The Research Data Alliance: Benefits and Challenges of Building a Community Organization

Francine Berman^{1,2,3} Merce Crosas⁴

¹Department of Computer Science, School of Science, Rensselaer Polytechnic Institute, Troy, New York, United States of America,

²Manning College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, Massachusetts, United States of America,

³Berkman Klein Center for Internet and Society, Harvard University, Cambridge, Massachusetts, United States of America,

⁴Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts, United States of America

Published on: Jan 31, 2020

DOI: <https://doi.org/10.1162/99608f92.5e126552>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

The Research Data Alliance (RDA) is a community-driven organization dedicated to the development and use of technical, social, and community infrastructure promoting data sharing and data-driven exploration. RDA is particularly important for the global academic community where research infrastructure is often ad hoc, may have a short shelf-life, and hard to fund.

At its launch in 2013, RDA struck a chord. Since then, RDA has attracted more than 9400 members from 130-plus countries and developed infrastructure used by groups all over the world. One of its founders is Francine Berman (FB), 2019–2020 Radcliffe Fellow at Harvard and the Edward P. Hamilton Distinguished Professor of Computer Science at Rensselaer Polytechnic Institute. Berman served as co-chair of RDA’s leadership council and chair of the U.S. region of RDA (RDA/US) for the first 6 years of RDA’s growth. Her leadership and organizational experience in developing and operating national-scale cyberinfrastructure (she is former director of the San Diego Supercomputer Center) and her broad community contributions have helped drive the success of RDA.

In this piece, Berman is interviewed by Mercè Crosas, Harvard University’s Research Data Officer with the Chief Information Officer’s leadership team and leader of Dataverse. Berman and Crosas met through RDA and both have deep knowledge of the benefits and challenges of running a community-focused organization. They share some of those insights in the conversation below. More information, including specifics on the RDA community and RDA’s ‘origin story,’ can be found [here](#).

Keywords: infrastructure, data sharing, interoperability, stewardship, research data, community organization, data management

Mercè Crosas (MC): You were part of the conception of RDA; why was it created?

Francine Berman (FB): RDA was created to accelerate the development and deployment of useful cyberinfrastructure that enables data sharing and data-driven exploration. Researchers (especially in academia) are often roadblocked by sparse or inadequate standards, models, common metadata, interoperability frameworks, etc., needed for their work. RDA was created to help remove those roadblocks.

MC: This is not a problem easy to solve by a single organization. What approach did RDA take from the beginning?

FB: RDA’s focus has always been on the pragmatic, rather than the ideal. Cyberinfrastructure developed by its working groups must be put into operation—‘adopted’—and solve someone’s problem, but RDA-developed infrastructure doesn’t have to solve everyone’s problem. The goal is to create and deploy infrastructure more

effectively as a means to an end. The hope was that RDA would provide a way to convene data professionals, researchers, and others to do useful work on specific problems, contributing to useful and usable global research data infrastructure.

MC: How did RDA get it started?

FB: RDA was actually conceptualized and catalyzed by astute funders in the U.S., EU, and Australia. They recognized the need for more enabling research infrastructure and the importance of encouraging the community to self-organize for this purpose. In the beginning, these funders identified an initial set of international leaders who could help found and jumpstart the organization. In the U.S., I was drawn to RDA's mission and excited to help actuate an effective organization. RDA's other seven founders and myself first got together by telecon in August 2012 and we got to work.

By October 2012, we had convened a group of 100 community leaders to discuss creating an effective organization for helping create useful research data infrastructure. Characteristic of RDA, the community used the October meeting both to envision the organization and to create groups to do the work. RDA was formally launched at Plenary 1 in Sweden in early 2013. Today, the RDA community has over 9,400 members from over 130 countries, all focused on data infrastructure. It has been, and continues to be, quite a ride.

MC: This can be considered a big success. What needs do you think RDA addressed?

FB: RDA struck a chord in the data community, particularly among academic researchers whose infrastructure may be sparse or inadequate. Part of the reason for this is systemic: Data infrastructure is often developed and utilized by individual researchers, specific projects, or domain communities to focus on particular problems and as a means to an end (new results or scholarly literature). Infrastructure maintenance, upgrade, and support may be inconsistent and deprioritized once the project is over, even though the infrastructure may still be useful to the team or other researchers. In some sense, this is an optimization strategy as academic researchers are incentivized to focus on new results, rather than on maintenance and support of infrastructure.

