
04,23,00,00,09     0    0    0  -63    0    0    0    0    0    0
0
04,23,00,00,10     0    0    0  -67    0    0    0    0    0    0
0
```

For Help, press F1 NUM

Inside one of the data files

Figure 2: A slide from Sarah's presentation demonstrating data management and archiving issues

As Figure 2 might allude to, this came with a wealth of data management and archiving issues. The data was stored in CD format, in a proprietary format with less than obvious file names. Whilst these processes were documented, this is still not anywhere close to an ideal data management scenario. Furthermore, the types of documents were .idl files which are outdated and unlikely to be backwards compatible with software used today. Research papers were published on this work, with tables displaying aggregated data, but the capacity to go back and fully explore the raw data to potentially reproduce results or experiments, or even verify the correctness of this work is not available. Admittedly hard copies of project reports still remain on paper, but similarly these do not provide access to the raw data, and are limited in usefulness by being paper copies. This demonstrates the importance of making sure your data is FAIR (Findable, Accessible, Interoperable, Reusable) [2] and of future proofing your data management and storage, particularly if you do have a requirement to use proprietary software formats. It also raises an important point that typically a lot of data collection, processing, analysis, and even blood sweat and tears can go into making one table or figure in a research paper, or one slide in a presentation and this doesn't always receive the recognition and appreciation equated to the work put in.

Continuing on her personal tale, Sarah shared an experience related to this problem that had a significant impact on her career trajectory. Sarah unfortunately got 'scooped'[1] early on in her career, and datasets that she had worked on for several years ended up being used in another research paper without the appropriate credit being given to Sarah and her fellow researchers. This spurred Sarah into looking for ways in which we can actually give the researchers and people who create the data that underpin projects and publications the proper credit that they deserve.

Sarah makes the point that data science is all about data, and if we are to strive for properly reproducible science, then we need reproducible data. Unfortunately, obviously not all data is reproducible, or able to be verified or collected retrospectively. For example, when working with environmental datasets, if an observation was missed, then it is unfortunately lost. Nevertheless, to the best of our abilities, we should be making raw data available for validation and reproducibility. It is important to have access to the raw collected data to confirm

[1] Scooped: Publish a story/article before (a rival reporter, newspaper, or broadcaster).

that you are performing valid science! Poor data analysis generates false facts, and false facts and inaccessible data can lead to a loss of all credibility if results can not be checked and verified.

This has led to a drive for open science, where researchers are encouraged to make their data open and accessible in order to facilitate the verification of the science that is being performed. This is a very positive step for publishing and data management, as it allows for fact checking and quality control, and indeed helps to reassure research funders that they are getting value for money. However, there is an understandable reluctance for researchers to spend years collecting and curating data for use in their research, only to then hand it over to governing bodies for little to no recognition or attribution. This raised a requirement for finding methods for suitably attributing and rewarding researchers for their work on datasets.

This led Sarah to start looking at data citation and publication. It's well accepted that publishing is a great way to share your research as a researcher, however the notion of publishing your datasets (either independently as separate entities for others to use, or publishing them linked to your paper) is still less common, but is starting to gain traction. There are however a number of different ways this can be achieved, although some have more potential drawbacks than others:

- **Cloud Storage:** It is tempting to store datasets in cloud storage such as dropbox and google drive, and as long as they don't invalidate your security requirements this is a decent short term solution. However, it is worth considering that you don't know the longevity of these products.
- **Website Publication:** One avenue for publishing research data is via a project website. It's quite common for research projects to have websites describing the project, the work being undertaken and introducing the research team. However, it is unfortunately equally common that these websites are only maintained throughout the duration of the project, and once the funding runs out, so does the website maintenance and often the payments to the domain/hosting service, running the risk that the website might just disappear.
- **Supplementary Material:** Datasets can be published as supplementary material in a journal paper. However, this is often less than ideal, both because the datasets are often too big for this to be effective, and also because it is then harder for researchers to find them as datasets in their own right.
- **Repository:** Datasets can be published in a disciplinary or institutional repository, which can make them more discoverable than via the website/supplementary material methods, but it does tend to depend on how much of a reach those repositories have and whether other researchers know to look for your datasets there.
- **Data Journal:** You can also publish datasets in a data journal. These are becoming much more commonplace in the academic publishing world. These journals invite submissions of data science outputs rather than traditional research articles (which can be datasets, software, code etc), and look for details on the context of the data collection, the choice of software, data processing decisions etc. Whereas traditional research articles would contain information about the analysis of the data, as part of their results, these journals provide a mechanism to describe the data and facilitate access. This has a lot of advantages, it means that the datasets are available much earlier for re-use than they usually would be, and it means that datasets can be formally cited via these journals, and reviewed through a traditional peer review process to provide them with more credibility.[2] Further, it provides researchers with the opportunity to publish their datasets in a way without necessarily needing to have the novel analysis of groundbreaking conclusions yet.

[2] https://www.strath.ac.uk/openaccess/researchdatamanagment/datajournals/

Sarah then moves on to emphasise the importance of making these datasets usable. Unfortunately, open doesn't necessarily mean usable; you can make your data findable, and have it open, but if it's of terrible quality, it doesn't have any associated metadata, there are no instructions on how to interpret it or use it, then it is essentially useless. However, not all datasets can or will be made open; ones holding confidential data about patients for example are unlikely to be made open, but nonetheless these should be made to be equally usable!

It is also worth remembering that data can be citable, even if it is not completely open, much like research papers behind paywalls can still be cited. If your datasets are confidential then they can be embargoed to a certain time period or frozen so that others cannot access them. However, it is still very worthwhile to make them citable, even if it cannot be open. It ensures that other researchers know that your dataset exists, who is responsible and where people can find out more about it.

To conclude, Sarah reiterated that datasets are hugely important. We need the data and to be able to verify the conclusions and knowledge that are the result of scientific experimentation and investigation. Good decisions require good data, and if we don't have good data then we will not be able to draw good accurate conclusions, which fundamentally goes against what science is all about. All researchers have a responsibility to care for the data that they produce, and these datasets should be transparent, reproducible and verifiable. Finally, datasets should be curated and published in such a way that if anyone (including yourself 20 years down the line) wants to use them, they should be able to understand how the data was created, curated and analysed, and be able to use them effectively and accurately.

Questions following the presentation:

Q: Can you think of a better example of shared FAIR data than its use by machines in developing COVID-19 vaccines in an unbelievable time?
A: *No, I can't. I think the COVID-19 situation and the speed that we've been pushing being able to do research work to combat the pandemic. I think this is a really good example of the power of open science and the power that we can have when we work together and we share data. There has been an awful lot of effort put into sharing data, and that's paid off in so many many different ways in this situation.*

Q: Do most journals publishing descriptor articles charge an APC (Article Processing Charge), and if so what is the average cost of an APC?
A: *Regarding journal articles in general I can't tell you the average cost because I haven't run those numbers. It depends on the Journal, so I think the majority of data journals at the moment are Open Access, so they do have APC's associated with them. Regarding journals publishing descritor articles, most journals don't publish descriptor articles. There are specific data journals like Nature Scientific Data, or Data in Brief in Elsevier, that are solely around publishing descriptor articles. Patterns is one of the unusual ones in the field in that we publish descriptor articles alongside original research articles, and we treat descriptors in exactly the same way as original research. There is no second tiering of the datasets or code as far as we're concerned. We are interested in the high quality, high impact data science outputs, which descriptors are.*

Q: One of the worries amongst many scientists is that article processing charges are pricing them out of being able to share their scientific discoveries in a reputable Journal. Some publishers, I won't name them, charge between $5000-$6000 for an article, and I've been worried for some time now that the same trend will start up with data oriented things such as data descriptions. In other words, it will become very, very expensive to publish your data descriptor so that other people can recognize it and see, and it will become almost as expensive as when it comes to publishing the story behind it, which I call an article, and that's my concern.

