Blog post: https://www.esipfed.org/esip-interviews/making-data-matter-with-matt-mayernik

Blog title: Making Data Matter with Matt Mayernik

Interviewee: Matt Mayernik, National Center for Atmospheric Research (NCAR)
Interviewer: Arika Virapongse
Date of interview: July 18, 2018

Blog preview / highlight: **“**One motivation for open data policies is to help science, and another is to ensure that science helps society and the economy.” - Matt Mayernik

Arika: Could you tell me about how you got started working in the field of data and informatics, particularly as it pertains to Earth Science?

Matt: I did my Masters and PhD (completed in 2011) at UCLA (University of California, Los Angeles) in the Department of Information Studies, which is the re-named library school that goes back 50+ years. My dissertation project was focused on scientific data and metadata--it was a mixed methods study looking at how people collect and manage data, and create metadata for themselves, as well as thinking about the bigger picture in regards to how they also share data with other people. The dissertation focused on data sharing, and specifically the metadata components of that. Most of my focus at that time was on environmental science, such as marine biology and geology. In any event, I didn’t get too far into Earth Sciences until I went to work at the NCAR (National Center for Atmospheric Research) library in 2011, and I’ve been there since then.

At the NCAR library, I focus on data management from a library/information science perspective. I lead research and service development projects related to data curation, and collaborate with many other data management people at NCAR, many of whom also come to ESIP (Earth Science Information Partners) meetings. Until I got to NCAR, I didn’t really realize the institutional scope of the Earth Science and Informatics community. Of course I knew what NASA (National Aeronautics and Space Administration) and NOAA (National Oceanic and Atmospheric Administration) were, but you don’t really understand the data aspects of these agencies until you really get into this community. When I got to NCAR, the world really opened up. I found that there was this whole community of people studying and building systems to support Earth Science data.

I met Carol Meyer--the previous ESIP executive director--at a data librarians meeting (RDAP-Research Data Access Preservation) in 2011 or 2012. Around that same time, she was invited to give a talk at NCAR about ESIP, and NCAR became an ESIP member. So that was how I started getting into the ESIP world.

Often graduate degree research is somewhat isolated and very focused. Is this why you didn't come across the Earth Science data world much during your graduate school studies?

I was mostly studying academic scientists. At the ESIP meeting, you can see that the academic presence is smaller than the government and nonprofit presence, like NCAR, because the academics tend to have their own priorities. Especially for academic scientists, they are really unaware that there are these existing data standards and systems that they potentially could or should be using. So my dissertation focus was on people who don't have much support for data management, asking: What do they do for data management? How do they create metadata? What does it even mean to say "metadata" to people who don't know what that is? For people who are completely in the science world, how do they do data management? These were the people who I was talking to, and they didn't work with ESIP or work with data standards. They are all just doing their thing science-wise. It wasn't until I got to NCAR, where data management is more institutionally embedded in these relations of organizations and governments that I started to see that there is a world of people who do do data stuff at a large scale.

So that must have been a pretty easy project because basically no one was doing anything.

Right. But even without data management, they were still able to do their science. So my question was, well, what is metadata even good for? They are doing their science, and from the outside, they have relatively poor, inconsistent, or uneven documentation. It's not that it causes them to not be able to do their science. It means that the metadata function is something different.

Do you think that things have changed much in the way that scientists work now?

I think what has changed is their awareness of the need to do data management. Places like ESIP have played into this. What has changed over the last 10 years is that scientists might understand and have a better awareness that there is a changing expectation for what they should and could be doing with their data. When I was studying this topic, the scientists were just starting to talk about data stuff. They were saying, "yeah--we are collecting lots of data and maybe we'll share it." Now, expectations have shifted. Scientists, particularly in the Earth Sciences, know that they should be doing something with their data because they are hearing these things from funders, journals, and professional meetings--like the American Geophysical Union (AGU) has big initiatives around data now.

So at the visual level (mind set and expectations), things have changed. At the practical level (day to day), however, that probably hasn't changed that much. Because if you are a scientist, and doing science is your job, doing this metadata stuff--they just don't know how to do it. You need someone to help you or you need an intensive deep dive into it, which of course they don't want or don't have time to do.

You mentioned that policies of funders, publications, and AGU, for example, are causing scientists to change the way they think about data management. Are there other things that have instigated this change?

It's a confluence of different factors.

