DDAHS 2026 summer school Workshop 5 Creating FAIR Outputs-20260716_103434-Meeting Recording

16 July 2026, 09:34am

1h 21m 41s

Amulya Chevuturi started transcription

Amulya Chevuturi 0:03

Has started.

Jasmine Hunter 0:04

Okay.

Els Dhiedt 0:05

Cool. I'll also share my slides. Hopefully everyone can see that. All right.
Um...
Otherwise, shout or put in the chat if you can't see my slides. So yeah, welcome everyone today to this 5th workshop.
Yeah, so here we're going to talk about how to create FAIR outputs and go a bit deeper into the FAIR principles. But first we'll kind of start with introducing ourselves for a bit. So I am Els. I am a data scientist here at the UK Centre for Ecology & Hydrology, but I'm also a data steward, so help and support scientists with their data management. And I'm joined here by Jasmine, who will also introduce herself.

Jasmine Hunter 1:06

Yeah, hi everyone. So my name is Jasmine. I'm a data steward at UKCEH as well. I also sit in the digital research group. So we do some research looking at different ways to get data management and digital research infrastructure to help other scientists.
Yeah, we'll kick on.

Els Dhiedt 1:24

Thank you.
So I'll maybe do a little bit of housekeeping as well. So we're going to share the slides. Well, the slides are already in the GitHub repo, so you don't need to

vigorously take notes. They'll be there. And we'll have some time in the middle and at the end for questions. So
You can address those there. But if you have any questions in the meantime, you can put them in the chat so you don't need to remember them all throughout.
And yeah, there'll be some kind of interactive elements. We'll do some polls and there'll be some messaging in the chat possibly. But also feel free to put up your hands or unmute if you want to actually
Say something. Yeah, Gianni.

Gianni Vesuviano 2:19
Yes, sorry, I'm back. I have no idea what happened to my internet connection. But I hope everything you're able to start recording and so on.

Els Dhiedt 2:28
Yeah, yes, that's fine. Amulya is recording, so it should hopefully be fine.

Amulya Chevuturi 2:28
Yep.

Gianni Vesuviano 2:32
Thank you, Amulya.

Els Dhiedt 2:35
Cool.
So today we'll cover these topics. We'll start with kind of explaining what FAIR is, then we'll go a bit deeper in how to prepare these FAIR data, how to publish FAIR data, and then how to prepare supporting documentation and metadata. And then finally, we'll share some of our
recommended tools and resources and obviously get into questions as well. So we'll start with the first section, what is FAIR, and yeah, go into the FAIR principles. But first we want to kind of gauge your knowledge on the FAIR principles.
So each of these letters stands for one of the principles, and Jasmine has just launched a poll. Please, yeah, pick your, which one you think is for each letter. And so there are slides, so you can have, you have to click the next one to click.
Yeah, vote for the next one.

Ohh, yes, polls are in the limits, indeed, so...
So I think most people have.
put their answers in. So in that case, we'll go to some of the results.
So the first one is findable, and I see that most people indeed have put that in. So findable is just literally, can you find the data? And that's by humans, for instance, by Googling or by machines as well. Then the second one, most people indeed, again, were

Els Dhiedt 4:52

Correct, is accessible. So that's once you have found the data, is it easy to access and download these data? This can be by download button, is it easy to find on the associated webpage? Or if it's really large data, it's maybe not really feasible to download everything. So APIs and user interfaces,
come into play here.
Then interoperable is indeed correct. It's a bit more a difficult one. But I see that most people actually put interpretable, but it's interoperable. So some people have got it right. But it's
So what it means is that it's easy, whether your data is easy to integrate with other data and workflows and applications. So kind of making it in a standardised way or so it can be actually integrated with existing workflows. But we'll go deeper into that later as well.
And then finally, R is reusable. Reproducible is obviously quite similar. So in a sense, it is a bit of a tricky one that we put in, but it is reusable. And so it means whether it's easy for the data to be
recreated and understood. So yeah, that means whether the provenance is fully described. So really well done. Thank you very much. So going a bit.
deeper into FAIR, so a little history. So it's actually something that has come from the scientific community itself, so a bottom-up approach. So people came together because open data was becoming, well, more and more important, but obviously and we'll come to it in the next slides, it also had to be in a way that could be reused. So that's why these principles came into being a decade ago.
So a little bit of an explanation of the difference between FAIR and open. So open data means that it's publicly available, but it doesn't always mean that it's FAIR, because you could just have data that's sitting in a university repository without any information about how the data was collected, about what variables,

