Episode 11: Dr. Steve Van Tuyl

# KL: Katie Linder SVT: Steve Van Tuyl KL: You’re listening to *Research in Action*: episode eleven.

# [intro music]

# Segment 1:

# KL: Welcome to *Research in Action*, a weekly podcast where you can hear about topics and issues related to research in higher education from experts across a range of disciplines. I’m your host, Dr. Katie Linder, director of research at Oregon State University Ecampus.

On today’s episode, I’m joined by Steve Van Tuyl, the Digital Repository Librarian at Oregon State University, where he manages the university’s institutional repository, ScholarsArchive@OSU, and participates in providing research data services to students and faculty. Prior to his work at OSU, Steve was a Data Services Librarian at Carnegie Mellon University and a Reference Librarian at the University of Pittsburgh. In a previous life, Steve was a Biologist with the USDA Forest Service, conducting research on disturbance impacts on forest carbon cycling. Welcome to the podcast, Steve.

**SVT**: Thanks

# KL: I’m so glad you’re here to join me. I had met you because I went to a workshop that you gave on data management. First of all, I’d love to start by talking about data management and what it is, because it seems kind of like a mythical creature. We have so much data now, how do we manage it? Is it even possible? But I love this idea of data management. Tell me more. What does it mean?

**SVT**: Data management is kind of a broad set of activities researchers take on when they start doing research. Really, anybody who’s doing research is doing data management because you have to do something with your data. Data management involves everything from identifying what data you have in a research project, who’s responsible for the data at the various different stages of the research project, and how you document and share your data to other researchers, either in your research group or to researcher who are interested in acquiring your data for other purposes. That’s kind of the broad definition of data management.

We usually try to tell people what we don’t mean by data management, also. So we’re not talking about things like analytics (how to analyze data), we’re not talking about database management (which is a very different creature), and we’re also not talking about how to design a research project. Really, those types of things are what researchers know how to do well for their own research domain, and data management, more broadly, is kind of a set of tools and activities you have to do no matter what project you’re working on.

# KL: So it seems like data management, to some degree, has gotten a little more complicated because we have several people that could be involved in helping you try to think about the boundaries around your data management. I think immediately of IRB and how now you may have to include a data management section in your IRB proposal, whereas maybe that wasn’t something that was there before. Or if you have grant money, there may be some requirements around data management or data sharing. Is this something that someone in your position would help researchers navigate, all these different kinds of regulations that might be placed on data from a research project?

**SVT**: Yeah, typically the work that we do with researchers when we start talking to them about research data management is to try to cover all of those externalities first, so to try to understand where they’re seeking grant funding from, or where their grant funding already comes from, and whether they have IRB-related issues or other research-ethics-related issues that they need to be dealing with. Then, very quickly, we can either identify what a funding agency is requiring them to do or not do, or point them directly to IRB and say, “IRB knows way more about IRB stuff than we do. That’s why they’re the IRB and we are this other set of services.” But understanding that landscape of what a researcher is doing helps us provide them with better guidance than we might have been able to if we didn’t understand that landscape ahead of time.

In some ways, I don’t think it’s actually more complicated. I think that the data management landscape has become more regulated, but that has made things a little bit easier in some ways. Because at least you know, as a researcher, what’s expected of you, whereas 25-30 years ago, if you told someone to write a data management plan, it might have been hard at that time to just come up with something, because you wouldn’t know what the elements of that plan might be.

# KL: One of the things I’d love to talk about are how researchers can set themselves up for effective data management. When I had attended this workshop that you gave here at Oregon State, you talked about granular things like how you label your files, which I’ll just admit, that kind of thing fascinates me. I’m the person who wants the organizational structure that is going to be the most efficient. But you also talked about things like effective data storage and backup, which I also think is crucial for researchers who especially have datasets that cannot be replicated and that really need to be stored and backed up in a way that’s really helpful.

# I’m wondering if we can just chat a little bit about, what are some of the most important ways that you think researchers can set themselves up to be really effective with data management?

