Data Sharing Seminar Series for Societies

Researcher Perspectives on Improving Data Sharing and Reuse

6 August Chat Discussion

10:15:18 From Jake Yeston to Everyone:

metadata metadata metadata is our location, location, location mantra!!!

10:18:12 From Shelley Stall to Everyone:

Usability rating…love that.

10:19:19 From Shelley Stall to Everyone:

I really like getting these recommendations for how societies and repositories can help make this better.

10:19:34 From Pedro Luiz Pizzigatti Corrêa - Universidade de São Paulo (USP) to Everyone:

Great to know that Kaggle is reference to scientific repository.

10:19:52 From Yvette Seger (FASEB) to Everyone:

Agree - I appreciate the directed recommendations!

10:22:59 From Shelley Stall to Everyone:

Internal sharing leads to public sharing — this is really key!!

10:23:06 From Bruce Wilson (he/him, ORNL DAAC) to Everyone:

Even thinking about me using the data a year from now has been a good start in making data reusable by others. Future me has often been irritated with past me about data management practices.

10:23:41 From Shelley Stall to Everyone:

Hah! “Future me has often been irritated with Past Me!!”

10:23:56 From Shelley Stall to Everyone:

I resemble that remark.

10:26:52 From Bruce Wilson (he/him, ORNL DAAC) to Everyone:

Data management practices are something that’s largely tribal, in my observation, in the sense that we learn a lot of things about culture and practices from the cadre ahead of us. Culture and knowledge are both important.

10:27:40 From Shelley Stall to Everyone:

+1 Bruce Community norms are very hard to adjust from the outside.

10:28:00 From Shelley Stall to Everyone:

I think is why societies are really key…we can help work with the communities!!

10:28:34 From Sherry Lake to Everyone:

Please do not leave out Libraries.

10:28:54 From Joel Gershenfeld to Everyone:

It takes wrangling with deeply embedded (typically unstated) assumptions in communities (such as seeing data as propriety or emphasizing individual accomplishments).

10:29:28 From Shelley Stall to Everyone:

+1 Sherry!!! AGU is piloting a Society and Library effort. We are really excited about that!!

10:29:28 From Bruce Wilson (he/him, ORNL DAAC) to Everyone:

That which reason hath not put into a person's mind, reason cannot dislodge — C.S. Lewis

10:29:54 From Jake Carlson to Everyone:

One next step might be to engage repository managers in these discussions. As they are the ones curating and hosting the data, they have a large stake in being a part of these research and data workflows.

10:29:56 From Sherry Lake to Everyone:

+1 Shelley - thanks.

10:30:41 From Shelley Stall to Everyone:

+1 Joel! Your work is really important here!

10:31:25 From Sherry Lake to Everyone:

+1 Jake - Libraries offer a range of services across the research Life Cycle - training, curation, etc…

10:31:37 From Shelley Stall to Everyone:

+1 Jake Absolutely!

10:31:43 From Paul Guinnessy (AIP) to Everyone:

This is fascinating, but I'm wondering a little about the disconnect? i.e. I'm not sure how true it is in other fields, but the majority of papers published in physics in US journals are from abroad, i.e. not the society members of the society publishers. So how do we promote open shared data internationally?

10:31:44 From Ixchel Faniel, OCLC Research to Everyone:

+1 on Libraries. They are a key service provider for those who produce and those who reuse data, so see/know the needs of both.

10:31:45 From Janet Wyngaard to Everyone:

Career recognition - tenure track relevance - HUGE reason

Data Sharing Seminar Series for Societies

10:32:35 From Sheenah to Everyone:

Cite data and producers will be of enormous benefit to the SRR and cores that generate the majority of data on campus

10:34:37 From Shelley Stall to Everyone:

@Ixchel!!! I'm so glad you're here. Sarah included your researcher at the top of the session!! So important to this understanding.

10:35:04 From Ixchel Faniel, OCLC Research to Everyone:

@Shelly glad to attend. Loving this presentation.