or what every column name means, for instance, in a table. So it might not be very easy to reuse. Whereas not an all FAIR data is not necessarily open. So you can have data that is FAIR, but for certain reasons, maybe it's sensitive data or commercial data.
It might be that it's behind a licence that you actually have to pay, or you need certain requirements to actually be able to access it, or you can only use it for certain purposes. And then also, I guess, importance in this day and age is AI ready data. So, well, you all know that large language models are making it easier for people to interact with it. I don't need these formal skills necessarily, because, well, you can just ask a natural language
question to a chat bot and it might come up with an answer. But obviously, the answer will only be as good as the data that it's supported by. And for this reason, it's very important that the data has been evaluated, validated, structured, governed and shared in ways that support this kind of responsible and effective use. And FAIR will get you partly there. I mean, there's other aspects to it that make it AI ready. But yeah, FAIR is definitely important in this aspect as well.
So, a little bit on the importance of FAIR.
So there are obviously some obligations. So first obligation is coming from funders. So if your research is funded by an institute or some type of body, they might have a data policy.
or they will likely have a data policy that will kind of influence the accessibility embargo periods and data repositories that you have to publish your data in. Then, might also be influenced by publishers. More and more publishers of journal papers are also now expecting researchers to publish their data alongside their journal article. And then finally, there's also legislation that will impact, yeah, FAIR, because certain data cannot be made available due to GDPR, for instance.
where, yeah, that's in Europe specifically, where sensitive or public data, public. Yeah, personal data cannot be made publicly available. And then freedom of information, like if data has been supported by public money, there is this kind of possibility for people to actually request your data.
and have access to it.
Then there's also obviously benefits. So it's not all just obligations. There's benefits, well, as a data producer, you can get credits and promotion of your work, especially for data authors. There are certain, like,
early career researchers or lab and field technicians might not be as involved in

writing papers, but they will be very involved in creating data and they might have opportunities to be lead authors. There's also more trust in your research because it's more transparent and you can also start building a network because people might come to you
because you're obviously putting out more kind of products of your research. It's also obviously beneficial for data users because they don't need to reproduce these data. It can be very expensive, but it's also easier to access and reuse other data. In the past, they might have had to actually
contact the.
data producer and then ask for the data and hope that it was well described or may have to ask a lot of questions to the data producer. But now it's just easy to get. Then also for the research community as a whole, because obviously it can build on existing science and it can also
create inspiration by doing this kind of open science, you can kind of inspire other scientists to do the same. And then finally, it's also beneficial for society because obviously there's better progress, but also the public has access to these data, which in the past they might not have.
And it's also transparent science, so they might have more trust in science as well.
And then we'll get to a little exercise. So how to achieve FAIR
So...
Yeah, how would you actually go about achieving FAIR? So we've got a Padlet here. Jasmine has just put it in the chat, but I'll just kind of start sharing that screen to explain.
So if you open the Padlets, let me know if you are not able to, then
We're asking you to kind of put your ideas of how you can actually make your data findable or accessible or interoperable or reusable. And you can add an idea by clicking plus and then type something. And you can also upvote someone else's idea or you can add a comment if you think, oh,
Actually, I would do it slightly different, or I would do this as well, if you think, yeah, so.
Yeah, go ahead, and
Get your ideas growing.
I will give.
Let's say, till...
Fifty-five.

Past 10, so that's about 7 to 8 minutes.
It's published data sets in a journal. That makes it obviously more findable because you're actually, well, hopefully also linking your data to
The ideal is that you actually publish your data as a separate resource and then link it to your journal paper. So if people find your journal paper, they can link it back to your data sets.
Open access publications, yeah, make it easier to find because obviously if it's not. visible anywhere there will not be a findable record.
It's also the database or the repository where you actually publish your data is also important as well, because, well, one where people actually are expected to find your data, but also, yeah, is it a reliable repository as well.
Um...
and well maintained.
So there's people saying, yeah, open access and not behind the paywall. So yes, ideally, if it's possible to publish, it's best to publish not behind the paywall, because obviously,
Yeah, as open as possible, makes it easier to reuse, but in some cases it's not possible. But then it should be clearly stated in a licence, which also makes it clear how it can be accessed.
includes metadata for reusability and deeds.
So it's clear kind of what every measure means, but also kind of spatial metadata, temporal metadata. There's loads of metadata or types of metadata that are very important for usability.
So, readable by anyone.
and talking about formats indeed. So yeah, if possible, use a format that is, yeah, one that everyone can use or you don't need special technical knowledge to use, but also that you don't need the licence for any software that
can only use this type of format. So that's very important. But again, here, if it's possible, you don't want to lose any kind of information in your data by transforming your
data to something more commonly used as well.
people making comments about GitHub and publishing their codes alongside the data set. It's definitely a good thing to do. Also for reusability, because obviously that makes it easier for people to understand how the data was actually produced and maybe make something similar.

as well in the future.
Full documentation for reusability, very important. And the reproducible scripts and data access.
Yeah, it comes back to, I guess, publishing your code, but also just making it that it's reproducible is also very important. Making sure it's commented well, that it's not using data that is not accessible.
International standards, yes, if there are standards.
for interoperability, then it's really good to use these because it will make it a lot easier for people to kind of integrate it in their workflows because they might also be using these same standards. And it will also be clearer actually what you are using. And.
I think people have stopped typing. So I'll go back to my slides. But yeah, that was all really good ideas. Thank you very much.
So just to summarise kind of the ideas that have already mostly been already proposed in the exercise. So findable, which generally means whether you're if you're actually looking for something specific, but discoverable is also important. So that's when you're not necessarily looking for something specific, but you might come across it kind of by accident. Then for this, you need a web accessible records and this should be in a well managed catalogue and have a unique identifier that makes it possible to link to these journal publications.
So it's easy to kind of find everything. I also have discipline specific keywords and descriptions.
Yeah, make it possible to actually find something.
And then accessible for this, you need the persistently working link and have no access constraints. So again, that's that no paywall, but or minimal and clear constraints if there is a need for this. For instance, if it's sensitive data or commercial data, and that has to be explained in a licence. And then if you have very large data or very complicated data, APIs and user interfaces may make it also a lot more accessible to, yeah, people with, yeah, not as much storage space or for people who are maybe not well-versed in formal analysis ways.
Then interoperable for this, we need accepted languages, formats and vocabularies. That was kind of mentioned as well. So using these standards that exists in your fields and formats that are non-proprietary. And because, well, it makes it obviously more widely accessible to people. And then also think about machine