**SVT**: One of the things that I would recommend, and you mentioned this a little bit, earlier, I would recommend giving your storage and backup solution a really… spend some time thinking about what you have in place for storage and backup. That’s the kind of thing that can very easily get lost in the fray of doing your research. If ahead of time, or periodically, you step back and look at what you have in place for storage and backup, that can save you a lot of headaches down the road. We have a whole list of data loss disaster scenarios.

# KL: I think everybody’s heard some horror stories from somebody about data that’s been lost.

**SVT**: And as soon as it happens to you, which it happened to me a couple of times, it suddenly becomes real. It’s kind of like how you pay your insurance company for insurance and it doesn’t really make a lot of sense until, suddenly it has to. So I would say that’s one concrete thing to do is to be very intentional about your storage and backup solution.

A second thing that I would recommend is to think about documenting your data from the beginning, so thinking about… for somebody who may want to use your data in the future, that might be you. Actually, in most cases that’s the researcher themselves. Or a graduate student that you bring on, or a colleague that you’re sharing your data with. That documentation is going to be so critical to their ability to use the data and understand what it is. In more of a practical way, you spend less time explaining things to people over and over again when it comes to understanding your data.

# KL: I heard this term bandied about quite a bit, “metadata.” Is that what you’re referencing here? If so, could you tell us a little bit more about what it is?

**SVT**: Yeah, I guess its metadata. Metadata is structured information that helps you understand what a dataset is or what a “thing” is, and how it might be used, and where it comes from, and all these different types of relationships that digital objects have, essentially. Typically, I think, when people think of metadata, it’s this lurking monster. There are lots and lots of metadata standards out there that your research domain may or may not use or recommend. Metadata oftentimes is encoded in something like XML, so it looks very complicated on the surface to users.

So what I try to do is to put that idea of metadata away and say, “What we really mean when we say ‘metadata’ is that you need to provide sufficient documentation for your data so that somebody can understand what it is, understand the context that it came from, so even an abstract from your project proposal provides enough context that somebody knows why the data exists in the first place, and the metadata should provide some description of how the data elements in your dataset were collected or created, or whatever it is.

Many, many of us have used other people’s data in our research. Those of us who have done that have a pretty good understanding of how difficult it can be to understand what the data is. Good documentation, while sometimes rare, can go a long way to helping people know what the data is, but also to not misuse the data because they didn’t understand the methods that were used to create it or collect it in the first place.

# KL: Absolutely. Sometimes when I think about metadata, I think on a much smaller scale about things like citations, and this is why we care about citations being correct because you need to be able to go back and find out what this was about, if someone’s referencing it or using it as an example in an article or in a research study. I think metadata is clearly more than that. It’s more detailed, there’s a context involved, like you were saying. But I feel like it’s one of those areas of research that could be kind of a pain, like citations can be kind of a pain, but getting them right is still critical to communicating your research and making sure your research, or your data, gets used by other people in the ways that are going to be most effective to furthering the field.

Well now that we have a nice strong foundation about what data management is, what it looks like, what some of the component are, data management plans can be pretty complex and have lots of moving parts. We’re going to take a short break, and then come back and talk about those in segment 2.

# Segment 2:

# KL: So Steve, data management plans. I’m seeing them everywhere on various applications I’m working on. I am seeing them requested by places like IRB, but also clearly grants and some other kinds of things I’ve been looking at recently. Can we start by just saying, “What is it?” and maybe, “When did it start becoming a thing to have a data management plan?”

**SVT**: A data management plan is usually a broad description of what data is going to be collected and what you intend to do with the data to meet the requirements of some “asker” (an entity asking for a data management plan). That entity may be you asking yourself what you’re going to do with your data.

# KL: One of the things that I’m curious about with data management plans is, “What is their origin?” Some researchers might say, “You know, I know how to handle my data. Why do I have to give this data management plan? Is it because there’s so much data?” Where is this coming from, these askers who are asking for these data management plans? What’s the motivation behind that?