10:35:48 From Suzanne Johnson to Everyone:

who is going to pay for this? are NIH and NSF prepared to pay for this and maintain the data?

10:35:54 From Shelley Stall to Everyone:

@Ixchel — me too. I feel like Societies are getting our marching orders. I'm SO excited.

10:35:55 From Bruce Wilson (he/him, ORNL DAAC) to Everyone:

Citation in the scholarly literature is important and useful, and there has been progress on this. Other uses are also important to track and that's an emergent field, I think (where libraries can really help). What is the value for my paper (or data) when it's used in an environmental impact assessment? That's often not captured in traditional bibliometrics. Data, in particular, often has applications outside scholarly literature.

10:36:24 From David Schultz to Everyone:

Sharing data can take a lot of effort from the scientists generating the data, and for large datasets it can occupy a lot of storage capacity. Although I value open sharing of data, does anyone have any quantitative measure that all this effort is worthwhile? How often are available datasets downloaded and used? I suspect that it is quite small, but I've not seen such evidence.

10:36:56 From Jake Carlson to Everyone:

+1 Bruce, we need better means of tracking use of published data both within and outside of academia

10:37:19 From Suzanne Johnson to Everyone:

What about the quality of data re-use? I am a psychologist and I can imagine problematic re-use of psychological data by non-psychologists. I would presume that is a problem in other areas as well

10:38:00 From Shelley Stall to Everyone:

@David The MakeDataCount.org folks are working on consistent measuring usage. An important element to that is if the data isn't documented, it won't be reused. A self-fulfilling outcome.

Data Sharing Seminar Series for Societies

10:38:02 From Richard Nakamura to Everyone:

Thank you!

10:38:06 From Jake Yeston to Everyone:

Question: did adopting electronic lab notebooks come up as a way of facilitating data deposition?

10:38:13 From Jeri Ropero-Miller, RTI to Everyone:

Since this 2015 mandate for data sharing and open access, how the landscape changed at all? Is it being monitored/evaluated? "Federally Funded Public Access Mandates."

https://researchguides.uic.edu/c.php?g=252224&p=2745823

10:38:27 From Bruce Wilson (he/him, ORNL DAAC) to Everyone:

@David — a hard problem. There is a very classic example of the benefit to making the Landsat data open. It's also an issue that the value of a given piece of data depends on the context around it — the positive application of what's described as the Mosaic Effect.

10:38:46 From Christina Chan-Park to Everyone:

@Suzanne, i think that's why it needs to be documented well. in the readme, it should also say not only how the data is used, but how it shouldn't be used.

10:39:07 From Jake Carlson to Everyone:

@David. Data Storage is a huge issue for data repositories as well. It's not clear how this will be accounted and paid for.

10:39:11 From Janet Wyngaard to Everyone:

Domains need to define standards/quality control/metrics... (no one else can really) But it's very hard to find domain experts interested/willing to do this given it's a rather thankless task- it's a service to your domain alas

10:39:35 From Zhong Liu to Everyone:

Producing datasets can be equally important to publishing research papers.

10:39:56 From Shelley Stall to Everyone:

+1 Zhong Liu !!

10:40:19 From Janet Wyngaard to Everyone:

+1 Jake - perhaps we need a survey of automation technologies available

10:40:21 From Bruce Wilson (he/him, ORNL DAAC) to Everyone:

The curation challenge is also hard. The value of any given bit of data over time is very domain-dependent.

Data Sharing Seminar Series for Societies

10:41:02 From Graham Smith to Everyone:

@David - it's an important question and I think there is a bit of an evidence gap here, but recent efforts (e.g. at RDA) are starting to focus more on tasks of data users, and how these are satisfied, where much focus to date has been on data producers and sharing.

10:41:04 From Shelley Stall to Everyone:

The FAIRsFAIR project funded by the European Commission has been a game-changer.

10:41:07 From Zhong Liu to Everyone:

You haven't addressed inter-agency and inter-organization data sharing and reuse challenges.