One part of it is that the internet makes it easier to share things. The internet has changed expectations that we can and should be able to share things more easily. We have this now ubiquitous distribution mechanism that we use to share many other things--pictures, emails, websites, maybe also stuff that we probably shouldn't be sharing. In some cases, it is difficult to not share because things are online. There is this thinking that we have the internet, so we should be sharing everything because it is so easy to share things in theory. So why aren't we doing that with data? There is a technological element: it is easier to do--in theory. I say "in theory", meaning that it's easy to exchange a file, but it's not easy to create metadata.

The other part of it is that there is a cultural and societal change in expectations--sometimes it's a good thing and sometimes its a bad thing--around public scrutiny of science. For some fields, like climate change, there is an obvious public scrutiny on science, and that has changed expectations around data sharing.

So at a large general scale, I would say that change is being driven by these two things: it is technically easier to share things on the internet, and there is this cultural shift in thinking that things should be available because they are publically funded, and public scrutiny of science is something that is politically more important.

When you refer to expectations around data sharing, do you mean that its expected that data are shared more?

Yes. Here is one example of that. When you look at open data-type policy proposals at a government level, they are supported by both sides of the political spectrum. Making data open is politically bipartisan. But when you dig down into why people of different political affiliations want data to be open, you see different reasons. But regardless, the overall bipartisan support for open data is an indicator of this changing expectation. Because we are providing all of this funding for science, those things should be open. The expectation has changed, because the internet is there changing expectations, so everyone is saying that we should be changing data policy as well.

What are those reasons that differ between the two political sides?

On both sides, its economic value. If the federal government provides money to someone/something that produces data--either as an individual or a big program like NASA, there is the idea that these data could have economic value either to the entity that creates it or a secondary user. So on both sides there is a big push to get data out into the open, so somebody can make money with them. The weather industry is a good example of that. NOAA produces tons of data and there is a whole weather industry that is built on top of this open data that is funded by the government. On the other side, around climate change, there is public scrutiny on whether you are hiding or trying to distort data, so at the government level there are some groups that want data to be open for that reason.

So at a broad brush, you could say that one motivation for open data policies is to help science, and another is to ensure that science helps society and the economy.

Going back to what you said earlier, that scientists know that there are changing expectations around data management, but they don't know how to meet them. So how do they go about figuring out what to do?

That is the challenge, and that is why movement is slow. You can't just tell people that they need to exchange data and it happens. My personal perspective is that it is more of a human challenge than a technical challenge. I think of it in the same way of writing software. I've done a little bit of coding, but I wouldn't call myself a coder--I can do a little bit of scripting in different languages. If it's something simple and I can find something online, I can do it myself. If I were in a situation where I needed a whole new application or significant new coding, it just wouldn't make sense for me to do it, because I would have to learn a whole new set of skills to do it. A better situation would be for me to find a colleague to help me or write a grant to hire someone who really knows how to code. So I see data management as a similar thing--it has its own distinct set of skills, knowledge, and expertise. To ask someone to do it and spend a lot of time doing it, is somewhat counterproductive if they just don't have those skills.

My perspective is that the data that are widely used and that are around for long periods of time are there because there are specific people who are tasked to make sure that those data are well curated. This is where there is a big resource challenge--finding people who can help the average scientist get their data into whatever shape it needs to be in to exchange it. Obviously, there are technologies to help make things simpler, like the internet. We have tools that can do automated metadata generation, for example, in regards to drones. So, there are tools to make this easier, but my perspective is that if you are scientist, you should have human and institutional support for data management, so the organization supports you spending time doing that (or you're getting your students to do that). The organization might provide storage space, funding, and someone who can help--whether that is at the university or someone at a data center.

There are ways to think about doing that, like at a community level. So maybe every university doesn't have 5 data managers. Maybe there is an organization that has a group that you could tap into. You could go to them and say, I just need someone for a week to help me get my data into whatever shape I need to get it into. It might not be a whole person, but rather 20 hours of a person.

Many universities offer a statistician who acts like a consultant to research projects. It sounds like it could be a model like that.

Yes, some university libraries provide data management/curation services like this, but not all have the expertise or capacity. As another example, for my dissertation, I used a university computing service that built websites, which I used as part of my research. I couldn't have done some parts of my dissertation study if the university didn't have a centralized group that could do website stuff on my behalf.

Where do you think that Earth Science data / informatics is going in the near future (within the next 5 years) or far future (next 20 years)?