A: *I understand completely, Open Access is really difficult. Its really important but the costs of Open Access are really, really difficult to deal with. I will say that you can publish your dataset without publishing a descriptor article associated with it. If you put your dataset in a trusted repository then you can get a DOI for it and you can cite it exactly like it was a journal paper, and if you have the right metadata associated with that dataset in that repository then you don't need to have a descriptor article published with it at all. So it is kind of up to you whether or not you want to publish a descriptor article to go along with it. But, I do completely understand the concerns about APCs and how we go about ensuring that people actually get to publish their data for their benefit, and publish other things as well for the benefit of the community.*

Q: Who are going to be the consumers of either the data descriptors, or the metadata which you said was kind of a replacement for a data descriptor? Is it going to be machines or people?
A: *Both, I'd imagine. Each depends on the data type. With machine readable metadata it is possible to use natural language processing to scan the thousands of journal abstracts to pull out ones relevant to your project. If we have machine readable metadata across the repositories, then we can start pulling together and synthesizing even larger scale datasets and bringing things together in the right sort of ways. This term is called data synthesis, which means bringing datasets from all around the world together to make the sum greater than the whole.*

Q: Should you focus your energies more on preparing better metadata rather than spending your time writing a data descriptor?
A: *It's completely up to you as an author and a data creator as to where you devote your time. It also depends who you are aiming to communicate with as well. If you want to communicate with other people doing standardization and communication between machines, then absolutely metadata is the way to go. If you want to literally tell your story about your data set, why you created it and make it available to people outside your original domain, then a descriptor is the more human readable way of doing that.*

4.2 Publishing and citing data in practice - Mr Jez Cope (British Library)

https://orcid.org/0000-0003-3629-1383

Figure 3: Jez Cope

The full video of Jez'z talk can be viewed here: https://youtu.be/PpMOkTnBMlI [3]

Jez Cope is Data Services Lead in the British Library's Research Infrastructure Services team, responsible for the Library's research identifiers service in conjunction with DataCite, and for implementing the Research Data Strategy. Jez has many years of experience in research data management with previous roles at University of Sheffield, Imperial College London and University of Bath prior to his work at the British Library. Alongside his work in research data management Jez is passionate about skills training for researchers and is a certified Carpentries instructor.

Jez's talk focuses on many of the key aspects in research data management, looking at the important considerations within publishing data, including: FAIR, open data, delivery mechanisms, data advertisement, citations and licensing. He also covers how these techniques can be implemented in practice and provides links to resources that will provide further information for interested parties.

Jez begins by outlining what is meant when talking about 'Research Data'. This relates to text documents, images, maps, spreadsheets, information from equipment, simply put it is any information used to support a research conclusion. From this we can extend to 'research data management' whereby we are managing this data to allow its use both now and in the future. This requires the data to be kept safe, documented, accurate, findable, useful, preserved and appropriately shared. All researchers do carry out research data management to a certain extent, but some researchers do more of it than others.

While the focus is often solely on the data, it is important to also think about the wider picture of the research. Although an article is the story, it isn't the whole story. To tell the whole story it requires consideration of the methodology, code, intellectual property, information sources, engagement activities and creative outputs. An article oftens contains these aspects only briefly, so it is important to think about how they can be communicated in a more detailed manner. Jez highlights the Wellcome Trust who incorporate this in their policy on requiring outputs from their funded research and actually require an outputs management plan when applying for funding.[3]

[3] https://wellcome.org/grant-funding/guidance/data-software-materials-management-and-sharing-policy

Some other frequently discussed areas within data are those of FAIR/FAIR data and open data. FAIR (Findable, Accessible, Interoperable, Reusable) is a set of guiding principles [2] to improve the reuse of scholarly data. It needs to be findable; publishing data that no one can find is pointless. It needs to be accessible; available in a form that people can get to it. It should be interoperable; using standard data formats with consistent standard metadata describing it, allowing integration with data from other sources. Finally it should be reusable; well enough documented that it can be used as the input to further research.

Open data is data which is available for others to reuse without a charge and without unreasonable restrictions, though there may sometimes be reasonable restrictions. However, Jez highlights that 'not all data can be open', and there are many situations, as also mentioned in Sarah's talk, where you are unable to make your data freely available for download. But, you can still generally create it in a way that is FAIR even if this isn't the free openness that is associated with open data.

Jez then talks about the opportunities that you can experience when making your research data more open and FAIR:

- **Raising your profile:** That includes you as a researcher, your projects, your department and your employer. Publishing research articles often takes a very long time, sometimes up to 12-18 months from submission. Making the underlying data available provides an extra channel to publicize your research, and stamp your claim on that research area and the discovery.
- **Protecting your credit:** Making your research available in a rigorous and open way actually reduces the possibility of being scooped, if someone tries to claim credit for your work you have the record of published data and research showing your claim to the work when you initially shared it.
- **Attracting new collaborators:** Allowing a potential collaborator to inspect your data is a powerful way to imbue trust. This allows them to understand the work that you do, how you structure your data, and how your work could potentially mesh with theirs.
- **Increasing citations:** While the number of citations you get isn't something that you should focus on, it is something that still holds some weight in academia. Making datasets available allows citations to be made directly to the datasets, but there's also evidence that an article published with available data gets more citations. If the data is available then the conclusions in the article are easier to probe and verify, making them more trustworthy. It's also easier to build on that work with new research which leads to citations.
- **Enable new research:** Availability of data can allow merging of data from several different sources, potentially with extra data that you've collected yourself. This opens up possibilities to unlock new knowledge that was present in the information but wasn't accessible because it hadn't been combined in the right way.
- **Improve transparency:** Funders, particularly government affiliated ones, are very interested in getting value for money. Making your data available makes your work easier to examine and is a way of demonstrating that you are doing the work that they're investing their money in.

But how do you publish data? Jez usually considers this in two separate sections: delivery, which is how you actually get the data from you to someone else, and advertising, which is how you make people aware and get them interested in using that data.

When considering delivery mechanisms it is best to pick only a single method. Putting data in several places can quickly become confusing for anyone looking for the data as there are several versions and also makes it tricky to amend the dataset. The best option is to submit your data to a repository or archive, as they are set up with long-term preservation and archiving plans in place so that the data will be available and accessible long after you've moved on and you won't have to maintain it.

Certain data may have complications with sharing, however there are often specialist repositories or different sharing mechanisms available to overcome these. With sensitive data, there are some data repositories and archives that have facilities with restrictions; this may be authentication, a vetting process or physically visiting a secure location. It is also possible to release the data under a data sharing agreement on request and if it's really big it may not be practical to be uploaded, which might result in sending it via a physical storage medium. However, this may have some costs associated with it, which could reasonably be expected to be covered by the requestor. This isn't a profit making enterprise, simply covering the costs. Something else that is becoming increasingly common is depositing a dataset alongside the facilities to analyze it. So that people can do their analysis where the data is rather than having to bring the data to where their analysis is.

There are also some other options that may be suitable in certain situations, however, generally these are options to avoid unless you have a really strong reason for using them:

- **Supplementary material:** most datasets are far too large to be represented in a table in a paper, and the pdf format supported in supplementary information is very poor for data re-use, so deposit the data elsewhere and then reference from the paper.
- **Personal websites:** if the department closes, or you move to a different job will the information continue to be available, or will it simply disappear?
- **Data available on request:** there are many better solutions and this can create a lot more work for yourself.

It has been pointed out that depositing data in a repository is often the best option and increasingly research organizations will have their own data repository which may be suitable for your data. However, if you are in a research area where data sharing is fairly common, there may be discipline-specific repositories or archives. These are often designed for storing specialized data types and understanding their preservation needs. They may also allow for querying and finding subsets within it. Jez highlights the registry of research data repositories[4] as a good place to find both discipline specific and general purpose repositories.