10:41:27 From Christina Chan-Park to Everyone:

@Jake, our repository has figured out how to get big data discoverable, and how to upload it, but the snag is that there is not an easy way to download it without timeouts. we were supposed to work on policies regarding paying for storage, but it's moot until we figure out how to download the data without using programming

10:42:06 From Zhong Liu to Everyone:

Funding agencies should include a data management plan in proposal activities.

10:42:26 From Surya Mallapragada to Everyone:

Thanks for a great talk. Did you see any marked differences between different disciplines in survey responses?

10:42:58 From Christina Chan-Park to Everyone:

at many institutions, the data repository managers are librarians

10:43:00 From Bruce Wilson (he/him, ORNL DAAC) to Everyone:

+1 Suyra’s question

10:43:02 From Brian Westra to Everyone:

This all takes a lot of effort, and I think there are a number of data management and data curation efforts and groups that could be leveraged in working on many of these issues.

10:43:07 From Xu, Zhihong to Everyone:

Is there any institution doing a good job in giving credit to researchers who share their data?

10:43:18 From Heleri Inno to Everyone:

you can always publish your data in data journal, quite new thing still, but a possibility

Data Sharing Seminar Series for Societies

10:44:13 From Sarah Nusser (ISU/UVA - she/her) to Everyone:

All, I was asked to share the link to the Holdren memo, which kicked off public access for articles and data in the US: https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

10:44:17 From Brian Westra to Everyone:

Part of the issue with tracking use/reuse is that publications also lack a standard approach to citing data.

10:44:20 From Jake Carlson to Everyone:

We have incorporated Globus into our repository (Deep Blue Data - https://deepblue.lib.umich.edu/data) that allows us to make large data sets accessible. It's the increasing demands on storage of the data (and covering costs) that concerns me.

10:44:48 From Carolyn Olson to Everyone:

@David, From a geologic perspective, well-documented data could be useful for many reasons - a couple ideas here: for additional regional extrapolation of a more site-specific study; as expenses rise exponentially for field work, existing data to use for other purposes than the original collection becomes invaluable. The likely reason that additional use of data sets seems small, is that as explained by the speakers today, if the metrics or metadata is lacking or less than what is needed, it therefore renders attempts to use the dataset worthless.

10:45:08 From Janet Wyngaard to Everyone:

@Xu, also like to know - is any institutional culture changing on this? Funders requiring this (great) but would accelerate things 10x radically if institutions valued this

10:46:54 From Howard Ratner to Everyone:

This group might be interested in the work being done by the Coleridge Initiative on Rich Context — the goal is to help US agencies use state-of-the-art methods to develop automated ways of finding out what datasets are being used to solve problems, what measures are being generated, and which researchers are the experts. Conference to come on October 20. Much more details to come! https://coleridgeinitiative.org/projects-and-research/rich-context/

10:46:58 From Richard Nakamura to Everyone:

Hove you run across informed consent problems in reuse of human subject data?

10:46:58 From Christina Chan-Park to Everyone:

@Jake, Globus doesn't currently work with dataverse for reasons beyond my technical understanding. we're hoping that they figure out a work around. but we're still figuring out how to have people pay extra for big storage so that they have the number upfront to put into their grants

10:47:06 From Brian Westra to Everyone:

It would be interesting to compare incorrect reuse of data vs. incorrect reuse of/interpretation of publications.

Data Sharing Seminar Series for Societies

10:47:31 From Bruce Wilson (he/him, ORNL DAAC) to Everyone:

@janet and @xu — there is an active effort at a couple of US National Laboratories, including ORNL, to go after “journal article equivalents”. How can we count all of the things researchers do besides traditional scholarly journals?

10:48:11 From Janet Wyngaard to Everyone:

@Bruce - sounds great, any public info on it?

10:48:14 From Jake Carlson to Everyone:

@Christina that's a bummer. Having Globus has been really important for us

10:48:58 From Sherry Lake (University of Virginia) to Everyone:

Re: Dataverse and Globus… it’s being investigated and “almost” there.

10:49:14 From Brian Westra to Everyone:

I think Univ. of British Columbia has done a proof of concept with Dataverse integration with Globus.

