Speaker: Joy Davidson

Video Title: Research Data Management

**Transcript:**

**Slide 1: Introduction slide - video summary (0:07 - 0:38)**

So as I said this first quick session is really just going to give you some key messages about FAIR, open data and research data management, because they can be confusing concepts to researchers, so we want to give you a bit of clarity on what you can share with your researchers if they have questions about this. So we'll give you that understanding of where they come together and where they're different. A couple of slides just to give you some drivers for FAIR and why it matters and some practical tips when you're talking to researchers about how you can encourage them to engage with FAIR.

**Slide 2: What is Open Data (0:38 - 1:19)**

So first of all what is open data? The ODI gives this nice definition and essentially open data is just that; it's data that is licensed in such a way that anybody can take it and do whatever they like with it. That may also mean that they might use it for commercial purposes. So in many senses an open data license can be seen sort of as a waived license as well. So it's the most open, the most permissive form of data sharing. And it's not ideal for everyone but if you can move towards that sort of an open data approach it certainly is you know the ideal that we'd all like to be working towards.

**Slide 3: What is FAIR? (1:19 - 3:16)**

In terms of FAIR data; I think you can see from the screen that the FAIR data principles are nice and they make a nice acronym findable, accessible, interoperable and reusable, is what FAIR stands for. And it is a very nice acronym and it sounds very simple. And it's you know something that most people can get behind at that high level, so I don't think there's very many people, whether you're a funder, whether you're working at an organization that does research or whether you're a researcher, who could really argue with the notion that data should be FAIR.

I think when you start to look at the points underneath each of those headers; so under findable there's you know four separate bullets and similar under each of the other words, you can start to see that it actually is a little bit more complicated when you start to think about doing it in practice. So while most people can get behind FAIR as a concept and are willing to to work towards being FAIR, it does take that kind of extra work a little bit to try and do it in a practical sense. And as you can see from the points here under each of the words findable, accessible, interoperable and reusable, it doesn't just apply to the data sets that are coming out of research but more importantly is about the metadata so the description of the data sets as well. So i think for FAIR to really be something that is practically implementable, a lot of it comes down to the use of good metadata. So that we understand the data and we can find it and use it, and the use of persistent identifiers. So when you start to look it in that sort of more technical side, you can start to get a sense of why researchers might find it a little bit more tricky to put it into practice. So hopefully after today you'll have a few kind of tools in your toolkit that you can start to help people to understand what it means in a practical sense.

**Slide 4: Open and FAIR data are not the same thing! (3:16 - 4:00)**

So those are the kind of definitions for the three often confused concepts. So this slide I think this is quite useful for sharing with your researchers because it does help you to see that open and FAIR are not the same thing. You can have FAIR data that is closed and you can have open data that's not FAIR. So I think it's it's good to get your researchers to understand that there are grades in fairness and grades and openness. The most value will come when you've got FAIR and open data, but I think it's important to remember that FAIR data can be closed and we'll come on to this a little bit later. So if you've got sensitive data you can still make it FAIR even if it can't be shared widely or openly.

**Slide 5: How do Open, FAIR data and RDM intersect? (4:00 - 5:11)**

This is another nice diagram to kind of think about the concepts and how they intersect. So open data is primarily what people would assume is for the benefit of others. So your research is available for anybody to make use of and it benefits the community more broadly. FAIR data sits sort of in the middle between open and managed data. Managed data is in everybody's best interest. Even if a researcher isn't planning to share their data with anyone, it's still very good that they should be managing it well, just to be keeping in line with research integrity and to stand up to scrutiny if any questions come back to them about their published results. But you can see that FAIR sits between the two, so you would hope that FAIR data is helping you to be well managed and to understand (even if it's just you) how to find, access, interoperate with and reuse your own data. But moving towards that kind of open as well. So basically it’s a nice diagram to show you how they sit together and where they fit together, in that whole research data management life cycle.

**Slide 6: Funder expectations (5:11 - 6:18)**

So some of the reasons for aligning with FAIR primarily have come from the funders. Over the last few years many research funders, not just in Europe but across the globe, have started to include references to FAIR and to data sharing and their policies. So there is an expectation: This is an example from Horizon Europe; there will be FAIR data produced as a result of getting public money to do research. And I think many funders are looking at this as a public good and they are looking to make sure that as much of the research that they fund, can be as openly shared as possible. I think the Horizon 2020 mandate ‘as open as possible but as close as necessary’ is being translated through into Horizon Europe as well. So again that is a really important thing when you're talking to researchers; to make sure that they understand that FAIR doesn't mean open and it's up to them to make sure what they can share and what needs to be protected and how they'll go about doing that.