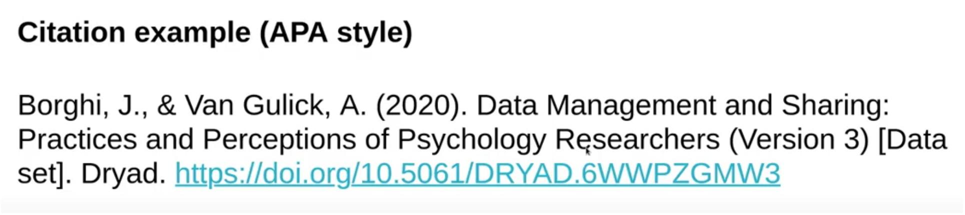
In contrast to selecting a single delivery mechanism, Jez advises picking multiple advertisement mechanisms. In fact you can utilise as many of them as you like, to make as many people aware of your data as possible.

- **Get a DOI:** Whenever possible, get a DOI for the dataset. This is a persistent identifier which will take you directly to the specific dataset, and is very helpful for locating the data. When publishing in a large repository or archive they will often assign DOIs for you.
- **Cite your datasets:** If you write a paper make sure to cite your own datasets. Increasingly journals are accepting that datasets are items that can be cited, but if you can't cite it directly, include a data availability statement, a few sentences saying where the data supporting the conclusions is available.
- **Write a descriptor article:** A short article describing the dataset you have produced. These are an increasingly available method of advertising your data and it allows you to provide more human readable context to that data.

[4] https://www.re3data.org/

- **Promote it at conferences:** present talks, posters, or discuss during networking.
- **Link from websites:** Don't upload it for download from your personal website but do provide links and a list of where you can get your datasets from.
- **Talk about it on social media.**
- **Share the links with colleagues and collaborators.**
- **Any other ways you can think of getting it out there.**

Citing your data should be fairly straightforward if you've followed all, or atleast most of, this guidance. Make sure you get a DOI and follow your department or publisher's referencing guide. Most major referencing guides now include guidance on how to cite datasets and if you use bibliographic database software (which we definitely recommend), they should be able to generate the citation for your dataset.



Citation example (APA style)

Borghi, J., & Van Gulick, A. (2020). Data Management and Sharing: Practices and Perceptions of Psychology Researchers (Version 3) [Data set]. Dryad. https://doi.org/10.5061/DRYAD.6WWPZGMW3

Figure 4: An example of a dataset citation in APA style from Jez's presentation

Figure 4 shows an example of a dataset citation, which looks largely like other citations you might be familiar with. It contains the key information; the creators, a publication date, a title providing information on what the data is about, version number for the dataset where applicable, in APA style you also flag up in square brackets that it is a dataset, Dryad is the location where the data is - essentially it's the data publisher and then you've got a DOI to allow easy location of the dataset.

If you don't use bibliographic software then Jez also recommends the DOI citation formatter tool[5] which allows you to paste in a DOI, not necessarily a dataset, then you select the referencing style and it will generate a formatted citation.

Another important topic to consider is licensing, which should be considered with respect to your own data, information you share and also information you reuse. Licenses specify what anyone other than the owner can do with it. It isn't a case of sticking software or data up on github and it's simply open. People can't do anything with it unless there is also a license that says 'if you've downloaded this dataset I permit you to do x'.

There are a number of different standard licenses available which can be reused. These are useful because they eliminate the need to create a custom license, but also because other people become familiar with the licenses and can see at a glance what they are permitted to do. For creative works the creative commons licenses[6] are widely used, for open data there are the open data commons licenses[7] and also some open government licenses relevant for open software. Looking at specific software licenses there are: The gnu public license, MIT, BSD and Apache licenses[8]. When working with software and data it is important to use licenses which are specialised for these object types.

[5] https://citation.crosscite.org/)
[6] https://creativecommons.org/licenses/
[7] https://opendatacommons.org/
[8] More information can be found from: https://choosealicense.com/licenses/

Across the different licenses there are common types of restrictions or requirements that may be encountered;

- **Attribution** - an attribution clause says if you reuse this, credit must be provided for it.
- **No-derivatives** - means that you can use and share this but it can only be shared in its original form. It's useful for opinion papers and things where something is stated as part of a whole argument and it shouldn't be quoted out of context.
- **Share-Alike** - this is an interesting condition which means that if you reuse the dataset then you can do whatever you like but the derived versions must be made available under the same license as the original.
- **Non-commercial** - this is a clause which says you can do what you like as long as a profit isn't made from it.

For most academic research Jez advises steering clear of this clause as there haven't really been any test cases and it isn't clear whether all academic research actually counts as non-commercial because the sector is increasingly having to act like profit making business.

Jez concludes his talk by highlighting that the topics covered in this talk are only a portion of the important considerations when looking at research data management in sharing and preserving data. A particular topic that should be of interest is data management planning. Some information, templates and examples can be found on the Digital Curation Centre's website [9]. However, that is just the beginning and there is plenty more information and training out there. If you are lucky enough to work in a university that has a university library, ask a librarian for recommendations! Managing information is something that librarians have done since the dawn of time and they typically have a lot of expertise in data sharing as well as the more conventional paper books and journals. University departments and libraries may often provide training on data topics, so keep an eye out for these.

Questions following the presentation:

Q: What, if any, validation would you expect a repository to conduct when data is submitted?
A: *So it goes right from one extreme to the other, general purpose big repositories e.g. Zenodo and figshare, do no validation. You upload, you click the button, it's published. As long as everyone knows and understands that that's what's happening, it's OK. Then there are various degrees along the line of validation, such as; vetting data, format checking, peer review. There are also certifications like CoreTrustSeal (https://www.coretrustseal.org/), which is a standard and a set of requirements that says this repository is being run in a particular way. It really varies a lot and if you want a particular level of service, you'll need to find their repository that provides that.*

Q: Can you explain how the no-derivative license is adapted to research outputs?
A: *So it's something I always mention because it's useful to be aware of. However, it's of limited use for a lot of research outputs. So first thing is that you can still quote pieces of a document that's protected under UK Copyright law, like an article, so the license can't prevent you from doing that. For something like a dataset, if you try to make it available under a no derivatives license you end up saying there's basically not a lot you can do with this data. You kind of render it a bit impotent.*

[9] https://www.dcc.ac.uk/

4.3 The (long) journey from supporting information to Publishing and Finding FAIR data in chemistry - Professor Henry Rzepa (Imperial College London)

https://orcid.org/0000-0002-8635-8390

Figure 5: Henry Rzepa

The full video of Henry's talk can be viewed here: https://doi.org/10.14469/hpc/7629 [4]

Professor Henry Rzepa since 1971 started as a synthetic chemist and then became a computational and information scientist and a spectroscopist. These research activities have generally generated large amounts of data and early on he became increasingly concerned that this vital research product was rarely treated as what is now called a first class scientific citizen. Since 2005 he has been trying to elevate his group's data to this status, using a combination of ELNs (electronic laboratory notebooks) closely coupled to what is now three generations of data repositories, to try to achieve its FAIRdom (Findable, Accessible, Interoperable and re-usable).

Electronic supporting information had its origins in the early to mid 1990s and it has evolved in a highly ad hoc manner since then. The concept of FAIR data arose about five years ago to try in part to rationalise the chaotic state of ESI. The talk will illustrate these developments by presenting a case study illustrating how one (either human or AI) might use the properties of FAIR to "F"ind some highly focused chemical spectroscopic and computational data. I will conclude by trying to unpick some of the supporting infrastructures which enable this and how the creators of the data facilitate this by using metadata to describe and then publish the data. The talk incorporates some elements of FAIR by having its own metadata and its own persistent identifier (as a DOI)[10], so that you can yourself Find, Access, Interoperate, Re-use and Cite it as appropriate.