10:49:17 From Bruce Wilson (he/him, ORNL DAAC) to Everyone:

@Janet. No. It’s an ongoing discussion. There is also some discussion along those lines for Research Software in us-rse.org (research software engineers). In some ways, citing and recognizing software is a harder problem than citing and recognizing data (understanding that software is data, and also more than data).

10:49:49 From Shelley Stall to Everyone:

+1 Sherry - good to learn!

10:49:57 From Xu, Zhihong to Everyone:

Yes, it is hard to define “ journal article equivalents”. But it will be a trend I believe. Collaborative effort!

10:51:43 From Janet Wyngaard to Everyone:

Someday: we'll have a impact factor for data based on its reuse count- pretty sure it'll be far more reflective of real world value than our current journal impact factors :)

10:51:47 From Howard Ratner to Everyone:

Are data forward journals like Scientific Data or GigaScience helping or non-starters?

10:52:05 From Peter Schiffer to Everyone:

To add to the cost question: what fraction of repositories are funded in a sustainable way (i.e., not on grant funding or relying on the budget of the university/society supporting costs in perpetuity)? Data could be stored and then lost if the funding source for storage disappears. Should there be a minimum anticipated term of funding before a repository is used?

Data Sharing Seminar Series for Societies

10:53:06 From Sheenah to Everyone:

Multidisciplinary team science involving multiple data sources creates additional complexities for workflow and tracking data. Graduate students need instruction early - coincident with scientific rigor and reproducibility

10:53:30 From Shelley Stall to Everyone:

For my grant with Belmont Forum - where there are four counties being funded - we have to hand in FIVE data management plans updated each year. The one we have for BF is 20 pages long.

10:53:47 From Bonnie Nelson, RTI (USA) to Everyone:

Yes, long-term storage/accessibility on private/institute repositories is unlikely.

10:54:51 From Janet Wyngaard to Everyone:

@Howard

Data journals aren't non-starters - at least personally, I've just found them to be ~"non-intuitive"/feels like a forced approach to publishing data as opposed to having my data deposited natively in the known domain archive for my data type that gets the data FAIR and into the hands of relevant possible re-users.

10:55:27 From Sherry Lake (University of Virginia) to Everyone:

Yes, @Janet- exactly my take on Data journals.

10:56:22 From Sheenah to Everyone:

The first publication on team science was 1996!!🙂

10:56:29 From Brian Westra to Everyone:

From COGR news summary: 8/4/21: Dashboard will track hiring and promotion criteria (Nature) A US$1.2 million grant will fund an effort to identify and publicize the criteria that universities around the world use to hire and promote researchers. The Declaration on Research Assessment (DORA), a global initiative to reform the evaluation of researchers, will use part of the funds to create an interactive dashboard that will shine much-needed light on a process that is often opaque and controversial... https://www.nature.com/articles/d41586-021-02145-x

10:56:59 From Howard Ratner to Everyone:

@janet @Sherry Are there elements in those journals that could be extracted for more general use for data citation

10:59:02 From Janet Wyngaard to Everyone:

Ja...perhaps a marrying of elements of the data journals with tighter coupling with domain/other archives is the final solution we need

Data Sharing Seminar Series for Societies

10:59:49 From Janet Wyngaard to Everyone:

Eg many archivies won't give you a DOI and often metadata is missing -

10:59:56 From Rebecca Alvania to Everyone:

Perhaps journals transition from evaluating data for publication to evaluating data/metadata for depostion

11:00:18 From Janet Wyngaard to Everyone:

What Rebecca said!

11:00:28 From Graham Smith to Everyone:

A word on data journals (FD - I'm at Springer Nature), these are envisaged as just one part of data sharing/data publishing ecosystem, working alongside data repositories and increasing the profile of data - but more widespread policies and data sharing practice for research journals are a key factor

11:00:31 From Bonnie Nelson, RTI (USA) to Everyone:

The inclusion of Libraries will help with the move to data PIDs and metadata.