There is a trend now in finding shared models--cooperative models--for infrastructure and services across organizations. The idea is that everyone shouldn't and can't be doing everything themselves. Whether that is shared infrastructure, like common back up systems, or a common membership for assigning DOIs (Digital Object Identifier). In the library community, there are a bunch of libraries looking at a shared curation model across universities where expertise is pooled and tapped into in different ways. So perhaps not every university needs to have expertise in every specific field, but instead they coordinate that. I think that will continue for a while, because the resources are reducing and the demands are increasing. So I think the only reasonable way to move forward is to look at cooperative models. That is definitely going to happen.

Cloud will clearly increase in importance for organizations. It is still unclear how it will fit with some data centers, but it is happening and going forward.

I am also of the opinion that many of these data challenges are systemic challenges that are inherent in science, or institutions of science. The challenges won't be solved, in the sense that they get solved and then are done. They are ongoing challenges. One example is metadata--creating metadata, getting quality and valid metadata, exchanging and standardizing metadata--these have been problems in the library and data worlds for a hundred years. Progress has been made, standards have been created, but it is still a problem for the data community. I think it's going to be a problem still in 50 years. Same thing with identifiers and identifying things--this was hard 50 years ago and and it is hard now, and it will be hard 50 years from now. So you have to try to deal with things as they currently manifest with the current technologies, the current data, the current stakeholders that are involved. You try to make progress and try to come up with temporary and interim solutions but the problems still keep coming up: How to have quality metadata? How get metadata in the first place? How do we identify the stuff that we have? These are systematic challenges.

So in terms of the future, I think that we are going to keep working on these problems in perpetuity. We will continue to make progress but continue to encounter them in new ways. So I think that the challenges in the future will be many of the same things, in addition to new changes that will happen.

You’ve mentioned collaboration a few times, and that its key to many of these new mechanisms moving forward. But it’s clear in what you’ve said that collaboration is also hard. What are your thoughts around that?

This is where places like ESIP are essential. I’ve been part of a group led by Joel Cutcher-Gershenfeld that meets regularly to work on concepts related to stakeholder alignment. We’ve been talking a lot about the idea of middle-out solutions. The top down model is that a funding agency or someone from above will dictate thou shalt do such and such. That works to a certain extent. Then there is the grassroots model where everyone is doing what they want to do and maybe some common solutions percolate. The idea of the “middle-out” or “middle-across” is using coordination to work in both directions. You use an organization like ESIP to build those collaborations by getting the relevant stakeholders together. You can percolate up to the funders, “here is what the relevant stakeholders think is going to happen.” Then you can percolate down to the grassroots to say that we’ve done some coordination already and we have some suggestions or ideas for common approaches to deal with whatever your problem is, and you don’t need to start from scratch. So I think that this is a real value of these types of communities, which in some sense work at different levels. I was co-author on a short editorial that describes some of our thoughts around the “middle out” solutions. https://doi.org/10.1038/543615a

Collaboration is just essential to everything that we do. I think that most scientists resonate with that. Because every aspect of science is collaborative to some degree. Like the dissertation, which is the ultimate individual project, even has a committee that is working with you. I think that that aspect is not a problem. Most scientists would love to have help on their data management, as long as there is funding and the burdens associated with it aren’t too onerous.

Is this where you see the value of the organizations that you interact with?

Yes. ESIP is a very focused community. A lot of the same people at ESIP also go to AGU, but AGU is such a huge and diverse meeting that it’s hard to get a focus on anything. ESIP has the focused community--the group of people working on the data stuff, so its a place where you can just dig into that. Like the ESIP data stewardship committee has produced data citations recommendations that are pointed to as the best recommendations for Earth Science Data Citation. There was a paper written in 2011 that came out of an ESIP working group that Ruth Duerr led on comparing identification schemes. To me it is a foundational paper. So that is the good thing about ESIP. It’s not just a talk-talk meeting. People are here to produce things. The outputs have been valued.

Are there other organizations that have filled a similar role?

RDA (Research Data Alliance) obviously. Very similar and sometimes modeled on ESIP. But the challenge for the RDA is to build a sustained community. There are certain groups of people who work together on certain things. Maybe they produce some outputs. But each working group is a little different. You can look at a group of people like the ESIP Data Stewardship committee and some people have been there for 20 years, and have been key participants for 10-15 years. That level of continuity is really valuable.

[Disclaimer: The National Center for Atmospheric Research is sponsored by the National Science Foundation. Any opinions or recommendations expressed in this interview are those of the interviewee and do not necessarily reflect the views of NCAR or the National Science Foundation.]