Henry actually starts his story not at the beginning, but at the end by presenting what he wants to achieve with FAIR-enabled scientific data and then works his way through the steps to show how this can be achieved and how subject areas can benefit from FAIR. The use-case in question is chemistry data problem and the idea is to 'To **F**ind and then **A**ccess any data that conforms to a specific set of properties, for the purpose of its **R**e-use by **I**nter-operating it for a different purpose than its original context.' In particular, it will be looking at raw NMR instrument data, rather than a visual representation, examining the ^{11}B nucleus. In addition to the instrument measurements you may also want calibration or specification information and experimental information about the molecule and its solvent. However, this situation isn't necessarily intended for a human being to perform the task of retrieving this information, but

[10] https://doi.org/ff6g

instead for it to be machine actionable, allowing integration with ML (machine learning) or AI (artificial intelligence) techniques.

Looking at how supplementary information has been presented in the past, and up until quite recently, lends us to believe that this goal of readily retrieving the information is not a simple task. The evolution of the electronic supplementary information (ESI) has been rather ad hoc since its introduction in the 1990s, leading to lengthy unstructured ESI in the form of pdfs. Henry highlights a 2017 paper [5] with an exceptionally long ESI, the longest he has encountered, coming in at 907 pages. This pdf contained a whole raft of methods used within the research, purely in text form, alongside 485 pages of spectra presented purely in an image format, with very little extractable information. This was accompanied by 20 crystallograhic (.cif) files, which do present structured information for a handful of compounds in the paper, however these still also require human understanding to grasp their actual meaning in the context of the paper, especially with the compound naming scheme employed. While it's clear that the authors invested a lot of time in generating this ESI, it isn't really very useful for anyone wanting to easily access and use the data contained within it, and unfortunately this is how the majority of ESI is still presented.

There is however movement from Journals to change the way in which data is presented and linked to in articles. Henry in particular mentions the Nature Communications journal which, along with other Nature Journals, developed a policy in 2016[11] relating to production of a data availability statement [6]. This required authors to include a section in their manuscript detailing the availability of the data. It strongly encourages, but does not mandate, authors to deposit their data in a repository that will generate a DOI for the data, as discussed in the previous talks in this event. This is how many journals are beginning to think and many journals will now require a section on data availability.

Looking beyond the availability of data to how the data is actually described and discovered, Henry moves on to discuss metadata and how this is applied to research objects. Metadata is data that describes an object, but doesn't actually contain the object data and they are specified using a schema to define how they are structured. Research data has a schema defined by the DataCite registration authority[12] and when data is published a persistent identifier, PID, will be issued for it by the registration authority.

Many researchers might be familiar with a resolver such as the DOI resolver[13], which takes an DOI, a type of persistent identifier, and resolves to a landing page, or finding aid, which is a human readable interface. However they might be less aware of other resolvers for PIDs, such as the DataCite content resolver[14] which, among other things, can provide direct access to the metadata for an object[15]. These are intended for machine access and aggregation / harvesting of the metadata. This provides the basis for new search engine capability which utilises this harvested metadata.

[11] https://www.nature.com/documents/nr-data-availability-statements-data-citations.pdf

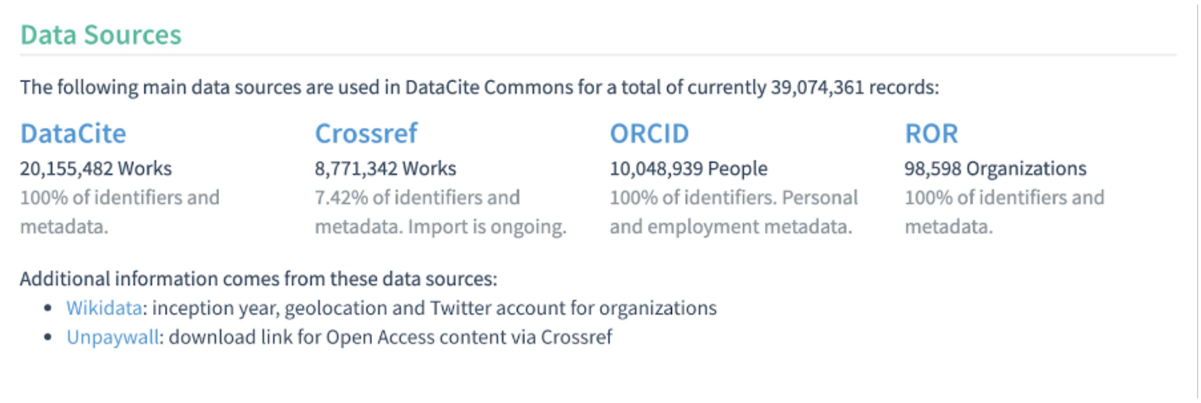
[12] https://schema.datacite.org/meta/kernel-4.3/

[13] https://www.doi.org/

[14] https://support.datacite.org/docs/datacite-content-resolver

[15] In fact the DataCite schema can be accessed via its PID https://doi.org/10.14454/7xq3-zf69 and the metadata can be directly accessed via the DataCite resolver https://data.datacite.org/application/vnd.datacite.datacite+xml/10.14454/7xq3-zf69

In 2020 DataCite and FREYA project announced DataCite Commons[16], which is a web interface allowing you to explore the PID graph. This graph combines information on publications, research outputs, researchers and organisations to produce a rich description. The graph hasn't been fully populated, but it currently contains almost 40 million records, Figure 6 shows the records incorporated for DataCite, Crossref, ORCID and ROR. Over 10million researchers have registered for their own identifier through ORCID and almost 100,000 organisations have their own identifiers. This starts to demonstrate what can be achieved through utilising identifiers and there will be a lot more in this area in the future.



Data Sources

The following main data sources are used in DataCite Commons for a total of currently 39,074,361 records:

DataCite	Crossref	ORCID	ROR
20,155,482 Works	8,771,342 Works	10,048,939 People	98,598 Organizations
100% of identifiers and metadata.	7.42% of identifiers and metadata. Import is ongoing.	100% of identifiers. Personal and employment metadata.	100% of identifiers and metadata.

Additional information comes from these data sources:

- Wikidata: inception year, geolocation and Twitter account for organizations
- Unpaywall: download link for Open Access content via Crossref

Figure 6: Data Sources included in DataCommons as of Oct 2020. Reproduced from https://doi.org/10.5438/mkpq-4w71 under CC BY 4.0 License.

Data availability examples

Returning to look at data availability statements, Henry presents a 2019 paper [7] from Nature Chemistry, which was one of the first articles he had seen with this new data availability statement. This article contains the normal sections that you would expect in a paper, but then additionally contains a section on 'Data Availability' with a statement providing a DOI for the data location. This meets the requirements from the journal and the funding council, however, how useful is this for examining the data?

The data provided in the repository is a .zip compressed folder at almost 5GB in size. To even explore what is contained within this dataset you have to download the full dataset. Once downloaded it contains 529 items, which are meant to refer to only a handful of molecules. This dataset is very much designed for humans to interact with, in conjunction with the other data and articles. Henry then highlights that if you were to examine the metadata as if you were a computer, you do not discover a huge amount about the data that might allow a computer to handle it. It contains some information about the author and affiliations, and a little information about the file type and size. The only other information it really contains is around the subject area which may be a little useful, but probably not of significant use. It is promising to see the data made available alongside the publication, however, this dataset might not be particularly suitable for machine processing which is the ideal goal.