**SVT**: One way to think about the motivation and where this is coming from is about return on investment for grant dollars. The specific dates always escape me, but right around the turn of the 21st century, NIH started asking for data sharing plans, I believe they were called (or are called), for very large grants, so $500,000 or more. In some ways, that may be one of these origins of a data management plan. I think we were seeing some requests for data management planning, very specifically before that, for very specific research domains or grant opportunities for very specific types of research.

Even going back to the 1960s we saw NASA talking about data management as an important thing because they realized they were collecting all of this data and started offering guidance on what needed to be done to make sure this data was usable into the future.

More recently, though, going back forward in time, probably about 5 years ago NSF started requiring a data management plan for basically every proposal that came in for an NSF grant (although there are some exceptions, I think). What those data management plans look like is kind of what we’re starting to think of as a normal data management plan for a grant, which is a two-ish page document that lays out these different components of what a data management plan is and says that you intend to do as far as data management is concerned.

In 2013, the office of science and technology policy of the White House took what NSF did and said, “We intend that every agency with over a certain amount of R&D money is going to require this type of data management plan, also.” There were some other elements of that mandate that came out around data sharing and publication sharing, but a data management plan was a big component of what they were asking for.

# KL: So this seems like it’s a… I mean it’s definitely a thing now. It doesn’t seem like it’s going to go away anytime soon. Actually, I have to say, for things like this, I kind of, to some degree, appreciate it. It’s forcing a level of planning and intentionality with data that I welcome because there’s so many different components of research that I think you don’t want to lose sight of the impact that long-term planning can have on things like data sharing. But let’s talk a little bit about, “What are the typical things that are included in a data management plan? What do people need to be thinking about as they’re preparing this document?

**SVT**: When we think about a data management plan, and especially one of these short-form data management plans that funding agencies are asking for, there are a handful of things that we try to make sure people are including in those plans.

The first thing is, “What data are you going to actually produce for a project?” That kind of sets the groundwork for what needs to happen subsequently.

The second thing is about how those data are going to be handled during the project. Are they going to be backed up and stored securely? How they going to be passed back-and-forth from site to site (if you’re in a multi-institutional research project)? Dealing with those operational elements of data handling is also important.

Then we start to get into questions about responsibility. When we say that the backups are going to happen weekly, who is going to be doing that? Or when we say that one or a handful of people are going to be primarily responsible for collecting and handling the data, it’s helpful to identify who those people are on the project. That seems to be not so much for the funding agencies as it does for the researchers to say, “Oh, now we know in our research project who to go to when there’s a problem.” Because otherwise it can get very confusing when something comes up and you don’t know who to talk to.

So who’s responsible is a big part. The next element that we try to include is documentation standards. That’s the metadata that we talked about before or other types of documentation for the data. If there are standards that exist, what do those look like and which ones are you planning to use? If there are not, maybe a brief description of how you intend to document the data so it’s usable to other people.

Then the last two elements get at data sharing and opening up your research to the scientific community. “How do you intend to preserve and make the data accessible to others?” And the second piece to that is, “What are the conditions under which you make that data accessible?” Because as we know, there are some types of data that are not appropriate for sharing openly, but many of those types of data are still appropriate for sharing in some context. If you need to be explicit about what context those data can be shared in, laying that out in the data management plan is pretty important.

So those are kind of the five or six major elements of a data management plan that we look for.

# KL: And all in two-ish pages.

**SVT**: Two-ish pages.

# KL: That seems like kind of a tall task for folks to complete.

**SVT**: Yes.

# KL: So we may have some listeners who are hearing all of this data management… data management plans, all of these different components, and maybe feeling a little bit overwhelmed. What are some of the resources that are available to help researchers think about these plans and how to write them, I would say maybe particularly for people who are beginning grant writers or who are coming back to it after a while and are realizing there’s some changes now in terms of what’s being required.

**SVT**: There are two major resources that I would point people to. If you’re at a research institution, it’s likely that your academic library or some other unit, possibly the research office, is providing guidance on data management to faculty and graduate students at the university. So those are the people that I would try to identify early in the process, because if you have an issue you might be able to talk to them about best practices, or tips and tricks for writing a data management plan.