**Slide 7: EOSC vision needs FAIR data! (6:18 - 7:30)**

And again I mentioned the European Open Science Cloud (EOSC). This is a really huge infrastructure development program at the European Commission level. It’s aiming to bring together all the different stakeholders, different sorts of infrastructures, funders, research performing organizations, people contributing to and using research infrastructure to produce and share data. I think it's been since 2016 that this has been getting invested in and we're starting to move more towards the implementations side now. So it's definitely still a key ambition for Horizon Europe and there's many projects that will be coming to improve the European Open Science Cloud. But essentially for this whole trusted virtual federated environment we need to have FAIR data so that people can find the research that they want to reuse and then to be able to um share their own data, so that it becomes this kind of self-fulfilling environment where people can find what they need when they want to use it and it just works.

**Slide 8: Make use of a data repository (7:30 - 8:56)**

I think the first, and probably one of the best, things you can tell your researchers to do to help support the fairness of their data is to make sure that they use the preferred repository at the end of their project. Preferably one that is FAIR aligned. And in that respect when we're talking about preferred the the first thing you should do is always go for a domain specific repository if there is one for that community of practice. It's important to use the discipline specific resources, primarily because they will do a lot to make sure that you're using domain-specific standards in relation to metadata and you'll get a persistent identifier applied to the data. And it does solve that sort of ability to make your data visible to aggregators and to allow the access and to allow reuse.

So we would always say domain specific is your first choice and if you don't have one of those then institutional repositories and we've given an example here from the university of cambridge; they have the apollo service. Or a generalist service and zenodo is something i think most researchers are familiar with and it is a good option if there's no institutional or domain specific repository to use. So that is the number one recommendation we would give to you is to make sure that the researchers do make sure they put their data into a repository.

**Slide 9: Finding Repositories (8:56 - 10:11)**

To find repositories (and there's loads of them out there) i think if you're in the role of a data steward and you're trying to help a researcher to find a suitable repository to use; we would recommend using something like a registry. Re3data is probably one that you've maybe used before. It's funded by DFG and i think the key thing that you can benefit from the re3data search functionality is that you can start to look at what sorts of things you want from your repositories. So if there are certain types of access, you know you might want some period of embargo, you can check to see what sort of repositories will support that; the kinds of licenses they use, whether they offer different kinds of persistent identifiers and certificates and standards. So re3data will also allow you to look for certified repositories, so repositories that might have core trust seal certification or similar. So that's a really good tool to help you to look for repositories by country, by domain and by some of these other filters and that can be a really useful tool when you're talking to researchers at the data management planning stage.

**Slide 10: Descriptions for humans and machines (10:11 - 11:19)**

The next thing that is really important is to work with researchers when they're maybe thinking (at the data management planning stage) about what sort of metadata they should be using. And again if you're using a domain specific repository they will already be using specific standards and that can help you to know which you should be recommending to researchers, if you're planning to use a specific repository. The key thing to remember with standards is that they should be readable by both humans and machines. So they should be structured and in a standard format. The ability to have machine readability just means that they can be found by aggregators and made more visible but we also want to ensure that they're human readable as well. I think finding the right standard is tricky and there's lots of different things that can help you to do it, so RDA and DCC offer some good things that you can use to find domain specific standards. But again have a look at the repository, if there's one specific repository you're planning to use, then check to see what standards they support and that can save you some time.

**Slide 11: Metadata and interoperability (11:19 - 12:20)**

Another thing that you can talk to your researchers about is the use of controlled vocabularies and ontologies. I think this is an example more from the life sciences sort of approach. So controlled vocabularies on the left side of the screen, you can see there are many ways to say humans and as you can see in that slide; so homo sapiens. Very different kinds of things. And I think you just want to have the controlled vocabulary and encouraging researchers to make use of controlled vocabularies can reduce ambiguity over what different terms mean in data sets. And ontologies are another good example of things to use. It just helps you to have defined terms and hierarchical ways of understanding how they fit together and we've included a few links here to things that you can use to to help you to find it; so GO FAIR offers some nice tips that you can use to help support researchers to find vocabularies and ontologies as well.

