Data Papers in the Network Era

Mackenzie Smith
Massachusetts Institute of Technology, kenzie@mit.edu

Mackenzie Smith, "Data Papers in the Network Era" (2011). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284314871

Data Papers in the Network Era

Mackenzie Smith, Research Director, MIT Libraries

Good morning. Again, my name is Mackenzie Smith, and I'm a Research Director at MIT libraries where I was, until recently, the Associate Director for Technology Strategy there.

I think you're going to see some interesting synergies between my talk today and the talk you just heard, because I'm also a linked data person and many of the things we're going to talk about build on some of the background that Mike Keller just gave you, hopefully in a useful way.

As background, the MIT libraries have been involved for many years in developing innovative tools for the content industry, particularly libraries like DSpace, the open source institutional library platform, and Simile, which is a set of open-source tools for linked data publishing and visualization on the web. I will talk a little bit more about that later.

More recently, we have been very involved in thinking about the role of primary research data in scholarly communication and particularly how to apply linked data standards and tools to research data, which is all my way of explaining why I am here to talk to you today about the concept of data papers and why that idea may solve some of the problems we have today in getting the full benefit of research in the network era.

Why data sharing is important: I'd like to start with explaining why this problem is of such pressing importance today and why I'm here to talk to you about it. The most immediate driver for research is mandates—research sharing, I mean. Many funders are requiring researchers to show their research data now, and there is growing pressure to provide better access to, and accountability for, taxpayer-funded research results. This is also driving a lot of the open access debate. One notable example of this is the National Institutes of Health (NIH) data sharing policy for any grants in excess of a certain amount of money, $500,000.00 in their case, and this policy has been in place actually since 2003. Another example: This year, we have a new policy from the National Science Foundation (NFS) about data sharing, and this applies to both PI's of grant projects and the research institutions they work for, which are the official grantees. The new guidelines include a mandatory data management plan which has to be part of every single grant proposal that gets submitted to the NSF. These plans are now part of the competitive review process, so with federal research funding that is continuing to get tighter every year, we now expect to see data management plans become a competitive advantage for PIs who do a good job with them.

The NSF guidelines are available online at http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpgprint.pdf. These guidelines say that NSF data management plans have to be explicit about certain things, in particular about the policies and provisions that you are making to share your data, including future reuse, repurposing, and redistribution of that data. The reason for that instruction from the NSF is to get more leverage for that expensive data that they are funding the production of to generate new research and to get greater impact from that funding, which may make pretty good sense to everybody. But from the researcher's perspective, the really big driver is credit. The core principle of the scientific method is that research should be reproducible to get the best possible science. While reproducible data-driven research is still very difficult to achieve in a lot of scientific disciplines, it requires changes to workflow and scientific processes and is a very good reason for researchers to want to share their data. So they have these two drivers: mandates from their institutions and funders, and also this underlying desire to do better science and make the research reproducible.

But if data sharing is such a great thing to do, and it is expected in so many cases, why hasn't it happened already, and why is it not routine today? And the reason is that it is still really hard. Even in the Internet age and with ubiquitous platforms like the web available, it is still hard. Most researchers don't object to sharing their research data; some do, but most don't. But things get in the way, a lot of things, like fixing data quality problems and documenting the data, usually after the fact when you've forgot-

DOI: http://dx.doi.org/10.5703/1288284314871

ten many of the steps that you took; losing control of your data and who is going to do what with it is a very serious concern for some researchers; often very serious confidentiality and privacy issues arise, like HIPAA regulations or protecting the location of an endangered species, and commercial interest from both private industry and universities in whatever intellectual property rights might accrue to the data. This causes a lot of confusion about the policies by which data can be made available, which is one of the things NSF asks you to be clear about. This relates a little bit to Mike's plea for open metadata. Metadata is in many ways the data of the library profession, and he and I agree with this proposal that would make this openly available to get the most benefit from it.