# KL: I’m wondering if they might even have examples that they could share based on what kind of unit it is.

**SVT**: Yeah. In some cases, data services folks at universities have examples at hand. Those may be fake examples, so a mockup of a good data management plan. At some universities, the research office has been in a position to share data management plans as examples (and they’ve been anonymized and whatnot), but you can get a sense for what a good data management plan looks like.

If those services don’t exist at your institution, there are a handful of organizations around that are also providing guidance for data management planning. The one that I would point out as my favorite is an organization called DataONE. They have a number of things that they do, but one of the things that they provide is guidance and educational materials around data management planning.

The second thing that I would point out, and this is a tool that I think almost anybody can take advantage of, is a tool called DMPTool (that stands for Data Management Plan Tool). This is a tool that was developed out of grant funding by the California Digital Library and a number of other institutions across the U.S. many years ago. This tool is a space where you can identify what grant stream you’re applying for, and then specific guidance is given to you inside the tool for how to write a data management plan to meet the expectations of that funding body.

The other thing that the DMPTool does that is really nice is that instead of just giving you just two blank pages and saying, “Write the data management plan,” they actually break the plan up into its component chunks. We talked about these five or six pieces that are important, so they break the plan up into those chunks so you’re not trying to write the whole thing at once, you’re trying to write these individual elements, which is really nice.

# KL: Those sound like amazing resources. We will definitely make sure to link to those in the show notes. We’re going to take another quick break. When we come back, we’re going to talk about something that Steve’s really passionate about: open science.

# Segment 3:

**KL**: Steve, I know one of the things that you are passionate about and interested in is open science. This is something I don’t feel like I know a ton about, so can you just start us off by telling me, what is open science?

# SVT: Sure. Open science is one of those terms that I think a lot of people have different definitions for. My flavor or feeling about open science is that it’s a movement that has been around for a while, because there are many research communities that have been very open with their data and with their publications, even since the inceptions of these research communities.

Some research domains are a little bit less comfortable opening up their research to the broader scientific community, so the open science movement has sought to make that easier for researchers who are not familiar with how to open up their research to others. When we think about opening up our research, we mean things like sharing publications in venues that don’t have paywalls, sharing datasets that are the results of our scientific process in venues that don’t have paywalls, and really giving back to the scientific community by making the things that we do available for others to use and reuse.

**KL**: By the name alone, I wonder, is this something that’s limited to particular kinds of disciplines in the sciences because it’s called “open science,” or is it going more broadly that that? Are we seeing things in the humanities as well under this umbrella of open science, or is it more limited?

# SVT: It feels a little bit more limited to science. Of course, social sciences are folded in there. In some ways, the idea is open to anyone doing research, but operationally, I think we see that the open science community tends to be more focused on types of research that have a familiar trajectory of research production. So you create data or you collect data, you analyze data, and you publish on that data. In some ways, that’s where it stops for most people. But what the open science movement would ask you to do at that point is to make all of that process available and to make it very clear to everybody what you did and why you did it and offer proof that what you’ve done is what you said you did. So by sharing your code, sharing your analysis scripts, sharing your data, and all of these types of things

**KL**: As someone who loves process, I love process, and I also love seeing examples of process because I think… particularly, when I think about working with novice researchers or being a novice in certain areas of research myself, that’s helpful for me just as a learner to see what other folks have done.

But I can imagine that, in something like open science, one of the real concerns would be about things like, to some degree, maybe things like credibility or recognition or how things are being cited, or even, “How do you put things like this on your CV?” If it’s not necessarily going through something that’s peer-reviewed. It sounds like it’s really disrupting some of those foundational components of academia and how people are recognized. Are there ways that people can engage in open science and be assured that their work is really going to be tied to them?

**SVT**: Yeah, that question is hard. Honestly, there are no assurances, but that’s kind of how academia is, in some ways. The open science community, though, and the broader data management community is thinking about these issues and trying to understand how to give credit where credit is due. There are a number of ways that we started doing that.