For the most part, the private sector cannot solve this problem. The 'market' of users for infrastructure may be very specific and/or small compared to the market for infrastructure products and services provided by the private sector, or the market for targeted open-source efforts for commonly used programs and systems. In particular, research data infrastructure may not find its way into larger communities where it can be supported and sustained. Yet without this infrastructure, modern data-driven research cannot move forward.

MC: Building a sustainable community around an infrastructure requires constant dedication, and the right timing, as we have experienced with Dataverse. Funding for data professionals is often challenging in the academic environment.

FB: Yes! Academic research infrastructure plays second fiddle to new exploration as a funding priority. It is hard to obtain R&D funding to maintain or improve infrastructure for the purpose of keeping it going or making it more useful to a larger user base. While new results advance the reputation of researchers and their institutions (often leading to greater opportunities and funding), developing or maintaining effective working infrastructure rarely has this outcome.

In higher ed, this generally translates into less recognition, resources, and job stability for the professionals who enable data-driven research. By bringing recognition to the importance of the development and use of research data infrastructure, RDA hoped to elevate the stature of data professionals in the typical university environment and beyond.

MC: When you and the other founders conceptualized RDA, how did you see the community coming together?

FB: In creating RDA's organizational structure, we borrowed liberally from organizations that had focused on infrastructure and/or had high community participation, particularly the IETF [Internet Engineering Task Force]. We believed that the heart of RDA needed to be its working groups—the groups that did the work that addressed infrastructure challenges. RDA would not dictate topics for these working groups. They would be 'bottom up,' i.e., self-organized: anyone that needed infrastructure to address a research roadblock could start or join a group. Shortly after the first plenary, we developed interest groups—groups that may study an issue before deciding which solution to invest effort in. This suited the kind of activity many of RDA's members saw as an important precursor to working groups.

RDA was initially conceptualized as a diverse, international, and community-driven organization by both its funders and its founders. In developing the organization, all of us had a strong commitment to creating this kind of a community. Each of the 8 international cofounders had had experiences with different organizations—some that had worked and ones that had not worked so well—and both good and bad organizational examples were valuable. Moreover, with RDA, we got to start from scratch. The experience of inventing RDA from the ground up was both thrilling and complex.

MC: What formal RDA organization did you end up with?

FB: The working groups and interest groups (today numbering more than 80) form RDA's core. Facilitating their work are an elected technical advisory board [TAB] (whose job is to vet and evaluate group charters and outputs), an organizational advisory board [OAB] (who represent dues-paying organizations within the RDA, focusing on organizational needs, infrastructure adoption, and business advice), and RDA's council (which is responsible for the overarching strategy and organizational health of RDA and plays the role of RDA's 'board').

Critical to the organization is RDA's secretariat, its administrative and operational arm, led by a secretary general, nominally the 'CEO' of the organization. Decision making in all groups is largely by consensus and

community-driven. All groups follow RDA’s principles, which include openness, balanced representation, harmonization across diverse infrastructures, and nonpromotion of particular technologies and products.

RDA was organized with agility in mind. We worked hard to determine best practices and were open to variance, evolution, and change if the initial organizational structures and approaches didn’t work as desired. For this reason, RDA’s “Governance Document” describing the organization is a living document and has evolved over time by RDA’s council and membership based on how things are working out.

MC: What do you think distinguishes the RDA community from other communities?

FB: I’ve always thought that something that both distinguishes RDA and that RDA really got right is its culture. Inclusion and diversity of its membership and throughout its decision-making processes are tremendously important in RDA and are ‘baked in’ to RDA’s leadership, selection processes, and meetings.

RDA is serious about gender diversity: There are essentially half women in all leadership positions (council, technical advisory board, organizational advisory board, secretariat, working and interest group co-chairs). Beyond gender, RDA is intergenerational (including people in various levels of leadership from all stages of their careers), ethnically diverse, domain-diverse, and has members from all over the world. RDA is also professionally diverse with members from all sectors, although academia is the most strongly represented (and perhaps represents the greatest need for data infrastructure development and support).

