ISSN 2432-1249
SPARC Japan NewsLetter
NO. 27 Nov. 2015

The 2nd SPARC Japan Seminar 2015

"Towards the new paradigm of science and scholarly communication environment
— e-Science, research data sharing, and research data infrastructures —"

Wednesday, October 21, 2015: National Institute of Informatics
12th floor conference room (Attendees: 100)

During the past several years, research data management and sharing have become critically important issues to academic and research institutions as well as to those in the scholarly communication community. In the past, data that were generated in the research process belonged to the researchers themselves or the research community with which they were affiliated. But the rapid development and commoditization of information and communication technology from the 1990s on—first with the Internet and later with the spread of wireless and mobile technology—led to an exponential increase in the volume and range of data in various scientific fields and at the same time greatly facilitated the processing and sharing of such data. These changes in the data environment sparked predictions of a major paradigm shift in the way we carry out scientific research—the so-called "fourth paradigm," also referred to as e-science, data-driven science, and data-intensive science. Meanwhile, an increasing number of academic and research institutions were embracing open-access policies. While the initial focus of open access was research papers, the concept has influenced people's thinking on research data as well. Over the past few years, there has been a growing push for sharing of data in the name of "open data" and "open science." Today, we are grappling with new challenges posed by the sharing of research data as we move toward a new scientific paradigm.
This seminar was organized with the aim of fostering an essential understanding of the issues of research data management and sharing among such stakeholders as researchers, research administrators, engineers, publishers, government institutions, and library professionals who support research, as well as others with an interest in the new scientific paradigm, and to provide an opportunity for discussion about how to create a research support environment that meets scientists' needs.
A summary of the seminar is given below. See the SPARC Japan website (http://www.nii.ac.jp/sparc/en/event/2015/20151021en.html) for handouts and other details.

Part 1: Keynote Address

Open Data is not Enough
Mark Parsons
(Secretary General, Research Data Alliance)

Part 2: Science and Research Data Presentations

Design of Research Infrastructure and Utilization of Research Data for Breaking through "Research Barriers"
Asanobu Kitamoto
(National Institute of Informatics)

Inductively Think about Impacts of Open Platforms on Research
Daisuke Ikeda
(Department of Informatics, Kyushu University)

Research data sharing in the field of solar-terrestrial physics
Masahito Nosé
(Graduate School of Science, Kyoto University)

Panel Discussion
How ought the research data sharing to be?

Moderator: Hideaki Takeda
(National Institute of Informatics)
Panel members: Part 1 and Part 2 speakers
(see above)

TAKEDA: Who should be given credit for research data? How should data attribution be handled?

PARSONS: In some situations, data citation can provide an incentive for sharing research. Accountability is another important function. In one RDA project, they're trying to break research down into all its constituent roles for citation purposes. The motivation for data citation has changed.

KITAMOTO: Ideally, I think that the purpose of citation should be acknowledging credit. Right now the focus is just on data citation to facilitate reproducibility, probably because this suits the scientific journals, which are very influential. As long as one can meet the credit acknowledgement requirements by citing papers, people are not going to bother about detailed data citations. But how to assign credit for data is a big issue. I've proposed that each scientific community establish a "shoulders of giants" prize. The point is that a researcher's contribution should be judged by the degree to which he or she has created "giants' shoulders" for others to stand on—not just through research papers but also through data sharing and development of data infrastructure.

IKEDA: In his presentation, Mark mentioned a dynamic-data citation working group. I first heard about dynamic-data citation at the RDA Plenary Meeting last March. This would allow you to cite subsets of a dataset and credit the generator of that particular subset. Some people may not think that's necessary, but from a technical standpoint, it should be feasible.

TAKEDA: Citation issues could take us rather far afield. Let's go back to the simple question, "Who should be given the credit for research data"?

NOSÉ: My work involves both research and data management. In our field of research, data sharing has become a matter of course, and we haven't given much thought to credit. But I think it would be best to acknowledge the people who collect the data, as well as those who manage it and provide it, so that we could use it as a metric in assessing people's work.

TAKEDA: Given that the scope of data reuse is bound to expand as we go forward, it sounds like we should be moving in the direction of clearly acknowledging credit for research data.

NOSÉ: The reason I've personally become involved with the digital object identifier (DOI) system and so forth is that I'm interested in giving credit to the people who provide the data so it can be used as a metric for assessment. It seems to me that if we create a culture of data citation, it will make itself felt in a more equitable acknowledgment of credit.

TAKEDA: With the current technology, it's possible to display any number of credits. So, it seems to me that attaching full credit is the direction in which data sharing should be heading.