One of them is that when you publish the data, you’re not just putting it out there on a website, necessarily, but you’re putting it out in a venue that looks more like a traditional publication. So datasets are often given DOIs (digital object identifiers), which we see typically on a journal publication, but now datasets have them so that you can do things like track what’s happening with that dataset. Is it being downloaded? Is it being shared? These types of things. Also, you have a persistent identifier for that data so you know that the data that someone’s using is the data that you shared. That kind of infrastructure is helpful for giving credit.

Likewise, there are people thinking hard about, “How do we cite datasets?” In the same way that we cite a publication, we need to think about how we cite a dataset, because that citation is really one of the currencies of credit in academia. So we try to translate that over to sharing data or sharing code or sharing other types of content that are part of the scientific process.

The other thing that we can do is we can try to use tools that aggregate all the different types of research outputs that you have into one place so that it can be very clear what you’ve done and who you are and what your research outputs are. One tool that’s like that that seems to be gaining a lot of traction is something called an “ORCID,” which is an identifier for researcher that allows researcher to aggregate under a unique identifier for them what types of research outputs they’ve created.

# KL: So for folks who are listening and are interested in learning more about ORCID, we will definitely link to it in the show notes. But just to clarify the spelling of it, it’s not spelled like the flower. It’s ORCID if you want to Google it and find out more. This was something I actually looked in to after I had learned about it from you, Steve, just to see what it was. One of the cool features of ORCID, once you set it up and you have this unique identifier, is it will help you to go out and search the Internet to see what are all the different things that your name is popping up with, and then you can click and say, “Yes, that’s me,” or, “No, that’s not me.” Similar to how you might use something like Google Citation. Is it Google Citation or Google Scholar?

**SVT**: Google Scholar.

# KL: A portion of Google Scholar is you can bring your research together also under your name, and it will go out and seek things out for you. But ORCID seems like a more authentic way, maybe is more the word I’m looking for, of correlating these things together in a very intentional way of gathering all of your disparate research projects and bringing them under one umbrella.

**SVT**: Yeah, I’d say that even as recently as a couple of years ago it was not clear to many of us what standard would rise to the top when it came to research identifiers. It was kind of a Betamax-VHS problem. So individual publishers and content aggregators like Thomson Reuters and Elsevier were creating their own researcher identification systems, Google Scholar has its own thing where you can identify yourself as researcher doing your own thing, and there were a handful of other entities out there entering this sphere.

But for various reasons, ORCID over the last few years has gained traction with a lot of funding agencies and publication bodies. Actually, with these other entities that were creating researcher identifiers, they all seemed to be converging on ORCID in a meaningful way, so I think just by the way things have gone it’s gained authenticity, to use your term. And it’s being more widely recognized, which is great. It’s nice to see that there’s one thing out there that people can use to uniquely identify themselves.

# KL: Steve, why is openness good for science in general?

**SVT**: Openness is, in some ways, fundamental to the way the scientific process is expected to work. We think of the scientific process as building on the shoulders of giants. We do our research based on research that other people have done and questions that they’ve asked and results that they’ve produced. We know that the whole system is flawed, not entirely because of intentional problematic behavior, but it’s just not a perfect process. The extent to which we can take advantage of openness to refine that process is pretty key to driving things forward.

We have a number of cases in the literature where, when researcher have opened up their data, amazing progress has been made in fields of research, or new science is done based on data and publications that have been made available to others.

To me, this really always goes back to, “This is how the scientific process is meant to be done.” Science is not done inside your head or just inside your lab; it’s done as a community. And while there’s competition in that community, which is helpful and healthy, ultimately the point is for the community of scientists around a research problem to come together and solve the problem together. Openness hopefully facilitates that.

# KL: Well I want to thank you Steve so much for coming and joining me today and sharing your expertise.

**SVT**: Sure, no problem.

# KL: And also, thank you to our listeners for joining us for this episode of *Research in Action*. We’ll be back with a new episode next week, and I hope you’ll join me then as well.