A good example of RDA’s commitment to diversity is the TAB election process, which is predicated on an algorithm that takes votes from the community and ensures that elected TAB members span multiple sectors and regions. This isn’t just an open-minded thing to do, it also ensures that TAB can reasonably assess new working group charters from across the board.

Diversity and inclusion are also goals of RDA’s twice-yearly plenary meeting agendas. RDA seeks to avoid ‘manels’ (panels with all men) and all male keynoters (or all keynoters in a particular area). In a real way, RDA’s focus on diversity is thoroughly embedded in the organization’s culture and new members have found RDA to be a particularly welcoming environment for data professionals of all types.

MC: What other characteristics of RDA are noteworthy?

FB: Beyond culture, RDA has distinguished itself by its focus on pragmatic action, openness, and collaboration in a crowded landscape of community organizations. Its focus on practical, timely, and ‘adoptable’ solutions that address specific data infrastructure requirements make it useful to specific groups, and when possible, to a broader set of research domains and communities that can leverage or expand RDA infrastructure solutions.

RDA is also serious about collaboration. Its twice-yearly plenaries provide a “neutral space” that attract other organizations interested in infrastructure. Many of these groups co-sponsor RDA working groups or interest

groups and hold side-meetings at RDA. A few years ago, CODATA [Committee on Data of the International Science Council] and the World Data System worked with RDA to co-locate meetings every two years or so, creating an “International Data Week” for members from all three groups and beyond. The first International Data Week was in the U.S. in Denver in 2016 and the second was in Gabarone, Botswana, in 2018. These meetings attracted a very broad audience and numerous organizations and side meetings. A great outcome was the development of synergistic connections and additional productive collaborations between multiple organizations. On RDA’s side, this increased the number of jointly co-sponsored working and interest groups and introduced new topics of interest to multiple organizations.

MC: Indeed, I have benefited from RDA’s collaborative approach by participating in a CODATA and RDA effort on research data management in institutions. With these collaborations everybody wins. . .

FB: I agree. This kind of strong peer interaction between RDA and other groups is also ‘baked in’ to RDA’s culture. I have always appreciated that RDA’s collaborations with external groups focus on what is really important—useful data infrastructure that supports the community—rather than ‘world domination’ or exclusive partnerships that might make it easier for RDA address its financial sustainability challenges but less open to serving its community.

I have also really enjoyed RDA’s breadth. Each RDA plenary is a fantastic *mélange* of data scientists, data professionals, domain scientists, journalists, policymakers, students, funders, and others interested in data. The energy and perspective of these diverse members and the opportunity to learn different approaches to challenging problems makes RDA meetings particularly dynamic and exciting.

MC: But does such a breadth also impose some organizational challenges?

FB: You bet! RDA’s breadth is both a benefit and a liability. When an organization is so broad and diverse, it makes it hard to articulate what success means. For some of our funders, success means that RDA has developed infrastructure that can become standards and incorporated into policy; other funders want to see a strong uptake in organizational membership and engagement as represented by RDA linkage in proposals and publications. For some RDA members, success means removing a research roadblock by creating infrastructure that can be deployed in their lab or on their project. Since RDA serves so many constituencies and can be measured by quite different notions of success, it is hard to target sustainability and other efforts to a single objective.

This also makes it hard to measure impact for RDA. Should we measure the amount of infrastructure? Its adoption and usage? The number of members and their demographics? This problem is exacerbated because as an international and multifocused organization, RDA encompasses a highly diverse set of political and professional cultures and agendas. Harmonizing those for impact, sustainability, and distinct stakeholders is hard.

MC: What kinds of data infrastructure has RDA generated or helped generate?

FB: Many kinds. There was a recognition by RDA, even at the beginning, that needed infrastructure will take many forms—code, models, practice, policy, common approaches—and indeed it has. RDA’s *Wheat Data Interoperability Working Group* developed a common vocabulary for the community to help harmonize diverse agricultural data sets. The *Data Type Registries Working Group* developed a model and expression for data types that could be read by machines as well as humans, enabling both to parse and understand the semantics, context, and assumptions behind data. RDA’s *Libraries for Research Data Interest Group* developed a set of best practices, online resources, and tools to incorporate data management into library practice. They published them in the highly popular “23 Things: Libraries for Research Data” document that has been translated into 12 languages and used by librarians around the world.