This is also true for the data of science, social science, and humanities research. Basically, we think that most data should be made openly available to get the most benefit from it. What stops researchers from doing this now is a lack of credit for all the extra effort and work that it is going to take to show their data effectively. The scholarly communication system needs ways to count that high quality data as a legitimate part of the individual's research record and a valuable contribution to science or whatever discipline you are a part of. This has happened in a few cases if you think of people like Craig Venter and Eric Lander and the human genome project; they have gotten a huge amount of credit for creating that data and then sharing it publicly. More on Lander's side that Venter's side, but that is another story. The right cause won. And a barrier to including data in the scholarly communication system is the lack of infrastructure that we have had, which I'm going to talk about in a few minutes.

So, first I need to say just a few words about what data is and what is the state of data, because I learned over the years that the data means very different things to every single person that you talk. As we well know, anything can be used as data for some purpose, and these days a lot of the public discussion around data and whether it should be shared or not is actually talking about business data like website click streams and public sector information, which is government produced data, whereas I am talking about data that underlies research, and particularly scientific research. So, research data typically includes things like observational data, which would be things like sensor readings, telemetries, and survey data. It also includes experimental data, gene sequences, and spectragrams. It can include media, which includes text images, audiovisual files, or a neuroimage, which is still an image even though it is created by an MRI machine. Simulations are a kind of data, and these are typically software or algorithms rather than numerical kinds of data sets. So, a key property of a lot of data is that it would be prohibitively expensive or impossible to reproduce. One of the big drivers of sharing data is that you cannot get it back again. If you're doing climate change research, for example, or if you're taking sensor data from the ocean to measure temperature and salinity and things like that, you have that data from a moment in time and you can't just go back in time to get a new sample, because once time has passed that data changes. So, you understand this is a very time-based type of data. Whereas other kinds of data, like genomic data, can be easily reproduced, so, in fact, there is debate about whether gene data should be shared and kept for long times because it is getting cheaper and cheaper to just re-sequence a genome then to store an old one and the techniques get better too. So, there's a lot of tension in the community over how long to keep data and that type of thing. But sharing in the first place is really not that controversial.

Also, keep in mind that way more than text, data can be in standard, or proprietary, or discipline-specific formats, like the FITS format in astronomy—it's very specific to that discipline. It can even be specific to a particular instrument, like one particular confocal microscope has a proprietary format that comes off that microscope that the maker of the instrument dictated and owns control of. Data also requires software to do anything, and that software can also be standard or common, like the language "R" for statistical processing, or it can be proprietary and discipline specific. So another important property of data is that typically without the software the data is useless. The distinction between data and software is getting very blurry. Data can't be neatly packaged like a book, so it has very fundamental differences from the kinds of content that we've historically dealt with.

Finally, what do our researchers want to be able to do with this data, to inform how we want to share it? Well, obviously they need to be able to find it (as Mike just explained) evaluate it, process it, analyze it, visualize it, and annotate it. Sometimes they want to reuse it, whether it is to validate an experiment or do new research of their own, either alone or in combination with other data, which is a very different set of requirements than what we have for text, for articles, books, and more traditional kinds of content.

On the last point about reusing data, that has become a big driver for data sharing for a few important reasons. First is cost. As I explained, a lot of data can only be produced once and it is also often very expensive to collect. Take an example like a neuroimaging study where every single scan with MRI cost a minimum of $1,000.00. You can get so many scans in your study, but you may not be able to achieve good statistical significance on your own. If you can combine your data set with other studies that did similar kinds of research, then you get a much bigger pool of data to do your analyses on and much more impressive and believable results. So, there's a lot of pressure in many scientific disciplines to be able to pull the data to get better results—that is a big driver.

The second is interdisciplinary; to be able to combine data from different fields, for example, climate change data with economic and population data, to look at the impact of policies and politics on climate change. Those fields do not talk to each other; their data is in very different formats but there is growing need to be able to combine it in order to perform important research.

And third is the growth of computational science, like building better disease models from large aggregations of clinical trial data which are seen with efforts like the Sage Bionetworks effort. If you haven't looked at that, it is an open access database in clinical trial data from all the big pharma companies who have decided that data is actually pretty competitive and that they will get more advantage by aggregating and sharing it so they can mine it than they would if they clung to their data and kept it private.

So, for whatever the reason, integrating data is really important, that it is extremely difficult and labor-intensive today and, in part, that is because data without meaningful structure and documentation is useless. It is just columns of numbers, you don't know what it means, and the only person who really does know what it means is the person who created it and maybe a handful of researchers who worked with them. Solving this problem is not something that third parties like libraries or publishers are going to be able to do after the fact. It has to be part of the research workflow somehow, and that requires better tools and some changes to current research practice. That doesn't mean, by the way, that there's not a role for libraries and publishers, and I'll get to that in a little while.