readability as well, so it can be easily integrated in workflows. So this is, yeah, this is more possible with a lot of metadata that's integrated in your data as well.
And then finally, reusable. That is, yeah, can be achieved by having discipline specific vocabularies again. So if everyone is kind of on the same page, what everything means. Appropriate metadata standards.
supporting documentation, so it's fully described what you do or what you have done, and then a clear licence describing the terms of use as well.
Some key considerations when you're actually preparing your data. So think about these community standards. So is there anything in your kind of area that is important to kind of
take on. So for instance, are there technologies and techniques, methodologies or techniques used? So by clear descriptions of sampling methods and maybe pointing to these kind of existing techniques and methodologies, making use of control measurements, there may be standards
that are widely used. And then also think about uncertainty and error reporting. There may be ways that it's kind of classically done in your fields that may be useful to adopt as well.
And some useful kind of links to find these standards, because you may not be aware in your fields, but FAIR sharing repository has standards for different kind of subjects. And then the DCC has a list of metadata standards.
So yeah, kind of explaining how metadata should be structured so it can be easily kind of integrated with different systems.
It's also important to think about your data formats. So, and there's different things that you should consider when actually deciding which format you'll be using during your research. So, well, think about the intended use. Is it just you?
Is it government? Is it?
Yeah, for analysis or yeah, all these kind of things that you might want to consider. Then, yeah, are there specific analysis requirements? Maybe you want to do something that needs a specific software and does that software require specific formats? Then might want to think whether it's useful to have embedded metadata. So it makes it easier to kind of...
Yeah, integrate with existing workflows. A good example here is NetCDF files, which are, yeah, type of raster, but with a lot more embedded metadata in it. So it's clear, like, for instance, what units for each variable, some
like specific terms for each of the variables that might make it a lot easier to actually

integrate. So you don't need to hard code everything in. And then there's also, you need to obviously think whether you actually have the skills.
to handle these types of formats. It might be nice to have these kind of complicated formats, but if you don't have the skills and don't have the time to actually learn these skills, it might be a bit difficult. And obviously also the software, do you have a licence for this software? You might also want to consider whether
Do you want the version control kind of in the format? Or will you just kind of do it more, yeah, kind of yourself by writing away new versions? And then finally, if you're then intending to publish your data, can it be published in this format or would you have to transform it to a different format?
format. Again, for these kind of reasons that it might not be accessible to everyone or reusable by everyone.
Then also controlled vocabularies have been mentioned already in the standards. So it can be useful to use standard terms. So everyone again is on the same page and these exist for different fields. Or if they don't exist, it might be useful to kind of
at least within your project to use something that is like standards for, yeah, for everyone. So it's kind of with all the data files, it's kind of standards. An example, yeah, here examples are given for like different
Yeah, kind of.
concentrations of things. So they have like a specific string of letters that would denote something specific. And yeah, everyone kind of understands then that that means this. And so they can really easily kind of put it in their workflow.
And then obviously, it's important to think about reproducible data transformations and methods, which was also mentioned during the exercise. So make your code or your research workflow reproducible by
yeah, documents and what you're doing, but also, yeah, for instance, if you're writing codes, adding comments to the codes to make it more reproducible.
So then we're getting into actually publishing your data. So you're coming at the end of your project or at a stage where your data is final. So what do you do? So in the past, kind of traditionally, you would have a journal paper and then at the end, they might say, like, if you want,
this data, contact me. It's not always very useful because, well, e-mail addresses often change because obviously people change organizations. If you don't actually find these people, then they might not find the data or it might be, yeah, in the not very well described way.

So what the data publication is, is actually having a statement at the end pointing towards a different resource, which is this published data set, which has a lot of kind of these extra features with it, which we'll go into
Yeah, in the next sections as well.
And one thing to kind of think about here or what is important is these persistent identifiers. We've already mentioned that, yeah, when talking about findability that you need this kind of persistence identifier or unique identifier to make it possible to kind of
points to a resource and there are different examples. So you might have heard of DOIs, that's for journal papers, but actually also for data sets. You might have heard of ORCIDs, so that's pointing to specific researchers. And if you're planning to stay in research, it's really useful to actually
consider making an ORCID account because it, yeah, it makes it clear what the resources are made by you. But also organisations have unique identifiers, RORs, and then physical samples might also have.
a persistent identifier, so it makes it really clear how things are linked with each other.
And then we'll also go into why publish data. So maybe we'll ask here in the chat, maybe you have published data in the past, and then you can give here your reasons why you did that, or maybe you can
Think of like in the future, why would you publish data or why would someone publish data? So please, yeah, put your IDs in the chat or either unmute or put your hand up. That's also fine.
Yes, Robyn.