Data availability 2020 ex - FAIR collection

Moving on to another example of data availability statements with a data collection that aims to be more FAIR Henry shows the data availability statement for an additional paper, in press at time of meeting, but subsequently published in nature communications [8]. The statement provides locations for two separate data collections, which are themselves comprised of datasets. This is more granular than the previous example as you can explore a specific file or dataset

[16] https://doi.org/10.5438/mkpq-4w71

without having to download an entire 5GB file.

Looking at the metadata for this collection you can find top level metadata for the collection itself, which contains, as expected, information about the author and dataset description, but also contains identifiers for each of the member components of the collection. The metadata for these components, referred to as filesets, can also be explored. This is where you can find more detailed information about the data, again in a structured format so it could easily be navigated by a machine.

Moving further into the exploration of the data, Henry explores the dataset information, which allows you to describe in more detail the information that is contained within the dataset. This allows more domain specific information to be encoded. In the example it contains Gibbs energy and InChI which are also linked to the schemas where the terms are defined, which allows a machine to determine what is meant by the terms. From the dataset descriptions you also have the ability to access the individual data files, with resolvers which also contain information on file types, which allows you to process the datafiles themselves.

Henry then demonstrates how this can be used in an interactive way, with a machine retrieving and displaying data for a molecule based upon the information presented in the linked records. This is referred to by Henry as a FAIR data table, and presents the data in a human readable, and recognisable, format, but it is also completely machine readable. This is an example of some of the elements of FAIR as the machine can find and retrieve the data but then also use it to perform additional operations on it, such as molecular mechanics calculations.

Packaging Data up
Frequently data is required to support some finding, a paper, or a specific piece of work, and in order to do that the data is required to be packaged up. In traditional ESI this has often been in the form of pdf, as discussed earlier. These pdfs often have little metadata or information about the information included in them, especially not in a machine readable format. There is movement towards the inclusion of a compressed directory, such as the .zip file outlined in the second example. However, this is still difficult to navigate, potentially not including metadata, a manifest or any index.

Other potential ways of packaging data include solutions in proprietary packages. Data formats and packaging is a topic which could form the whole basis of a separate talk. However, Henry briefly raises a project they have undertaken with Mestre Nova data, whereby the data is packaged up with a license which allows the data from that dataset to be opened even if the user does not have a full license to the program. This tries to address some of the access problems within FAIR, in that it is not only the data that you might require, but also adequate programs to process the data.

Packaging up your data in a dataset with a manifest and metadata to describe it is a more FAIR way of providing a collection of data and allows for human and machine processing. This is what Henry has described throughout the last couple of sections of this talk. These datasets can then be easily linked through their DOIs, or through the creation of a finding aid.

Finding data using the metadata
Now that Henry has gone through some of the issues with data publication and an explanation of their data publication solution, he revisits the objective of searching for specific data. In this search for data it allows for several search terms to be combined with boolean operators. For example searching for information on a molecule with a specific InChIKey, and for a specific

type of NMR data. However, these queries can become quite complex and it would be tricky for a user to formulate the correct search query. In order to capitalise on the power of searching, better interfaces are required, potentially with natural language processing, which will allow a human to easily formulate a search that can then be converted into the machine syntax.

An example dataset in the Imperial repository is shown, which goes through the different levels of a collection, showing the NMR datasets, and the packages which include the licensed datafiles. The search is then run on the datacite platform which queries its 20 million+ entries and you can see that the search query results in the same example dataset that Henry was showing previously. This shows that the navigation can be done by both a human and also by a machine.

Wrap up
Henry concludes his talk with a couple of take home messages to consider.

- Metadata and PID's are central to FAIR publishing and data citation. You should familiarise yourself with them.
- Metadata for data declarations is not well populated in most contemporary journals. This can be demonstrated by picking an article in your domain and seeing how populated the metadata is for a data declaration, if there is one present.
- For most data repositories, including generic ones such as Zenodo and Github that have been mentioned in the discussion of this event, the metadata offered is only at a basic level. It should be something that is generated automatically, as if the authors are required to create it that is an additional manual step and is likely to be ignored.
- Metadata should be structured against schemas, as creating rich subject metadata opens up that data to being harvested by algorithms. However, the subject areas must agree on how their metadata should be expressed, which is no small task.
- The harvesting of metadata is opening up a range of new commercial avenues, including the microsoft academic graph and is likely to continue to spark new chemical innovation [9] as graphs continue to connect up the diverse range of scientific objects."

Henry concludes by saying there is lots of new development still to come and to quote the American president Ronald Reagan, it would be fair to say 'we ain't seen nothing yet'.

Henry also acknowledged some of his colleagues with this statement: "I have to express my greatest thanks to two people at Imperial College, Matt Harvey and Simon Clifford, who both recognized the importance of these areas 15 years ago, which is when I started collaborating with them and what we've been able to publish in that 15 years is wonderful and I couldn't have done any bit without their help, enthusiasm and inspiration."

Questions following the presentation:

Q: A manifest is an excellent concept and acts as a useful aggregator of different datasets associated with a thing. How much expert effort did it require to generate your nature Comms example?
A: *All the manifest is done automatically by the repository. I don't have to worry about that from my point of view as an author. We actually work with an electronic lab notebook, and that's what I sit in front of most of the day, organizing my research, and so do my students, and the ELN has a publish button. When we've decided that the dataset that we've collected or generated is ready to be shared, all we have to do is click on the publish button. That generates the metadata, generates the manifest, does everything. There's a bit more organization required, but essentially it's all automated, so we don't really ask humans to do this, because humans are very, very capable of making horrible errors, which will throw everybody off. So the less human*

involvement, the better I say. We do the difficult bits which is actually design the experiment and conduct it and the data is handled by the machine.

Q: What ELN are you using?
A: *Back in 2005 when we started, there weren't really any ELNs available. The high performance computing division at Imperial College had just got new management. One of the new management, Matt, asked what you'd like us to do and I described what we now know as an electronic laboratory notebook. Now that's exactly what he went away and built. It took us a further two years of backwards and forwards before it actually started working quite the way we wanted to, but it wasn't a difficult thing to do. Right from the outset we said we need to put a data repository at the other side of this. I want you to produce a publish button and I won't have to worry about it. So yes, I'm afraid it's a bespoke repository. It works well for my line of work but it probably wouldn't work very well for pretty much anybody else's line of work. It was a collaboration between the researcher who knew more or less what kind of tools they needed, and a service provider who was willing to listen to what the researcher had to say and to act upon it. Which is why I gave my very profuse thanks to Matt and Simon. So it doesn't really help you with which ELN it is, but I suppose it does indicate how it came about.*

5 Panel Session

Following the presentations we ran a panel session where members of the audience were able to ask questions to the panellists and participate in discussion. The panel consisted of two of our speakers from the earlier presentations joined by Nushrat Khan, a data research specialist from research data services at the University of Bath.[17] Some additional comments were provided by our third speaker Jez Cope for some of the pre-prepared questions.

The panel members were:

- Dr Sarah Callaghan - Patterns
- Professor Henry Rzepa - Imperial College London
- Ms Nushrat Khan - University of Bath

The panel was chaired by Dr Samantha Kanza & Dr Nicola Knight. The questions asked to the panel were a mixture of pre-prepared questions and questions asked by members of the audience. The report content below outlines the questions asked and summarises the discussion and responses that followed these questions. This also includes discussion carried out in the meeting chat.