RDA’s infrastructure is as diverse as the organization. Even though there is a fairly broad spectrum of infrastructure types, each infrastructure output solves someone’s infrastructure problem by design. The flexibility and agility that we tried to build into the organization is directly related to the broad scope and different kinds of infrastructure produced by RDA’s members.

MC: RDA brings multiple disciplines together. Do you have a cross-disciplinary example that integrates with the humanities?

FB: One of my favorite groups has been the digital ethnographers who found RDA to be a great venue for developing and leveraging infrastructure that supports historical and ethnographic research for anthropologists and humanists. They created the *RDA Digital Practices in History and Ethnography Interest Group* to explore and discover critical infrastructure components and mechanisms—metadata standards for researcher-created primary data (e.g., field notes and formal interviews), citation practices, digital exhibition protocols, etc.

The co-chairs told me that RDA was useful for them because they could learn and use best practices for their data without having to reinvent the wheel, and doing their work in the context of RDA helped them create specific and needed infrastructure in collaboration with experts from around the globe. This is exactly the kind of role RDA wants to and should play, but it also illustrates how the benefits of RDA are often intangible and hard to capture as impact objectives.

MC: Following our work originated by Force11 (another community organization focused on improving knowledge creation and sharing) on data citation principles, there was another successful output from RDA. . .

FB: Yes, the *Data Citation Working Group*’s infrastructure has gotten a lot of uptake and its outputs include a model and protocols for citing data sets that change dynamically over time. The group came up with an approach to ‘timestamp’ evolving data sets so they could be appropriately cited in publications and used correctly for reproducible research. It’s an interesting case because the group’s ideas were first peer-reviewed

and published in the scholarly literature. They then used RDA as a venue to develop and disseminate their approach. To date, their model and protocols have been used by researchers in forestry, oceanography, biomedical sciences (electronic health records), environmental science, and a variety of other domains.

MC: How does the RDA community relate to data science? And how can it help data science?

FB: RDA is an organization strongly related to the broad and varied aspects of data science. Many of its groups focus on the use of data in domain research and the stewardship and preservation of data, however, all topics in RDA are ‘bottom up,’ i.e., suggested and created by interested members. In essence, everything in data science related to infrastructure is fair game.

Data science is a broad field. Many people think of data science as primarily machine learning or statistics. There are many facets to data science: machine learning and statistics to be sure, as well as the challenges of organizing and storing data, visualizing and communicating about data, computing with data, stewardship and preservation of data, and the linkages between data science and the broader world: policies and regulation about the use of data, ethics, business analytics, and the use of data in the commercial sector, experimental design, and research.

Interestingly, machine learning and statistics are not currently well-represented in RDA. RDA would welcome new members and groups in these areas, and I believe the interaction between RDA’s current efforts and more AI and statistics-focused areas would provide a great opportunity for synergy and innovation.

MC: How can the reader learn more about this broad definition of data science, which I think is key to data science success?

FB: Rob Rutenbar and I led a study for NSF [National Science Foundation] on “Realizing the Potential of Data Science” [<https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf>] to map out the next decade of opportunities in data science research and education for NSF planning purposes. We used this broader definition and talked about research, curriculum, infrastructure, and administrative challenges for universities helping to evolve and advance this fundamentally important area.

MC: What lessons could other scientific or data organizations learn from RDA?

FB: There are many things that RDA really got right and also many things that RDA’s leadership and community are still working on. RDA’s focus on a culture of diversity and inclusion has really worked and extends the notion of diversity beyond gender and ethnicity to geography, age, native language, profession, sector, and other things.

Another thing that RDA has done right is to focus on peer collaboration. RDA has worked hard to partner with other organizations successfully to get things done. Many organizations, including CODATA, WDS [World Data System], Force11, ESIP [Earth Science Information Partners], AGU [American Geophysical Union], and

others have worked with RDA and many more have used RDA plenaries as a ‘value added’ environment for collaboration and synergistic meetings.

A third thing RDA has done right is to keep the focus on impact, through infrastructure adoption programs, early career programs, prestandardization efforts, open dissemination of infrastructure, etc. These efforts have helped RDA maximize the effectiveness of its efforts.