JR **Jones, Robyn (MECP)** 33:06
Hi, so I haven't really published data, but for work, I work for the provincial government, and we get a lot of like submissions from industrial dischargers, and they've gone out and collected their own data. And so I

JR **Jones, Robyn (MECP)** 33:25
I find one of the benefits of other people publishing their data online is that it allows us to kind of check what is being reported versus what others are seeing and can flag for us when there's like maybe issues with how the company's proceeding with their monitoring. So I do like that. And I think another benefit that

Jones, Robyn (MECP) 33:47
I don't know if you touched on it in your presentation, but I think there's like an environmental benefit because it's pretty resource extensive to go out and do monitoring. And so there's like also the environmental benefit of it. We can share the data.

Els Dhiedt 33:58
Yes.
Yes, yeah, there's definitely.
a positive that you don't have to redo, like it doesn't only cost time of people, but obviously it also has a lot of resources that you need to actually do these data collections. And yes, to compare your results. So I guess it's less of a
I guess, research kind of neat, but it's more of a kind of, yeah, does these kind of data that these people submit make sense that that's really, yeah, useful.
Other people are saying collaboration.
and contribute to scientific knowledge. I think, yeah, here again, you're not just keeping it to yourself, you're kind of making it open to everyone. Reproducibility, yes, make it possible for people to really make it clear that your way of doing is transparent and
can be reproduced.
Yeah, visibility and impact indeed. And that actually comes back to some examples that we'll give as well. But yeah, so why we do it is obviously because, well, it kind of leads to this open data, which leads to open science, which makes it possible to do Well, better research and more transparent research, and...
Yeah, we'll have here some case studies of why it's useful to publish data. So here is 1 is by Michael Tso at the UK Centre for Ecology & Hydrology has kind of published these soil moisture dry down kind of
paper and that was solely based on data that was published by others. So.
Without having the need to actually start from scratch and having to produce all these data, he was able to kind of take these resources, analyse them, and create this new
kind of resource. And then also another case study is maximising input, so kind of benefits for the producer themselves. So we have an example here is the Plynlimon field site, which is situated in Wales.

And there is a lot of hydrology and water quality data that had been collected already since 1968. And yeah, has loads of spatial and time series data sets. But only data from 2011 till present was actually available in
well, publicly available here in the Environmental Information Data Centre, hosted at UK Centre for Ecology & Hydrology. So in the past, people wouldn't be able to actually access these data unless they were collaborating or whether they ask for these data.
But recently, early last, well, late last year, these data were finally published. And yeah, that has already led to quite some impact for, well, the people involved in these data collection. So there's already a lot of views and downloads, but also it's
being reused within the Floods and Droughts Research Infrastructure.
programme, which is also partly kind of involved in this workshop. And then there's also, yeah, a seminar series that is going to be run for the Royal Geographical Society here in the UK. So there's some new collaborations
that came out of it. And then finally, there's also the idea to create a small grant scheme to make it able for people to reuse these data as well.
So we're now coming to the break, but first we'll ask whether there's any questions. At this points.
Then we can, but there's also time at the end, so...
If you can't think of any yet, but we'll think of any in the future, then we can come to it.
Yeah, feel free to put them in the chat or raise your hand.
If there are aren't any.
Yes, we can go into the break.
Which is 5 minutes, but I would say.
Till 20 past, does that make sense? Yeah.
Cool. So see you back in 6 minutes.

Jasmine Hunter 46:07

Okay, we'll probably get started in a moment. I'm just going to quickly check Els that you can see my slides all okay. Then we swapped over. Yeah, fantastic. So yeah, I'm going to take you guys through the second half of this and then we'll have some time at the end for questions as well. But we're really going to
dive deep into what actually publishing FAIR data looks like and how you do it. So that's starting with the what to publish.

And with that, as we've been doing before, we're going to start with a quick scenario poll. So I want you to imagine you're in this project, you're studying ground water level and soil moisture in the Wear catchment in the northeast of England. And as part of that project, you're collecting various data throughout. It's quite a long project, right? You're going to do a few different things. So the first thing you've collected is some ground level
oil moisture from sensors in the field, and you've stored that straight out of the sensor in the sensor specific format. That format requires a licence to be used. So to help your wider team use that, you've converted it into an aggregated CSV format, so you've got that as well.
You then also find, as you're searching for other datasets and related research, you find this freely available weather station data and satellite data
that's available for download, that has a licence that's available for you to use. So you combine that with your ground truth field data to create a model that predicts groundwater level soil moisture. Now that model only takes 5 minutes to run. It's very quick, so you can do it while it doesn't require much computer processing power.
What we want you guys to think about is, out of these, which ones are you think of long-term value and which ones would you actually publish?
Els is going to put a poll up. You can click multiple options. There is also a none of the above and all of the above.
But have a go and then we'll discuss.
And I should say I've clicked none of the above just so I can see the results. So that one's me.
Alright, I'll give you guys another 30 seconds or so. Still have a few people to click submit.
No problem, Nikola. If you want to add some ideas into the chat, feel free. Otherwise, that's all fine.
Okay, well, we might call it there. So if we start from number one, no one said publish the straight from the sensor. You'd be correct there because it's in a non-proprietary format. You need a licence. So it's just not that accessible. Whereas #2, the accessible one that's open on any software is a lot more interoperable, right?
So that's the one you'd want to publish. Looking at #3, a couple of people said yes. In this case, we would say, no, don't publish this one because someone else has already published it and you're not changing anything about that data set, right? So