[17] Now at UCL

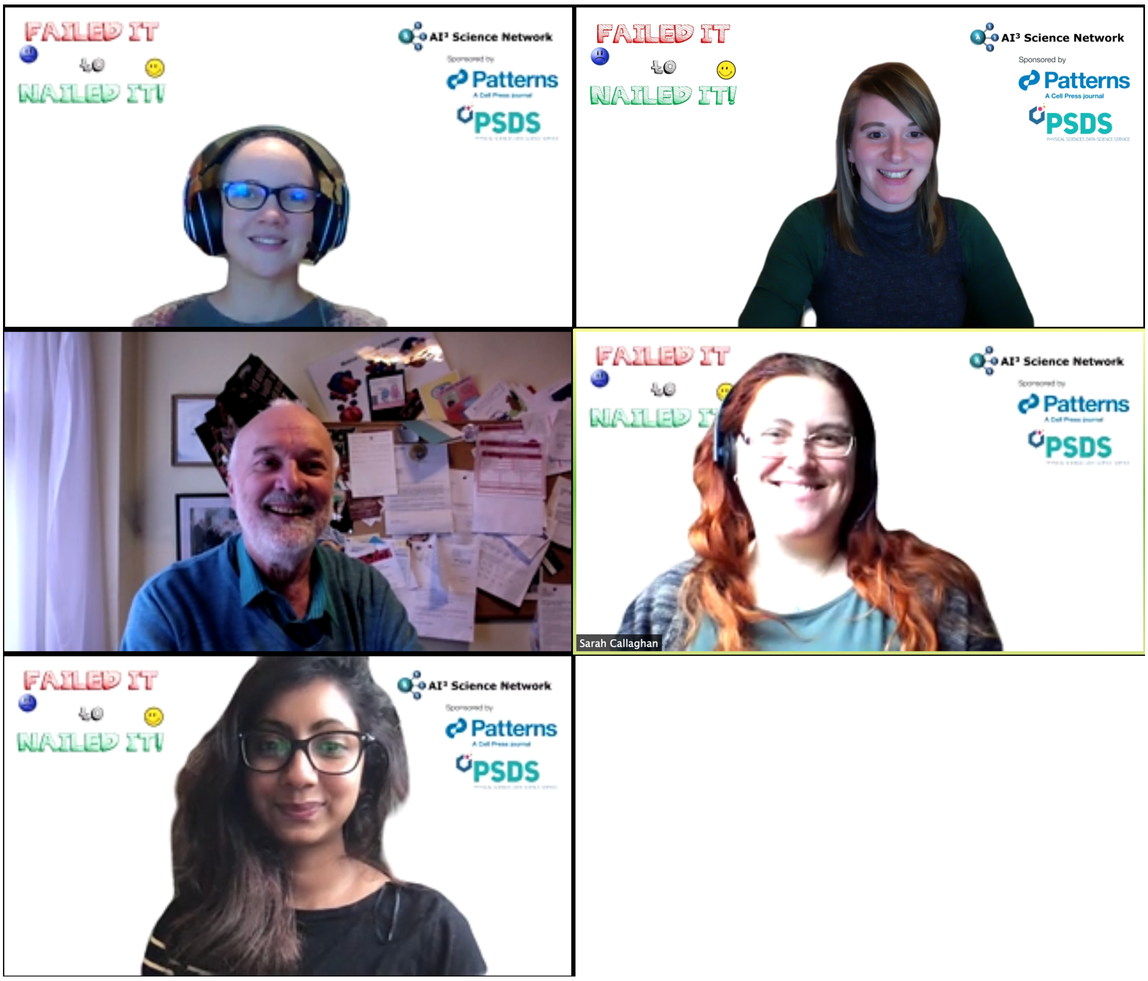


Figure 7: The Panel members

Q1 - What excites you / are you hopeful about for the future of data publishing?

- We have been working on data citation and publishing for a long time and it feels like now the work is starting to pay off. Data citation and publication are becoming more common practice.
- The increasing number of datasets being cited, being published and being recognised as first class research outputs.
- Researchers utilising open datasets to create new research avenues.
- How FAIR data can play a pivotal part in data preparation and analysis. A particular example of this is handling the data surrounding the creation and testing of COVID vaccines in such a short space of time, and this inspiring talk given by Barend Mons from GO FAIR[18].
- It's really promising that data publishing hasn't been captured by a single stakeholder, but instead the researchers, institutions, funders, libraries and publishers are coming together to work towards better data publishing.

Q2 - What about publishing data on GitHub or Bitbucket to be truly open?

- It was mentioned how GitHub doesn't have any persistence and longevity guarantees, which is potentially a serious issue, but there is the connection between GitHub and Zenodo where you can essentially take a snapshot of a GitHub repository and dump it

[18] https://odileeds.org/events/northernlands2/talks/open-data-saves-lives-barend-mons

into zenodo, and it gets stored there. This does have the complication of course that the authors can decide later to remove that particular snapshot. So the DOI landing page will still be there but it doesn't actually guarantee that the underlying software code, or data ported over to Zenodo will actually be there.

- One of the things that I found interesting is that data is fragile. GitHub could go away, or the data could be in a form which is unreadable in five years time, but the metadata may be immortal, we hope. In fact DataCite say that it is their expectation that in 50 years time the metadata will still be there even if the data isn't and perhaps someone will find the metadata useful even if they can't access the data.
- There was discussion about how best to set up repositories or other systems for making open data available along the lines of arXiv for preprint servers, and I think that's a really good idea, but it's a really hard problem. Papers are relatively easy to deal with as they are essentially PDF documents, and don't tend to be that large in computing terms. But you can very easily have a single dataset that's multiple terabytes or even petabytes.
- We somewhat have data repositories for orphan works already with figshare and Zenodo. But looking beyond that it's a matter of how we actually manage them, and how we fund them, how we make them sustainable long term, because storage is not free and neither is bandwidth. We've been riding Moore's Law and dropping storage costs for a long while now, but people are seeing on the horizon the looming physical restrictions on the media. Soon we're going to run out of space and the capabilities of silicon to actually store data. So we have to come up with some new clever ways of thinking about this.
- I confidently expect in the future, like 10 years, 20 years from now there will be a subspecies of data scientists who are the Data archaeologists they will be going into old archives and old formats and trying to extract old data out and then put it into more usable formats. A kind of Indiana Jones of datasets, I'm convinced that's going to happen sooner or later.

Q3 - What are good practices for encouraging data citation and how can we capture how people are citing or reusing data?

- DataCite is working on doing this and they have some very fascinating insights into how people are citing data. An important part of their PID graph is data citation. They decide to actually recognize the importance of showing people how data is being used, being cited, how often and what outcomes there are. So this is very much one of their briefs and a fascinating one to watch as well.
- The ways in which scientists cite data can vary quite a lot. I think the best way to encourage scientists to cite someone else's data is to have data access statements, although these need to be good quality. A recent publication looked at thousands of Open Access publications and how they included data and access statements. Even when people often include access statements they don't even have a DOI or persistent identifier, sometimes they just say data is available on the website, but that's not really useful and doesn't give enough metadata.
- So in terms encouraging citation, we ask the authors to do it and we encourage them to do it and we check to see if they have done it and sometimes uptake can be variable. But generally if we're asking them and prodding them, nudging them gently, they do tend to do that.
- What we do, and many other cell press journals are moving towards, is that the submission process requires authors to provide details of their datasets and new code, which are then passed on to the reviewers. This allows the reviewer to actually evaluate those datasets and code as well. All our papers come with a data & code availability statement on original research articles, and I encourage very strongly moving away from the 'data available on request' version of the data availability statement because quite honestly, it doesn't work particularly well, and it's a pain for the lead contact five years from now if somebody

comes along and goes, where is that data? And they go, where did I put it? So we try to encourage people to use a repository for their own benefit, as well as for the readers.

- I want to encourage people to do data citation as a more formal thing because it enables those tracking metrics that were mentioned earlier, however, the citation metrics are not the be all and end all of metrics.