Reusable data is all of these things: structured, versioned, well-documented, so that you know exactly what you are getting; formatted for long-term access so that you know it won't disappear the next time you need it; archived somewhere, presumably in the library or an archive; findable and citable, to Mike's point, you need ways of being able to figure out if this data exists in the first place, which is not trivial; and legally unrestricted or with a very clear usage policy so that you know what you can do with that data once you find it.

This brings us to the main concept of this talk which is the data paper as a way of solving some of the problems I've just described. The data paper is "a formal publication whose primary purpose is to expose and describe data as opposed to analyze and draw conclusions from it." This is a quote from a paper on data papers published by the Narrow Commons Project, which is part of the Science Commons part of Creative Commons (http://neurocommons.org/report/data-publication.pdf).

The point here is that data papers are like traditional research papers in some aspects: they are formally accepted, they are peer-reviewed, they are citable entities, and so on. But, in other respects they are very different from traditional research articles because they are not about the research, they are about the data. If data papers catch on, we will start to see sets of papers about particular research projects, some which are more analytical and some of which are more technical-semantic. Just in case you

think I'm inventing all of this, data papers have been around for quite a long time. For example, the *Journal of Physical and Chemical Reference Data* publication from the American Institute of Physics started in the early 1970s to describe data about physical and chemical materials of general interest. This is still in publication; we subscribe to it at MIT. So the concept has actually been around for quite a while. But, the older journals that date from the print era tend to be not particularly useful in the modern environment—or, not as much as they could be—because what they do is visualize the data to a print format and then publish that as a PDF page, so what you're getting is a static visualization of the data rather than the data itself, but, it's getting at the concept that we are talking about.

More recent forms of data papers are taking more advantage of the Internet and the web, like supporting data downloads. So, take *Ecological Archives* from the Ecological Society of America. It's a modern publication. The data itself is open access, but what you see is that you can only download the data, that's all you can do with it, and the documentation here is very complex and completely unstructured. This is not something a machine can help deal with; you just have to read this long, long, long description of the data and then download it, so we can do better.

There is also an effort going on at National Information Standards Organization (NISO) to come up with a new standard for supplementary files for peer review published research articles. This is also a necessary step, but it's really focused more on the paper and the data is sort of a decoration of the paper in this case. It's not really a first-class object of its own; it is just trying to help standardize how this particular linking gets done. This brings us to some recommendations for independent scholarly publications of data sets. What we can envision data sets becoming in the near future.

This is a paper from Jonathan Rees at NeuroCommons (http://neurocommons.org/report/data-publication.pdf), and he is trying to identify the key components and requirements for a formal data publication. He claims and recommends that published data should have certain properties: be organized, peer-reviewed, and have established quality-control measures. This is not something new to the publishing world; we would expect this in anything that's considered a formal publication. It needs to create a citable entity, something the other researchers can refer to and know will still be there in the future. It needs to establish cross-linking mechanisms with the traditional papers to enforce that they are different but related—the set that I was describing a moment ago—it needs to specify what required documentation is needed to make the data really usable so these would be new standards for documentation metadata for papers in addition to the ones we're familiar with to support discovery; it would supply standard and very importantly interoperable legal licenses to the data sets and examples of those might be the Creative Commons Zero Waiver of Rights so there are no IT claims made on the data at all or various kind of attribution licenses, usage licenses, and other techniques. The point here is that it needs to be normative so that people are sure that they can combine data legally. And then finally, we need an archiving strategy in place so that the data, like the papers and the metadata for the data, stay around long enough to become part of the scholarly record.

One thing I think we can all agree on is that whatever this infrastructure is for data publishing, it has to be web-based. And to achieve the degree of data interoperability that we want, we need to look at linked data, the set of web standards underlying the semantic web that we've been given such a good explanation of just a few moments ago. So what would that infrastructure look like? There are three kinds of infrastructure that I'm going to talk about now that are key to this idea publishing data, that are already happening, and that we can invest our time and effort in leveraging and building out.The reason I'm here today is to kind of light a fire here and see if we can get more progress in these areas.