it's already available. You don't want to just copy that one and read.
Publish that.
Looking at #4, yep, a few of you said yes for the model code. Definitely, it's really good to publish a model code if you can. The best way to do this at the moment, I think, is on GitHub and then version it using a DOI via Zenodo. So that gives you that persistent link and citation for your code and sets it at a point where it's like, this is a finished product.
versus like a work in progress type thing. And you can do versioning with that as well. Number 5 is probably the trickiest one, and it really depends the answer here. So in this case, we kind of said no, because it's really, really quick to run. It's not that, you know, like someone, everything else is already available.
And you can just recreate that data set by running the model with the input data. But if it say it was a... it took a high performance computing or quite a lot of work to actually run that model, you would probably publish an output so that people could just view that and it's available to more people.
But yeah, thanks for doing that. So this is kind of the...
The quick checklist, if you like, for data of long-term value, it obviously does depend quite a lot on your specific data, but essentially not everything. If it can be really easily remade, it's not of long-term value.
Um...
So data you have generated is really good. So this can incorporate third party data in the creation of it, as long as you've checked the licence of that source data and it allows you to reuse that.
Raw vs processed is not straightforward. Sometimes the raw, you know, is more useful. You want to work out what the people in your community are going to find most reusable. And often that is processed and often it's a snapshot, right? You want a complete data set. You don't want to be publishing just random days throughout a year.
you want like a whole annual snapshot if you're doing long-term monitoring. Partly because it saves you work in, you know, you can publish it in one go, but also doing it quick enough that you're not just hoarding all this data and waiting 10 years to publish.
So now we'll look at where to publish. So you might have heard this term throughout this session called a data repository. It can also be called a data archive. It's just a centralised place where data is stored, organized, and made available for you. So this

can be physical. If you're thinking, you know, like soil cores, they're stored in a archive, samples, that kind of thing, or it can be online and digital for digital data. The purpose of most data repositories is mainly to share data, but also to preserve it in the long term and archive it for future use. It can also be used for data management and data analysis, depending on the organization.
There's quite a few around and there's different types of repositories. So you can search for repositories in your field using the commonsdatasite.org. And then it just, you know, you kind of have to go through and choose which one suits your research the best. Things to consider in that are firstly, your funder data policy requirements. For example, if you're NERC funded, you might need to publish your data in a NERC repository.
Other ones might have a requirement that you just publish it openly, for example. Does it meet FAIR is a good thing to consider, and partly that will be you having to supply that information to make it FAIR, but also that they'll give you a DOI and that kind of thing. Certifications can be a good thing to keep an eye out for. One good one is the Core Trust seal.
That's kind of an international certification for some repositories. And then finally, your intended audience. So are you looking for like a general repository that's open to everyone, or are you targeting a specific field that maybe you know everyone always goes to access?
So as I said, yeah, there's these types. There's the general repositories. This is your Zenodo, your Dryad, your Figshare. They have different levels of offering. Some don't require you to add any metadata. Other ones do. So consider that. And then you have specialist repositories. So one's focused on a specific field or data type.
A key example is the UK Data Service, which specialises in social data sets, or the World Glacier Monitoring Service.
We also have the NERC Environmental Data Service Repositories. So Els and I are part of the EIDC, the Environmental Information Data Centre, which is one of five NERC repositories. We focus on terrestrial and freshwater data, so everything in that remit that's funded by NERC comes to us. You also have CEDA,
who do atmospheric and earth observation data sets, BODC who do marine data, NGDC who do geoscience and geological data sets, and then the UK PDC who do polar and cryosphere data. You also can, there's a NERC EDS catalogue, but you can cross-reference all of these in one go if you're searching for data.
In one of those remits.

Okay, we're going to move on now to supporting documentation and metadata. This is really key in the reusable. And again, we're going to start with an exercise. This is going to be quite quick fire. So, you know, fingers at the ready on your mouse. We're just going to look at are these data practices FAIR or not FAIR.
So Els is going to put up a poll. The first one is you do not produce a separate document to publish alongside your data because your methods are described in the paper. FAIR, not FAIR or not sure?
All right, 4 responses, 100% not FAIR so far. Does anyone want to have a go at explaining why, either in the chat or putting your hand up?
Yep, Sam.

ST **Sam Tulloch (PGR)** 56:26
I'd say you'd want a readme file or something similar with the data because it's separated from the paper. So you want something that says, oh, by the way, here's what it is with the data specifically.

 **Jasmine Hunter** 56:29
Huh.
Yep, perfect. And Sebastian, anything to add?

YS **Yelesen, Sebastian** 56:45
Like in my experience, I've actually tried to reproduce somebody's results before, and you find that the method section often gets kind of distilled down for this, for this, you know, because you'll get people reading your paper and be like, this is too long. But then you'll be thinking, okay, how can somebody actually reproduce the results though?