- There's a slightly discredited metric called the H-index, about 10 years ago, which is basically the citations of your articles. When putting in your promotional request or salary increase, whoever was about to judge your performance would ask 'what's your H-index'? If you're 35 years old, it should be this by now etc. We don't have an h-index for data citation, and going by that, perhaps we shouldn't. However, for the next five years, in order to get people citing data, we ought to give recognition for data citations. So, if 10 people have accessed 10 of your data citations that will be a data h-index of 10 with recognition for doing that and then you get another one for 20 etc. Let's give people a gong of some kind, even though it's perhaps a meaningless gong, to just encourage them to cite their data. I'm not sure it's necessarily the best way, but it worked for 10 or 15 years, the h-index was the Holy Grail of your academic ladder, and maybe it would work, maybe it wouldn't. Someone would probably develop something better.
- I think the issue that we have with citing data is that it's still not amalgamated into the same systems that are used for citing articles at the moment. I know Clarivate was looking at doing the data citation index, they've been looking into it for some time, but I haven't heard about how far they've got with it. They did launch something, but it was a paid for service, which didn't help with their uptake.
- There was discussion about DataCite acting as a way of doing citation counts, or a group who would manage data citation counts. But I don't know if they've managed to do that at all. They have certainly illustrated their PID graph, and citation count will be a node in that graph and this could possibly be refocused to show who the most cited researchers are for data.
- People do get competitive as soon as you put a number on anything and compare people, they start getting all kinds of competitive and thinking my number has to be better than their number. It's human nature.

Q4 - I saw a citation schema recently which worked on adding extra information on the context of a citation in a paper (builds upon, critique of, related work etc.), do you think something similar to this could work for data citation?

- I know the citation schema that you're talking about, by Dave shotton at University of Oxford. He's done some amazing stuff and emphasized it's the quality of the citation that matters. You can cite something because you think it's rubbish or because you think it's great, there's no information included about it. He's really done a lot of work in elevating citations into first class scientific object in their own right, just like data. It's worth looking into what he's done leading the way in citations, and he could probably also tell us how to encourage people to cite data [10].
- The dataCite metadata schema, in the optional metadata, there is the opportunity for people to link their dataset to another dataset using the 'is related to' identifiers and I believe other bits of that schema allow you to do 'is partial', 'is a subset of' and other things like that.

Q4a - If I were to join several datasets and build on that to create an additional dataset, you then potentially have a combination of multiple datasets, which all in themselves have citations, which then gets into a very complex object. Is it necessary to build upon this and have extensions of how we describe the data to ensure that people can get recognition and acknowledgement where their data is reused?

- I know that Martin Fenner ran a datacite session several months ago with exactly that focus. How to exploit the datacite schema to provide this sort of information. The trouble is that no one could agree on how to interpret the schema, each person there interpreted the schema in a different way, and after about 2 hours of discussion there was general agreement that we would abandon that approach as being incomprehensible to most people, so it's a really difficult nut to crack I think, as to how do you give various qualities to a citation that everybody can agree upon and then make inferences from that agreement? It's a tough nut and I think they gave up temporarily trying to do that.
- It's definitely not an easy question. The phrase that gets thrown around a lot is attribution stacking, the fact that if you've got datasets where you have to acknowledge the previous creators and you merge them together, then you have to include all the other people who were involved as well and that rapidly gets unwieldy. But it's not beyond the realms of possibility because there's been precedent with the many Journal articles published from the Large Hadron Collider, where the list of authors is actually longer than the entire contents of the paper, so we can deal with this sort of thing. It's not necessarily a technological problem, it's more figuring out what the conventions are going to be, and I think that's where the problems lie.

Q5 - Do you have any suggestions or recommendations for how you might encourage people to publish their data using these good practices?

- Make it easier for them to do the right thing. Researchers are busy, they've got lots of stuff on their plates, they want to publish because that's the way they communicate with other researchers and get the impact, knowledge and recognition that they deserve. But writing and publishing an article is difficult, writing an article with datasets associated with it is extra difficult. I acknowledge that fact so it's up to the people who build the systems if we want to encourage good practice, then we should make it easier to do the good practice than it is to not do the good practice and I think that's just a general rule of thumb.
- I'd like to describe an experiment that we did this January with first year students just before COVID restrictions. We had 170 students doing their first set of practicals, with an experiment designed so that each student was making a molecule completely new to science, and we thought we will really motivate them by getting them to publish the data for that molecule. They recorded the data, we taught them how to publish it, they published it, they got a DOI, and within three months of joining Imperial College, they were a published author. Admittedly of a dataset rather than an article, but that proved very motivational. Of course you might ask, what was the quality of that data? Because they were inexperienced chemists, they might not have made a very pure molecule and we were publishing the data for an impure molecule. Some purists would say, that's rubbish data, you shouldn't be polluting data with this kind of poor quality data, and on the other hand someone else could analyze the data and say, oh yes, but it was pure. We didn't want to make that decision about the quality of the chemical data. But we wanted to motivate students, and it certainly worked, and so that's one of the ways in which you can engage with students right from the start, one of these days we'll be doing it at school, why wait until University? Unfortunately, then Covid restrictions started and we don't know whether we're going to run it again next January, we were hoping to do this every year with a completely new set of molecules.
- To follow up on that notion of good quality data versus bad quality data. This is one of my pet peeves, because data are data. They are their own thing and often quality in data is determined by the use to which you want to put the data. Back in my radio propagation research, my time signals from satellite would get very, very noisy. There were fast fluctuations due to the rapid scintillations in the atmosphere and there were

slower fluctuations due to the effects of rain and clouds etc. So I would have these two different time signals in my time series. For a researcher studying scintillations the rain signal was completely unnecessary and bad data as far as they were concerned, for the researchers studying rain, the fast fluctuations were bad data. But yet that's how the data came. It is what it is and it depends what use you want to put it to. The time that you get bad data is when you can't actually use it afterwards, when it's not been properly described, stored or file formatted etc.

- On the notion of bad data, my favorite thing to describe is the 16th century ships captains who sailed around the world collecting data, writing temperature and atmospheric pressure and weather conditions in their ships logs. They had absolutely no idea that 400 years later those measurements would be used as ground truth for climate models. We do not know what potential uses data could be put to in the future. For those students there is a small but non zero possibility that one of them might have accidentally found something really, really important in there. I mean, it's pretty small, but it's still possible.
- We did have a discussion about whether to publish this data, with some saying that the students shouldn't publish this data because we don't know how pure the molecules are and that will confuse everybody. This is a discussion that is ongoing in many places where data publication is raised, we don't want to publish it because it may not be good quality data, and so we don't. But possibly one of the most important points that we've made this afternoon is that data are data. Someone else can decide if it's good or bad data. Someone else can decide whether it needs curation or not. But that's not the function of the person publishing it.

Q6 - Do you have any views on movement towards people publishing 'unsuccessful' data?

- What does it mean by unsuccessful? If the outcome of doing the experiment that generated the data was unsuccessful? You know that might still be of interest to someone. There's been a long discussion in chemistry about publishing a Journal of irreproducible results. We got this result but we couldn't reproduce it, but here's the data maybe it's useful. It's never really succeeded, that discussion will be ongoing.
- I've heard something similar in a Journal of negative results and I think that's actually really important. There's been movement in this in the clinical trials space, in the fact that people who do clinical trials are expected to register the trial when they start the trial, rather than having the situation where they registered the trial and discovered that the drug that they're trialing doesn't actually do what they said it would and then quietly shoving the results in the desk drawer somewhere and failing to publish. Whereas if at least the trial is registered, people can go back and check what happens and that helps deal with the bias in only reporting the new and exciting science.
- I think we could do with a culture change in science as a whole, where reproducing and verifying scientific results is considered to be as important as getting there first, because we do have a reproducibility crisis and going on at the moment in a lot of different areas and we need to sort that out really.