The first is identifiers. As Mike explained, the web requires identifiers for resources on the web, or entities on the web, and those are called URI's. This is absolutely even truer for linked data than it was for traditional content on the web. In online journal publications we've seen some new identifier systems emerge that were developed for publications like the Cross Ref DOI's, but for data papers, we're going to need more kinds of identifiers, in particular, people. Mike gave a very eloquent description

of authority files from libraries, but the truth is that they're not useful on the linked data web because they don't have URI's. Yet. There is an effort that has started called ORCID, the open researcher and contributor ID, which will become a registry of people that have globally unique URI's associated with them but that you can start to use in publications. This initiative, ORCID, actually came from the publishing community with the help of some libraries including MIT, and it is launching next year. The idea here is that all universities and publishing houses would join ORCID and make sure that every researcher they are dealing with has one of these unique identifiers. What is behind this identifier is a profile for the researcher. In the profile data could be library authority data. That would be a fantastic way to seed this registry, but without the URI all that lovely authority data is not usable on the linked data web.

In addition to people, we need identifiers for institutions, and there is an effort at NISO called I2 Institutional Identifiers. I don't think it's quite as far along as ORCID, but it's absolutely necessary because in order to apply credit to researchers, you'd need not only URIs for the individual researchers, but also for the institutions that they work for since they move around a lot.

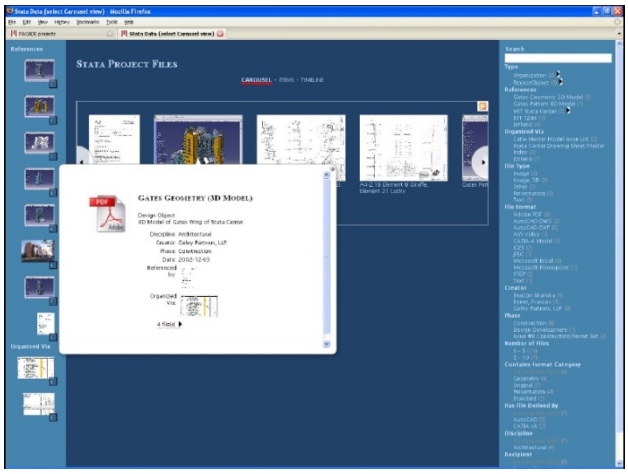
And finally, we are going to need identifiers for data sets, similar to articles but with some very important twists like versions of databases, which we did have to deal with a little bit in the article world, but it's much, much more prominent in data. And then you've got subsets of data sets, such as your big genome database from which you want to refer to just one gene or a set of records you pulled out. And you've also got data sets that were derived from multiple data sets, so aggregations. So, anyway, there are lots of variations of what you need to be able to name, but we need standard identifiers to do that, and fortunately there are two. CrossRef DOI's can be assigned to data, and some data producers are doing that now, and then the Data Site Initiative is one that the library community has invested quite a bit in including the British Library and the California Digital Library. These are both good efforts; they both use the same underlying URI syntax of handle so it is a good direction to go in for identifying data itself.

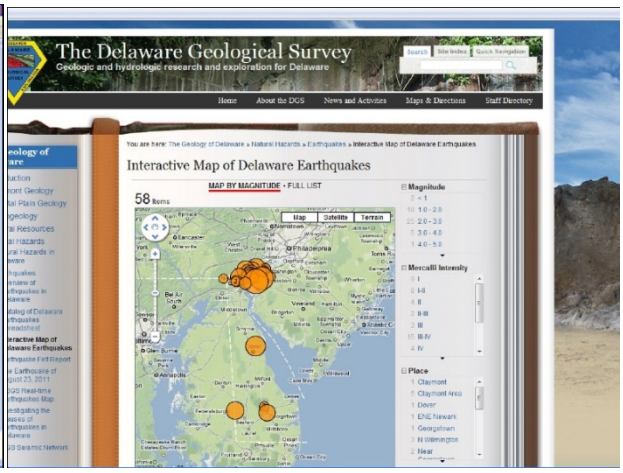
As I said earlier, research data can take many forms and can be encoded in lots of different ways depending on the discipline that is doing the research, depending on the source of the data, and depending on the tools that are available, among lots of other factors. So that works within a discipline. You can be as quirky and bespoke as you want to, as long as you know everybody who is looking at your data and trying to do similar work. It falls apart when you're trying to do interdisciplinary research, which is becoming more and more of a driver of research in general. So we have to start thinking a bit broader than just our own boundaries, and that applies to libraries as well. It also wastes the fantastic data sharing infrastructure that we already have in the World Wide Web. Web standards like XMLN anthologies for different data types could support much easier data integration and leverage all kinds of great tools that weren't designed for science, but could very well be used for it, just like researchers now rely on Microsoft Office, which was not designed for scientists but which of course has proven to be a critical tool in their toolkit. Things like Excel are the most popular data structuring software in the world, sadly. So, at the technical level, just at the technical level, data really is just data, whatever type it may be, and doesn't require quite so much custom infrastructure as a lot of researchers would have you believe. So, the first step is agreeing to share the research data using common web formats and developing new ontology's to structure that data more effectively for interdisciplinary research.