MC: As chair of RDA/US, what was done in particular in the U.S. that could be reused by other organizations?

FB: In the U.S., one thing we’ve done well is to focus on enabling and empowering U.S. leaders to benefit from and participate in the RDA. Funding from the NSF, MacArthur, NIST [National Institute of Standards and Technology], and the Alfred P. Sloan Foundation has been used to support RDA plenaries in the U.S., pilot RDA programs for early career professionals and infrastructure adoption, support working group chair meetings, and provide participant support for U.S. members who stepped up for leadership in the international organization and need to attend RDA’s international conferences to do their work.

Participant support is perhaps a small thing, but international travel is expensive and this support has been fundamental, enabling RDA’s U.S. members (especially those who don’t have institutional funding) to take full advantage of, and fully participate in RDA. This means that U.S. working group and interest group co-chairs and U.S. members of TAB, council, and the secretariat, as well as OAB co-chairs can show up at plenaries, provide onsite leadership, and do their work.

MC: What can be improved?

FB: There are many areas where RDA can improve: The business model hasn’t kept up with RDA’s growth and the challenge of providing adequate administrative support for an organization with 80-plus groups and 9,400-plus members has become critical. In particular, RDA needs a different business model than the one it was ‘born’ with (i.e., initial funding from visionary program officers who felt personally committed to its mission) to sustain it in the present.

The initial conceptualization of RDA as an open and low-barrier-to-access organization has made this harder. RDA was initially developed to be free to members (so that membership dues that would not disadvantage one group or geographical region over another), have modest plenary registration (to keep costs down for members from different economic environments), and to rely largely on volunteer efforts.

Along the way, we created a paid position for the secretary general but still have in-kind and volunteer efforts in the secretariat—fractional efforts of 9 or so people that only amounts to roughly 2.5 full-time FTEs, excluding the secretary general. It is hard to support adequate effort for business development, all of RDA’s legal, accounting, and other business-focused services must be outsourced, and RDA’s bank account typically

contains little more than the salary of the secretary general. As a business, RDA continues to struggle; as a community-driven organization, RDA has been a great success. The organization is seriously dealing with the conflict between these two realities and its resolution must be part of its future.

RDA is also challenged to successfully expand beyond its current membership. Engaging new international regions is hard because the successful cocktail of regional leadership, regional financial support, and a substantial regional community—the success strategy that’s largely been used in the U.S., EU, and Australia—is hard to jumpstart.

MC: Finally, what do you think should be the next focus for RDA?

FB: I think there are important opportunities that I’d love to see RDA embrace as it goes forward. Artificial intelligence, autonomous systems, social networks, and the Internet of Things all run on data and require considerable infrastructure (both social and technical) to ensure that data is shared appropriately, privately, and ethically. What data standards, strategies, and infrastructure are needed for these environments? How do we support reproducible data-driven research in the highly decentralized and nontransparent environments we are encountering? As research topics and techniques evolve, it is important for RDA to be prepared to create and deploy the data infrastructure needed for both current environments and the future research scenarios we are likely to encounter.

I’d also love to see RDA have more of an impact on the academic organizations that support data-driven research. Dataverse’s participation in the RDA has been important because Dataverse represents a serious effort to host and manage the data of academic communities. It represents a real model for institutional commitment and data community influence. It’s been important for Dataverse leadership to participate in the RDA: you all serve as exemplars of approaches and strategies that can be helpful for RDA membership, and hopefully RDA provides a source of new ideas and collaborations for Dataverse.

The same thing is happening with academic librarians. A number of librarians in academic environments have brought important strategies and problems to RDA and have brought back new practices and collaborations to their home institutions. Active participation in RDA by those linked closely with institutional data cultures can help improve the data environments where we work and support academic data science more broadly. I would love to see more of this in the future.

MC: Many thanks, Fran, for providing such an insightful account of RDA, and most importantly for helping to create and lead it. I look forward to seeing its further growth, which can benefit all of us and the data science community in general.

Disclosure Statement

Mercè Crosas and Francine Berman have no financial or non-financial disclosures to share for this interview.

©2020 Mercè Crosas and Francine Berman. This interview is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the interview.