TAKEDA: Who should be involved in supporting data sharing, and how?

NOSÉ: Scientists in the domain have to be involved in the undertaking to some degree. If you don't understand the content, you can't provide good access to it. But it also requires people with broader expertise in the handling of big data [data curators].

PARSONS: I agree with Mr. Nosé. It can't be done without specialized subject knowledge.

KITAMOTO: And I agree that data curators are important. But I like the analogy of physical infrastructure, which requires expertise in architecture, civil engineering, and other fields. I would think construction of data infrastructure involves collaboration among comparable specialists, including experts in information science.

IKEDA: In terms of figuring out how to share research papers, there are now well-established institutional repositories and digital libraries like arXiv.org in various domains, but some of the disciplines that came late to the idea of digital

access are still struggling with it. I don't think it should be the researchers' job to figure out how to share their data. I've suggested that the process can be divided into two stages, with universities and research institutions being required to store the raw data, and specialists taking charge of curation.

ADACHI (floor): In Japan, the issues of preserving and sharing data are all tangled up with the problem of research misconduct. Is that true in the West as well? I'd also like to ask about the problem of protecting personal information with open databases.

PARSONS: I agree there's a need to separate these concerns. But there are complex educational issues and ethical boundaries that need to be considered when it comes to sharing data. Data citation can help with accountability, but it isn't a panacea. I think data sharing should be carried out at the organizational level in a centralized manner.

IKEDA: I think we can let the institutions handle those issues. Of course there are personal data and other kinds of data that can't be made open, but I think it's possible to use the power of information to distinguish between data that can be made open and that which can't be. At the same time, a completely open approach to science isn't really viable as a business model, is it? I think the reason institutional repositories have proliferated to this point is that we've left the decisions to the institutions.

ADACHI (floor): A database needs ongoing maintenance or it becomes obsolete. You can't just leave it to a curator. The successful databases are those that were developed organizationally and are centrally administered. I think Japan should have the infrastructure to support that. We can't ask individual researchers to devote their efforts to data maintenance.

TAKEDA: I think we're agreed that data sharing shouldn't be left to the researchers. I don't imagine there's just one answer. We should be aware that there's such a thing as professional data curation, and we need to actively tap into the resources of computer science. And I imagine we also need to clarify the role of the professional community or discipline and make sure everyone is working as a team.

TAKEDA: What form should data-use licensing take in these cases?

PARSONS: Data that are collected with public money should be viewed as a public good. In such cases, I prefer something like the Creative Commons Zero (CC0), which puts the data as much as possible in the public domain, within certain ethical constraints.

KITAMOTO: I think there's an infinite spectrum of possibilities, from closed to open, but since it's hard to deal with an infinite number of choices, we probably need to specify a finite number of models. As an option for licensing, I think CC0 is at the extreme open end of the spectrum.

IKEDA: My feeling is that access control systems are more important than licensing for open data.

NOSÉ: Where natural science data are concerned, I think it's clear that data collected with the support of public funds should be open. But we still need to acknowledge priority rights.

Part 3: Research data infrastructure of Japan

Presentations

Research data infrastructure of Japan
Takafumi Kato
(Japan Science and Technology Agency)

Database for upper atmospheric science ~Activity of the IUGONET project~
Yoshimasa Tanaka
(National Institute of Polar Research)

Sharing Data Sets as Research Resources
Keizo Oyama
(National Institute of Informatics)

Introductory Guide of Open Data for Administrative Staff
Nami Hoshiko
(Kyushu University Library)

Panel Discussion

What is the needs of researchers for the research data environment and how should we deal with?

Moderator: Kei Kurakawa
(National Institute of Informatics)
Panel members: Part 3 speakers (see above)

KURAKAWA: How do we go about developing our research data infrastructure? I'd like to hear your views from the standpoint of your respective domains, looking at the historical development and how user needs are changing.

KATO: Our efforts at DOI registration for research data thus far have focused on identification at a basic level. The metadata are also fairly simple and general, so a given domain may not find it that useful. In terms of our immediate goals, I think we want to develop something that can be used to link different types of data, such as papers and raw data, and data in different fields, while leaving the details of application to the domain-specific databases.

KURAKAWA: The Japan Science and Technology Agency's Japan Link Center [JaLC] could be considered a big infrastructure project, I think. I assume the registration of DOIs was expanded from papers to data in response to users' changing needs. Is that correct?