Which brings me to another aspect of data publishing infrastructure, and that is what you do when we reach the data itself. What do you see once you go and look at data that has been structured in this linked data standard way? Web based and linked data compliant visualization software allows researchers to explore this linked data that they have retrieved, ideally, in different modalities, which means, for instance, on a map, on a timeline, in facets and so on, along with associated metadata and documentation so they don't have to jump from the data paper to the data set to the article as discrete things, that they are all connected together. And, in the same way that an e-journal is pretty useless without a web browser to display text and allow readers to navigate through the article, data browsers are going to be the key to success of data papers

and published data sets. I would argue data browsers are the key to success of linked data in general. Web browsers are not natively able to support linked data. We need another generation of web browsers, and we call those data browsers.

One example of an existing data browser is an open source platform called Exhibit, which we developed at MIT, so I will use that as my example. And, the Exhibit data browser is being used for all kinds of different data archives. This one is TB Commons which is in a new public archive of clinical trial data on tuberculosis drugs that Eli Lily has published, and this example supports searching metadata and visualizing the data itself. So, you can seamlessly navigate between the metadata and the data itself which become less and less a meaningful distinction in this kind of environment, as Mike showed with Freebase. By the way, two of the principal architects of Freebase, David Huynh and Stefano Mazzocchi, were at MIT and wrote Exhibit with me on this project before they went to Metaweb and then Google so there's a lot of history that is coming out of the library field, which is great.





Another example of Exhibit as a data browser is this set of data from a large building project, which includes all the architectural 3-D CAD models and 2-D drawing sets. Another example is earthquake data from the state of Delaware where you can see it's being visualized on a map, but you have all the facets of the data such as magnitude, impact, and place which is important to folks in California.

And the point of this is that all data can be converted into linked data, and if you do so you get the benefit of these tools. Tools like Exhibit can support and visualize any kind of data, anything at all, as long as it can be converted into this linked data model. Anything can be converted into the linked data model, but it is not trivial to do that as we have learned over many years of doing it. And, one of the things we're missing are ontologies for how you would structure that data. To reflect back on Mike's talk there are no ontologies for library metadata today. I've looked. There are some emerging, but for some of the standards that we are quite used to, like Ferber, there really is no conventional ontology. So, another thing we're missing from our environment is a registry of ontologies or schemas for these kinds of standards that we all believe in and trust in and contribute to, so I am hoping that will emerge in the near future.

I hope that I've convinced you that data papers might make sense and that there is an emerging infrastructure that will allow us to publish them. Which brings me to the last topic, which is who is going to do all this work that allows for formal data publication on the web? This is not a trivial set of tasks we've assigned ourselves, and the answer is very complicated. There are many players involved in the scholarly communication and data correction ecosystem who will need to be involved in data publishing. Researchers are at the center as they always are, but there are components needed from all of the other stakeholders. I'll go through a few of these, but, among these are institutions who employ researchers and have some responsi-

bility for making sure the researchers are compliant with federal and, in particular, research grant requirements. You have funders who are making up a lot of the roles and trying to demonstrate the impact of their funding on the public good. You have data centers who have built up to professionally manage this data but may not have all this other infrastructure that I've talked about like some way of visualizing the data. They typically allow you to get to the data, but that's it. You have technology companies developing tools in this space. You have societies who are trying to kind of help researchers understand how to evolve, many of whom are closely related to publishers trying to figure out how data fits in. Some publishers are welcoming the chance to take the data along with the research, others are running as fast as they can the other way. Nobody is sure whose job it is to store and archive and manage this stuff over time. So, as you can see, there are lots of players involved who need to have roles in this, so I'm just going to talk about a few possible ones.