- It might be similar to the clinical trial thing, but I'm sure I've heard something similar proposed about journals where you could propose an experiment you wanted to do. The Journal could choose to accept or reject it and then you had to publish it irrespective of whether it was actually successful or not, which I think could be quite useful, because then you're taking away some aspect of the negative of unsuccessful results, because they're just as important. When I did focus groups as part of my PhD I looked at digitising research, so a step below publishing it where they're not even digitising certain parts of it. A lot of the bits that didn't get digitized were the things that didn't work, so the failed experiments or the initial versions of protocols that didn't work before they made

substantial tweaks to them. I think it almost goes all the way back to generating that work in the first place that we need to start changing the attitude around making that available, even if you're just sharing it with your group. So you say I tried this and it didn't work. So if you're going to try it maybe try looking at it from a different angle, or even see if you can reproduce it if that's where you want to go. But it's nice to be able to share those things as well because I think that's just as important.

- I think it is important to make sure that there's essentially enough of the metadata and the context around it to allow people to make the decision about whether they think that it is high quality for their purposes or not. With a lot of these new technologies like applying machine learning and algorithms to the data. You need a full picture of the data, both 'successful' and 'unsuccessful', otherwise the data input just isn't going to be as useful.

Q7 - I have a question on quality from a repository's perspective. We want quality published data (a description of the data and how the data can be useful for someone else). I recently asked repositories whether they check their data and how they do it. Most institutional repositories said they use librarians or other information professionals and very few discipline specific ones had some automated checks. I guess we want to move more towards automated service as some institutions lack funding for the human resource required, do you have any examples of any institutions developing such services, and how should we move forward?

- In my previous job before I took on the role of editor in Chief, I was working for CEDA[19], a data repository, and we talked about this quite a lot. Now CEDA is a domain specific repository, so most of the staff have expertise within the Atmospheric Sciences, Earth observation, climate modeling etc. So they knew what they were looking at from a scientific point of view, but we were still discussing the idea of different levels of review of data. So you have the technical review of data asking questions such as; Is the metadata submitted with the dataset complete as far as the metadata schema in the catalog was concerned? That's fairly easy to answer, just see whether all the metadata tags had something in them. Then you would have the scientific question e.g., Are those tags actually appropriate for the data? So you could have the technical check, does the metadata tag have an entry in it and does that tag conform to the controlled vocabulary that is associated with that particular measured data entry? Those things are easy to be checked automatically by humans or machines, probably machines, it's quite boring doing that. Finally whether or not the actual scientific value is correct, e.g., if somebody has air temperature in a data field, tagged with air temperature and they've put air temperature at mean sea level and it's actually air temperature 100 meters up, then that's something that will not get picked up by the automatic checking because they're both proper controlled vocabulary terms, but they might be wrong in the scientific context. It's the yes/no questions that are easy to answer and they're easy to automate as well.
- It's a lot easier for people to automate a system if they're in a domain specific repository with well established community standards that everyone agrees on. When you've got the institutional repositories who have to take everything from the institution or just general repositories, that's when things start getting more complicated, because if you're dealing with astronomy to zoology and everything in between, the core metadata that you can mandate suddenly gets very small indeed. The specific ones you start struggling with finding controlled vocabularies for what they mean and all the rest of that, so that's where things become problematic.
- As we go along as we're training researchers to automatically think about how they describe their datasets and think about using community standards, which is going to

[19] https://www.ceda.ac.uk/

help a lot. It also helps that a lot of funders are now thinking about these things as well, and are talking about using community standards or developing new standards for data sharing and metadata, and if the funders are involved that generally means that there's a bit of money to go towards these sorts of things as well. It's a long and hard road, but we're getting there.

- You put a finger on a key point there, because money has to come in at some stage and someone's got to pay for whatever level of quality control you're going to do. It doesn't happen very much, but it's still money and if you care enough about the data, you can assess its quality. But I suspect a lot of people there won't be anybody around to care about most data. Which is why most data will eventually evaporate because no one is interested in it. Whether that's right or not I don't know, but that's the practicality of it.

6 Participants

Participants attended from a wide variety of backgrounds due to the online nature of the event. While the majority of attendees were from the UK, there were a number of registrations from other countries.

7 Conclusions

This event captured a large degree of useful information on how to future proof your data management and storage for publication and citation, and there were several key areas of advice from our expert speakers and panellists.

A key area of discussion that came up in every talk and the panel was the matter of making your data FAIR. Henry Rzepa noted that "Metadata and PID's are central to FAIR publishing and data citation. You should familiarise yourself with them". Your data needs to be Findable and Accessible, as it doesn't matter how high quality it is if no-one can locate it or gain access to it. It is generally advisable to store all of your data related to a publication or project in one place. It also needs to be Interoperable and Re-usable. Making your data open doesn't immediately make it useable. Whilst it is important to provide access to raw data to be able to confirm that your science is valid, and to provide reproducible science, this data still needs to be useable to be useful. Publish useable raw data in a common data format with descriptions and instructions, and demonstrate how you collected/cleaned/analysed it so that your work could be replicated. Additionally, it's important to note that not ALL data can or should be open (e.g. personal data) and that there might be situations where you wait to make your data available until after publication.

In line with knowing when to publish your data, there were some key recommendations about where to be cautious with data management and sharing. Sharing your data is a useful thing to do, however there is always a risk of getting scooped if you share too early; although there is also a similar issue of researchers producing material similar to your work before you if you take too long to get your data out there, so there's a fine line to walk with respect to when to share. Another cautionary tale was to be very careful about using proprietary software, it's not just about how to get your data info the software and what it can do with that data, you also need to be able to get the data back out again and not lose it to a costly (inevitably potentially outdated) proprietary format.

Another common area of discussion was around attribution and recognising effort. It is important not to underestimate how much effort goes into the tables and figures in papers, and indeed in collecting all of the data in the first place. There should be proper attribution and reward for the researchers who produce datasets, and if you are using someone else's dataset then it is really important to cite them. Further, researchers should put their datasets in data journals to make sure that it is easy to get the credit they deserve for their work, and should ensure that all of their data have DOI's whether that is through data journals or using free systems such as Zenodo or FigShare. There is a long way to go in the pursuit of data sharing and publication, but we are encouraged by the degree of progress that has been made in recent years.

8 Related Events

Details of the other events in the Failed it to Nailed it data seminar series can be found here:
https://www.ai3sd.org/ai3sd-online-seminar-series/data-seminar-series-2020/
Each of these events will have video recordings and a report associated with it.

Details of other AI3SD events and events of interest can be found on the AI3SD website events page:
https://www.ai3sd.org/ai3sd-events/
https://www.ai3sd.org/events/events-of-interest/

References

[1] Callaghan S. AI3SD Video: Data publication–a personal tale. 2020.

[2] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3(1):1-9.

[3] Cope J. AI3SD Video: Publishing and Citing Data in Practice. 2020.

[4] Rzepa H. The (long) journey from supporting information to Publishing and Finding FAIR data in chemistry. 2020.

[5] Lopchuk JM, Fjelbye K, Kawamata Y, Malins LR, Pan CM, Gianatassio R, et al. Strain-release heteroatom functionalization: development, scope, and stereospecificity. Journal of the American Chemical Society. 2017;139(8):3209-26.

[6] Announcement. Announcement: where are the data? Nature. 2016;537(7619):138.

[7] Jones CD, Simmons HT, Horner KE, Liu K, Thompson RL, Steed JW. Braiding, branching and chiral amplification of nanofibres in supramolecular gels. Nature Chemistry. 2019;11(4):375-81.

[8] Rzepa HS. A Thermodynamic assessment of the reported room-temperature chemical synthesis of C2. Nature communications. 2021;12(1):1-3.

[9] Computation sparks chemical discovery. Nature Communications. 2020;11(1):4811. Available from: `https://doi.org/10.1038/s41467-020-18651-x`.

[10] Willighagen E. Adoption of the Citation Typing Ontology by the Journal of Cheminformatics. BioMed Central; 2020.