**Slide 12: Use identifiers and metadata to link to related outputs (12:20 - 14:00)**

The most important thing i think for researchers to really benefit from making their data FAIR, is to try and make sure that they're using persistent identifiers for as many of the different outputs that they're creating. And that just enables these kinds of links and i'm not sure if many of you have heard of things like research graphs or persistent identifier graphs, but this is new technology that is allowing different outputs to be linked together through the use of persistent identifiers like DOI's. Essentially you can have something like the Orchid which is a researcher's DOI, it's an unambiguous way of researchers being able to say who they are, if they then include that then you can understand how you can start to move between the researchers published papers which will have a DOI, data sets in a repository that also have a DOI. You can start to make up these links between the various different outputs. And that's where you get real benefits in understanding providence and how something came to be. And i think it's also important to consider is not just papers and data sets but also things like software. If the researchers have created software and maybe deposit it in Github, then you also want to include that the identifiers for that so you can start to build up that picture. So it's really good to get researchers thinking about providing these links and to help them to understand that there's benefit for them in that, because it really does drive um understandability of the data and that then generally helps to improve your chances of somebody picking it up, reusing it and citing it.

**Slide 13: FAIR Closed data (14:00 - 15:03)**

So I mentioned before that there's a difference between FAIR and open. So for those researchers who will all of a sudden say: ‘I can't share my data - I can't do FAIR’ I think it's important to make sure that they understand that even if the data can't be shared it can still be FAIR. It's it's good to have it well managed and to use standards and all of the rest of it even if it isn't going to be shared widely. But there is in many ways, you can start to think about how you might share selectively if you've got sensitive data. And we're seeing the advent of data vaults and safe havens that are coming up at the organizational level but also at the discipline specific level. So even if you can't share it widely, think about how you can share it in a certain context and how you might go about doing it. So it's good to get researchers to think about how they might share and to what extent at the data management planning stage. And if they do have these needs, thinking about it then, how you would find a suitable repository to meet those needs?

**Slide 14: Consider - not all data needs to be kept forever (15:03 - 16:52)**

I think the other big misconception we find with researchers is that, there's a misconception that: all data has to be kept forever. And that's absolutely not what any funder is saying, nor any research performing organization policy, as far as I know. There is definitely a need to put the effort into those data that you want to make FAIR because it does take a bit of time and effort. At a basic minimum we would suggest that any data that supports published findings should be retained and that should support research integrity. There may be a case to make for keeping more data than that because it might have value or potential reuse value. But again you might need to consider that in a cost benefits analysis sort of approach to see if you can afford to keep it, do the costs outweigh the benefits? But I think the important thing to remember is even if you know you've identified a certain number of data sets you want to make FAIR and you've done the work to FAIRify them and they're in a repository, you might have a retention period of five years. At the end of that period you might get rid of the data in line with any kind of retention scheduling that you had. But the metadata about that data should still be in the repository. And I think it's just that is a really nice thing to have so you don't get dead ends. So if you've found a publication and it references a data set that has rightly been destroyed after a certain retention period, you don't want to come to a dead end. So having that metadata record there just gives you that understanding that you know what happened to the data, so even if the the data no longer exists the metadata should be.

**Slide 15: Seems like extra work! Why should researchers care? (16:52 - 17:04)**

I think when you do talk to researchers, you might find a lot of people start to think ‘this does sound like an awful lot of extra work’ and ‘why should i bother to do it’.

**Slide 16: Focus on the benefits (17:04 - 18:34)**

And I think when you're talking to them it's always better to focus on the benefits rather than - ‘you need to do it because your funder says you need to do it’. There's lots of good reasons to manage your data well; it helps you prevent data loss. There's been studies showing that if you share your data along with publications you get more citations. If you've got well-managed research you can share negative results. So if you did your research and didn't get the answer you were expecting, if it's well managed and you understand you know that you've got something that could be useful for someone else, there's more journals now that are coming up to share negative results. So i think it gives you more opportunities to publish and to get credit for your work. It certainly makes reporting easier. So we all have to spend a lot of our life (when we're working on these kind of funded projects) doing reporting, well-managed data and FAIR data can make that a lot easier. And as you'll hear on wednesday, it does help to support research and integrity and the validation of results. But also I think as we've seen with COVID 19, that's been the best case study we could have asked for. I think people see in a real world setting how collaboration and sharing of data in a quick and timely fashion is really leading to changes that’s helping us all. So I think it's really good to focus on these benefits so that researchers don't just do this as a box ticking exercise, but really engage, because it does lead to actual benefits for them and for society.