First of all, the researcher's role doesn't really change here. It is similar to traditional publishing in that they are responsible for creating the data or collecting the data in the first place and providing some of the metadata about it, like its structure, the methodology that produced it, what software was used, who else contributed, and things only they could know. Like traditional publications, researchers will need to be tapped for many of the editorial and peer-review functions of publishing. Only other researchers would presumably know if a data is methodologically valid. Other people like data centers may have the staff who could technically validate the data and make sure that it is complete and syntactically correct, but they would not necessarily know that the data was produced in a particularly good scientific way.

Next, we have publishers, and I would say a lot of societies too. There are several roles that I can see and maybe you'll think of others, as we saw publishers can put out data journals just as they have traditional journals, especially if the data is already available it is archived in a trustworthy repository.

Second, publishers can require data deposits—require is the key word there—into trustworthy archives. For example, in the life sciences, publishers for many years have required submitting genomic data into GenBank before they will publish a research article that refers to that gene. And more recently the Driad Project has enlisted all of the major evolutionary biology journal publishers to mandate data archiving as part of the publishing process into a trustworthy archive. So this role of requiring the deposit of data is a natural one that publishers can take on, although I understand that there are concerns about not adding more mandates to authors than necessary, so has to be something in the discipline has already kind of bought into.

A third role for societies and publishers is in the data accreditation area, organizing peer review and quality control required for usable data—not necessarily doing the selection and peer review but organizing that as they always have for traditional publications. But I do want to say, once again, that data has a very different intellectual property framework than we've had for traditional publications. For example, there is the fact that most data cannot be copyrighted. There is no law for copyrighting facts in the United States, so if the data is based on factual information, like sensor readings, it cannot be copyrighted, so we just cannot rely on the traditional mechanisms. We have to have new mechanisms to make sure that there is a sustainable business model for publishing data.

The role that many research libraries are exploring now is in the area of data curation, that is, collecting, cataloging, archiving, preserving, and providing access to the raw research, the primary research material, which is again a very natural extension of the role we've always had for primary research materials in all kinds of fields. Some libraries are also embracing the work of creating the new ontologies that I mentioned for data and working together with researchers to structure their data so that it is more interoperable. So this would be a library like the Oregon Health Sciences University Library which is creating ontologies for material in certain kinds of life science disciplines and then helping researchers figure out how to transcode their data into that new ontology. I think this is a really, really good role for libraries since they have a lot of that experience at structuring data from other areas.

Another natural role for libraries, of course, is outreach, education, and support to local researchers

who are struggling mightily with data management and what the best practices might be in their field. So, a couple of libraries I can think of are already required to sign off on every data management plan that is submitted to the NSF from that institution, because they may not understand the details of the data, they can see if the researcher has hit all of the points that the NSF asked for like a clear data usage policy and having it archived in a professional spot. So these sorts of data services are a very natural extension of traditional library work, but it does require that libraries get more involved in the research lifecycle than they have needed to traditionally.

Finally, we need technology companies and research institutions to help develop and support the web tools for interoperable data browsing, like Exhibit, and others that are out there. Without these tools, all of this linked data is no good: These tools are starting to emerge, but we really need to push them along and make them a priority, invest in them, get feedback, and use them so they become better—just like the web existed before Mosaic. The data web is out there, but it won't have any impact if we can't do useful things with that data. That is where we really need to put the effort next.

In closing, data papers have the potential to provide researchers with better incentives, methods, and credit for the data sharing they already know they need to do. But whether or not they become a commonplace part of the scholarly communication system, just having this idea of the data paper is really helping us understand how the changes in research affect us in the scholarly communication system and where we can most usefully invest more of our effort to address these changes and hopefully achieve the vision of a truly global interdisciplinary and large-scale data commerce.