 **Jasmine Hunter** 56:55
Uh-huh.

YS **Yelesen, Sebastian** 57:05
So obviously, when you try and work off of the method section, that's often not enough. It's often things get left out and left on the cutting room floor, which is why it's important to have something else for people that are actually trying to piece together what you've done so that they can reproduce the results.

Jasmine Hunter 57:19

Mhm.

Yeah, yeah, exactly. So we said pretty all the same thing. We said it was bad for reusability because you only have so much space in your paper, right? You're going to miss out on important details and you also want it located with the data set rather than over here somewhere in a random spot. Okay, next one, Els.

You publish your data with a licence that does not permit commercial use, for example, CC-BY-NC, as that was required by your funder.

Okay, so we've got 7 responses, all yes at the moment. Does anyone want to share why or why not?

No, that's fine. We can carry on. So we've also said this is good for usability. It's FAIR practice. The licence clarifies to the person looking at your data set how they can use it. So you're setting those conditions. It's a bonus that you've met your funder requirements. Key thing here, you remember at the very start we talked about FAIR versus open.

This data set would not be completely open because there's no commercial use, but it's still FAIR because you've set out those conditions in a clear way.

Yep, exactly. They funded the work, so you've agreed to that.

before you've started.

Okay, number 3, your data contains detailed farm location information, so you publish only the aggregated summary data.

Okay, so far we've got some unsures and some FAIR practice and one not FAIR practice. Does anyone want to have a go at saying why or why not in the chat or talking?

Yeah, in the chat, we've got Mayuran, detailed farm location information is sensitive.

Yep, yeah, definitely. So there might be legal reasons why you can't share that information.

And we said the same thing. So we said this is FAIR practice. It's good for accessibility and reusability because there's valid reasons, in this case legal, to protect the sensitive information. So they've shared what they can and they've described why they're doing that and what they've shared. So that data set is FAIR on its own.

Again, not completely open, but it's FAIR because there's valid reasons to do so.

Okay, next one. Only use new words developed in your project as the keywords, such as the name of your newly developed sensor and the project acronym.

Five responses so far.
Give it another 10 seconds.
And that one, everyone has said not FAIR. Does anyone want to say why not?
Yeah, Robyn.

Jones, Robyn (MECP) 1:01:34
One I feel like it's just not nice to use, make up all your own words, but also because then you can't really compare it to other work and it's not like, I guess, accessible then.

Jasmine Hunter 1:01:43
Mm.
Yep, yeah, definitely. And I can see a few people in the chat saying people aren't going to know what they are. And if they don't know what they are, they're not going to search for them. They're not going to find your data.
So yeah, we said the same thing. It's really important to use relevant keywords that are widely used. And that's where those controlled vocabularies can come in handy. But also just looking at other data sets and what other people have used, if you're unsure, can be helpful. Okay, and I think this is the last one. So you add in a data availability statement to your paper.
Telling people to e-mail you if they want access to the data.
FAIR, not FAIR, I'm sure.
Okay, it's a bit of a mixture for this one. Does anyone want to say why they chose what they did?
Yep, we'll go Sebastian first, I think.

Yelesen, Sebastian 1:03:03
I've actually emailed people looking for the data and yeah, I put unsure because I've got to, I'm going to be honest, they never, they never, they hardly ever write back. I think one of them I'm still waiting for two years, two years from now. So
Yeah, it's a bit hit or miss in terms of actually getting access to the data.

Jasmine Hunter 1:03:27
Yeah, it's a lot harder and Sam.

Sam Tulloch (PGR) 1:03:30
Yeah, ditto pretty much everything Sebastian said, as well as, like, if...
I hate checking my emails. I'd never let any, like, I'd rather just put it up somewhere and go find, like, it's here. Leave me alone. So that's my personal thoughts on it is like, I don't like reading my emails. I'm not going to give people an excuse to continue to e-mail me.

Jasmine Hunter 1:03:53
Yeah, you're just you're setting yourself up for extra work, I think, um, and Tlakale said as well, e-mail addresses change, 100%.
Yeah, so we said bad for accessibility, reusability, findability. Data should be published with a DOI and have all the metadata with it, because it's a lot of work just getting random emails, especially if you haven't made those things. It is important to say so that data availability statements are really good practice to include in your research paper.
especially if you've already published your data set. You can put the link straight in, people can click on it and access it. If you're worried about people using your data before you publish, you can add embargoes on your data set as well. So if anyone has questions about that, ask us at the end on embargoes.
Okay, that's it for the quickfire questions. We're going to look a bit more into supporting documentation. So supporting docs are just the documents containing the information that allow your data to be understood. Whereas metadata can be more than that, right? It's the data about the data. So it's structured information that give context, that allow viewers to use it, interpret it, and reuse it. So this doesn't have to be in a document. It could also be in a README table. You know, it could just have a table of the variables and that kind of thing, or the spatial location, all that stuff.
Why do we create it? Firstly, to not forget what you did or why you did it. I don't know about you guys, but if I look back like four years ago to a random lab day, what control did I use? I could not tell you. If I've written it down, I could find that info straight away or with maybe a bit of searching, but nearly straight away.
Secondly, it helps researchers interpret the data so they can use it straight away, and also to find and retrieve the data. So if someone's looking at a specific region, they can search for that region. And because you've included that in your metadata, the