Show notes with information regarding topics discussed in each episode, as well as the transcript for each episode, can be found at the *Research in Action* website at [ecampus.oregonstate.edu/podcast](http://www.ecampus.oregonstate.edu/podcast" \t "_blank).

There are several ways to connect with the *Research in Action* podcast. Visit the website to post a comment about a specific episode, suggest a future guest, or ask a question that could be featured in a future episode. Email us at [riapodcast@oregonstate.edu](mailto:riapodcast@oregonstate.edu). You can also offer feedback about *Research in Action* episodes or share research-related resources by contacting the *Research in Action* podcast via Twitter @RIA\_podcast or by using the hashtag #RIA\_podcast.  Finally, you can call the *Research in Action* voicemail line at 541-737-1111 to ask a question or leave a comment. If you listen to the podcast via iTunes, please consider leaving us a review.

The *Research in Action* podcast is a resource funded by Oregon State University Ecampus – ranked one of the nation’s best providers of online education with more than 40 degree programs and over 1,000 classes online. Learn more about Ecampus by visiting ecampus.oregonstate.edu. This podcast is produced by the phenomenal Ecampus Multimedia team.

# Bonus Clip:

# KL: In this bonus clip for Episode 11 of Research in Action, Steve Van Tuyl talks about why you should expand your two-page data management plan. Take a listen:

# KL: I think one of the things I’ve found, as someone who does grant writing, is I’m always incredibly thankful to myself upon receipt of a grant that I have gone into extreme detail in the application about the kinds of things that the grant money will do, the timeline for those things. As I’m writing the grant and doing that work, I’m always thinking ahead to, “Your future self will thank you if you end up getting this money because you’ll be able to hit the ground running.” And oftentimes you get the grant and it’s like, “Well, go ahead and start.”

# And if you don’t have a plan, it can be a little bit scary. Data management plans, to me, sound very similar to that. It’s really trying to plan ahead for the eventualities that might happen with your data and to really be ready to hit the ground running when that data is collected in terms of how you’re sharing it with colleagues, how you’re backing it up, those kinds of things. That makes a lot of sense to me in terms of planning ahead for research projects to make sure they’re going to be effective.

**SVT**: You’ve touched on a couple of things that I think are really important. One of them is that these data management plans that are being requested by many agencies now are relatively short in comparison with the amount of information that would actually be useful. But, it is useful to have a plan in place when a project starts. So there’s this balance there. You don’t want to write a full-blown data management plan for a project that hasn’t been funded yet, but at the same time it’s very helpful to have one once you get going.

One of the ways that I like to try to think about this is that you can write a brief data management plan for a proposal that really covers some of the basics, but doesn’t get into a lot of really fine detail. But once a project is about to start, or once it has started, it can be very helpful to stop for a minute and take that two-page data management plan and try to operationalize it. By that I mean to expand it to include a lot of that detail before the project gets too far underway so that you have a better sense for what to expect from yourself as the project moves forward.

# KL: That’s a phenomenal idea.

**SVT**: Yeah, it’s kind of a two-phase process because one of the things that we see, or are starting to see, is that it can be very easy to play this kind of “compliance game” when it comes to data management. What winds up happening with that, which I like to call the compliance-side economics of data management, is that you write a data management plan at the beginning, and then the project ends and you do the data management plan outcomes at the end.

So there’s that whole swath of stuff in the middle that is this black hole that you need to fill. So if you want until the end of a project to do the data management work, it feels like a huge burden and you could run into problems if something happens during the project. So that’s where taking that two-page plan at the beginning of the project and operationalizing it so you know what to expect from yourself over the lifespan of the project is very helpful.

# KL: You’ve just heard a bonus clip from Episode 11 of Research in Action with Steve Van Tuyl talking about why you should expand your two-page data management plan. Thanks for listening.

# “Research in Action” transcripts are sometimes created on a rush deadline and accuracy may vary. Please be aware that the authoritative record of the “Research in Action” podcast is the audio.