KATO: There was some talk of big infrastructure, but the main idea was to create a platform to ensure that information originating in Japan wouldn't be lost or overlooked. Meanwhile, with the emphasis shifting from papers to data, we've become aware that people need data identifiers so that they could track citations and use that data for evaluations. So, we're also working on the assignment of DOIs to research data to help facilitate quantitative assessment.

TANAKA: In the past, it was possible for scientists to write research papers just by analyzing data they collected individually, but today it's assumed that you get better results using a wide range of data to verify a phenomenon, and that's the prevailing style of research. That's why there's a growing demand for IUGONET.

OYAMA: In terms of the development of databases, there's been a huge change in scale and precision. In the past, the technical and cost constraints made it necessary to create well-organized data carefully. Nowadays, particularly with the rise of new statistical methods, the mainstream approach involves analyzing huge volumes of raw data from different angles in hopes of coming up with something. For another thing, research in information science used to have a narrow, technical focus, but nowadays there's more emphasis on research spanning different kinds of media or exploring the interaction between information and society. Human beings and society have become subjects of study for information scientists.

HOSHIKO: At the library, we've received queries about creating a public database from the University Research Administrator and the administration division, but at this time we don't have a good handle on the needs of researchers themselves.

KURAKAWA: What are some of the practical hurdles and considerations we should be aware of with regard to data management and sharing?

KATO: One problem is that validation of metadata hasn't made sufficient progress because there are so few use cases. Also, we want to make sure communication flows smoothly along all the various routes that have been established.

TANAKA: With regard to IUGONET, we worry about licensing and attribution. IUGONET itself isn't responsible for setting data sharing policies; each participating institution establishes its own data policies. As things stand, we're pretty much operating on the honor system, and there's no quantitative monitoring, so we feel some pressure to address that issue.

OYAMA: We make a big point of clearly explaining the conditions for use when people submit data.

HOSHIKO: Since we instituted a discovery service, there have been more requests for digital images of rare books and the like, and this has made us more conscious of the importance of good data management.

--Attendee feedback----------

(people affiliated with university libraries)
- Hearing what kinds of data are actually being handled, I could tell that the level of involvement varies a lot by discipline.
- I didn't understand everything that was said, but I was able to get a better sense of the state of the field and current trends. It was good to get a perspective on data sharing from people in the scientific community.

(university educator)
- It reminded me that there are important issues to be addressed, such as acknowledgment of data compilers and funding.

(university researcher)
- I was able to acquire some information in preparation for next year's RDA Plenary.

(other library staff)
- Since the context varies by field, it took a lot of effort to follow the direction of the discussion.
- I had wanted to hear about open data and open science from the researchers' standpoint, so it was very helpful.

(other university/research staff)
- I was able to get a good picture of the state of open data and some actual examples.

(others)
- I learned a lot about the RDA. I appreciated the topics and the way it was organized.

---------Afterword----------

For the past three years or so, the institutional repository movement has more or less plateaued, and just as I was thinking that research data might be the next big thing, I got a request from the SPARC Japan planning committee to help with this seminar. As a researcher specializing in digital libraries with a focus on author-name aggregation, I still feel out of my depth when it comes to the subject of scientific data curation. Scientific data take the form of spreadsheets full of things like observed variables and latent variables, and if one doesn't understand the model, the data are impossible to understand. When I dipped into some textbooks in the field in an effort gain a better understanding of those models, I had to go back and review my math, and I felt I was sinking even deeper. Unlike metadata for documents like books and articles, metadata for scientific data describe the specific models used in each field. The reason library personnel are easily able to handle metadata for document repositories is that the data described are in the form of books and other text-based documents, which are a librarian's field of expertise. I wonder if the time will come when we can package scientific data in such a way that professionals other than researchers in the field can manipulate it.

Kei Kurakawa
(National Institute of Informatics)

Planning and taking part in this seminar gave me an opportunity to think about some big issues of research data management that I have yet to incorporate into my day-to-day duties as a librarian. The keynote presentation by Mark Parsons offered a fascinating picture of the RDA's activities and future directions. The researchers' presentations provided easy-to-understand explanations of a wide range of research data along with specific examples of data management and sharing, to help bring the subject closer to home. I would like to thank everyone whose participation and cooperation helped make this seminar possible.

Nami Hoshiko
(Kyushu University Library)

It was very stimulating planning a seminar in collaboration with scientific researchers. My impression was that scientists have fairly low expectations of library professionals when it comes to the subject of open data and open science. But I think that we library professionals should participate actively in such discussions and work to put our libraries in the best possible position to support scientific research. I think this seminar provided an impetus for that.

Shigetoshi Kajiwara
(Hokkaido University Library)