spatial location, it'll pop up in their search.
It also ensures proper attribution of the output, so credit to you, people can use the citation, complies with data sharing policies and avoid misinterpretation. You'll see a lot of these are the same as the benefits of publishing in general and the benefits of FAIR. They all kind of link in with each other.
Okay, so we have another poll
so you found the perfect third-party data set, what information do you need to reuse it effectively? And I've got some hints here. So this is just in the chat or by speaking out. So what types of metadata do we need?
Location? Yep, definitely. Latitude, longitude is always really good.
Units, yep, otherwise people aren't going to know what your data is. It'll just be a bunch of random numbers.
Time step, yeah, what resolution for sure, and when as well. So how frequently are you collecting data?
The date, yep, for sure.
Well labelled. Columns, definitions, yep, definitely. If you're not sure what a column is, you can't actually use it, right? And now we're getting into the how as well, yeah, what processing was done. For example, did you convert it into a new format or something?
Did you need to calibrate it to account for your instrument that you used?
Give 30 more seconds. Any final ones? I think we've done most of these. There's a couple missing.
Yep, scale factors for sure, and the source, definitely.
And.
The person with a question mark, I think we haven't done. And the stars with the checklist. Yep, who collected or what instrument? So what collected? Definitely.
All right, we'll call it there. So first thing we want, I actually don't think we did mention this, but the data itself, you do want that. So those files should be whole and complete. They should be reusable format, and they should be QA'd to ensure the accuracy and integrity. And you should include those QA steps.
quality assurance steps in your supporting information as well. So that just helps people, you know, feel a bit more confident about your data. We then have descriptive metadata. This is the who, the what, the where, the how, the why, all the kind of stuff about how that was collected and where.
Next, we have structural metadata. So this is more about how to use the data set. I

think people mentioned the format, the units, what the columns are. It can also be if you have multiple files in your data set, it could be in a folder structure. So just telling people how to access things.
If you need specific software, even if it's open access, you can include instructions on that if you want to.
We then have administrative metadata. So this is how it can be used. So this is where you'd have your licensing, any funding acknowledgements, you know, authors, institutions, who to contact if people want to get involved, because that might spur some new collaborations, which is always good.
And then we have relationship metadata. So this is how it links with the other data sets, code, your research paper. You might also include versioning if the work you're doing has multiple versions, as well as linking to any source data sets. Particularly important to do that
if those data sets require you to acknowledge their use. So just make sure you check that. You can include their citation on the data set record page.
Okay, we now have tools and resources. These are just some extra things. And the first one is us, it's ourselves. So you get to that. But the EIDC has a bunch of guidance on data structure, formatting and metadata contents. You can see here on the main page, it can tell you different acceptable formats that are non-proprietary.
How to do your variables, what to do for missing data, and how to organise your tables in a FAIR way.
You also have some information about metadata guidance, what to title your data sets, really important for findability. Make sure you're talking about the type of data set and not your research. And finally, we have a supporting documentation, and this includes a downloadable supporting document template. You don't have to deposit with us to use this template.
Please go download it and use it for whatever you like. One thing I'd recommend is to download it now and while you're doing your research, just add in stuff as you go so then when you get to the end, you already have a skeleton of what your supporting information would look like.
Some external resources. So the Turing Way is a UK kind of community-based project through the Alan Turing Institute. They have a Turing Way book that's all about research data management and just transparent and ethical data science. So that's one good tool.
We also have the Elixir RDMkit. So Elixir is a network, I think, across Europe and the

UK in the life sciences domain, but a lot of their tools and resources are available. They're applicable across different fields. So I'd recommend having a look. They have stuff on the data life cycle on specific
tasks that you're doing. So for example, if you're looking to learn how to version control in Git, they might have a page on that. And they also have a FAIR cookbook that's kind of like recipes for data management and FAIR resources. So that's quite a good read as well. And finally, we have the ARDC. This is the Australian Research Data Commons.
They have a whole bunch of resources for FAIR data. I'd particularly recommend the FAIR self-assessment checklist, so you can answer some questions and just check how FAIR your data set is before you publish it. They also have a free self-paced FAIR 101 course that's good to use. And there's also a section
on FAIR software that I think is really good. If you're working with code and that kind of thing, it's a good thing to consider.
And that's the end. So thank you all for listening to us talk and getting involved with all the polls and everything. It makes it a lot more fun. We've got, I think, 13 minutes or so. If anyone has any extra questions they'd like to ask us, we can have a discussion on anything.
And I'll...
Stop sharing my screen.
If we don't have any questions, would anyone like to share maybe something they've learnt in this workshop or in any of the workshops that's kind of stuck with them a bit? Something you're going to take away?
That's always nice to know what which bits, you know, are really helpful.
Yes, Sebastian.

YS **Yelesen, Sebastian** 1:14:17
I thought I'd say, I thought the FAIR one...when I was first reading the itinerary for this for this summer school a couple of months ago, I think it was the FAIR data use was one of the areas I was more interested in.
Because, um, yeah, I've just kind of noticed actually how...
And how difficult it is to actually, like, obviously.
like how difficult it is to have well, like fully reproducible, fully traceable and like, you know, FAIR data use whenever you're doing any sort of any sort of research and like how difficult it is. Like, I'll read papers done by very

big, you know, big institutions like the Met Office and still think like, wow, like I would not, I would not be able to reproduce their method at all. They haven't like, they haven't like, like there's so many areas where it's not been explained how something's been done. And I was just thinking like, OK, how can I

Yelesen, Sebastian 1:15:26
like, how can I improve my own writing to make it clear without also making my writing like so detailed, dense and incomprehensible, just because they're too, they're too, like they're too
opposing forces. So it's tough. But yeah, I was going to ask, do you think reproducibility is like getting worse or getting better in like, you know, hydrology research or meteorology?

Jasmine Hunter 1:16:01
I would say it's definitely getting better. I think there's still work to do, but I think you're finding the fact that people are hearing about reproducibility as a requirement is quite a change compared to say 20 years ago. Like it's becoming a lot more mainstream, if you like. I think particularly in earlier career researchers, they're learning about that from the start.

Yelesen, Sebastian 1:16:15
Yeah.

Jasmine Hunter 1:16:25
rather than needing to like retrofit what they've previously done into these. But yeah, there's definitely work to do. And I think particularly in transparency and around, you know, AI coming in, we've really got to keep logging what we're doing so that other people know how something was made.

Yelesen, Sebastian 1:16:30
Yeah.
Thank you.

Jasmine Hunter 1:16:46
No problem. Els, did you have any extra points on that?

Els Dhiedt 1:16:51
I think that kind of covers it from what I kind of see. Indeed, it's getting better. It's obviously, in certain fields, like within research, it's more possible, but then obviously industry, they have obviously other things that they need to kind of take into consideration. So it's less of a kind of active practice there.

Jasmine Hunter 1:17:02
Mm.

Els Dhiedt 1:17:12
Um, but yeah, definitely getting better, I think.

Jasmine Hunter 1:17:16
Yeah, yeah, 100%. I think, yeah, your point on that, like, the work involved is a very valid one. We only have so much time. So I think just doing as much as you can and really organising it in a way that makes it as easy as possible is important. But also, I think, yeah, some people starting to do this will make other people then... oh, I need to do this as well. So I think it's kind of a broad collective.

Yelesen, Sebastian 1:17:43
Yeah.

Jasmine Hunter 1:17:43
I'm just having a look at the chat. Yep, thanks Martha. Interesting to know. Yeah, data archives could be physical. So it's probably because they haven't been called data archives. They would have been called, you know, physical sample archives or something. But you can have all sorts of all sorts of data.
Producing, publishing, sharing.
Is there such a thing as ensuring outputs are FAIR as well? There is some, there's some recent work that's looking actually at making training materials FAIR, so that's quite interesting, as well as FAIR software. I can't remember what the acronym for FAIR software is, but there has been one made. I'm not sure if that is for papers. Don't think so. Or for like, yeah, written, written things.
Appreciate the links. Thanks, Robyn. That's yeah.

Amulya Chevuturi 1:18:35
Sorry.
Just to jump in, in the sense, you're right. I just wanted to add about the papers being FAIR. Technically, the peer review process should take care of that, ideally. But yeah, if it doesn't work, then yeah, we won't have the FAIR principles for publications.

Jasmine Hunter 1:18:38
Yeah, that's a good point actually. And the peer review is kind of, it's such an inherent part of science, I think, and starting to get the peer review data sets as well is really good.
And.
Yep, brilliant. Thanks, Adi.
FAIR training, yeah, because there's a lot of reuse of training materials as well, rather than having to remake them, it's always very good.
Any other thoughts, questions?
No, that's all good. We'll probably end it there then. You'll get a copy. Well, there is already a copy of the slides. We'll share a copy of the Padlet exercise that you guys did in case you want to look back at that, as well as a list of all the links. And I think, Amulya, I know you're going to share the questions asked. Is that right?

Amulya Chevuturi 1:20:06
Yes, yes, the links that are posted on the chat and the questions asked and answers, I have tried to take in as much as possible. And Gianni will be sharing them post day workshop kind of thing. But first of all, we would like to thank both Els and Jasmine for delivering such an awesome.
workshop. So please, big round of applause for all of them. And thank you all the participants also for giving all the attention. Avintha, you have your hand raised. Is that a hand raised or is that a clap?

Avintha 1:20:47
Sorry, it's a clap.

Amulya Chevuturi 1:20:50
Cool. But thank you once again. Anything else you would like to say?

Before we round off, I think the next session, now we have a lunch break for an hour, and the next session will start at 1pm UK time. That is in an hour and 5 minutes. So one hour and 5 minutes later, we'll see you all again.
Thank you once again, Els and Jasmine, for a wonderful workshop. And I will see the rest of the participants in an hour. Bye. Have a good lunch. Thank you.

Jasmine Hunter 1:21:27
Thanks, everyone. Enjoy the rest.

Els Dhiedt 1:21:28
Thank you.

YS

Yelesen, Sebastian 1:21:28
Thank you.

Amulya Chevuturi stopped transcription