July 2018

# “If we build it they will come” – Ways of User Involvement in Information Infrastructure Development

*Report on a co-located event at the 11th RDA Plenary, March 23rd, 2018*

## INTRODUCTION

At the 11th Research Data Alliance Plenary Meeting, the German Council for Scientific Information Infrastructures (RfII) organized a co-located event to discuss practical ways of obtaining successful user involvement in the development of research data infrastructures. Speaker Lars Bernard (TU Dresden and RfII member) introduced the event, briefly presenting the background to the RfII and summarized the recent RfII recommendations to establish a National Research Data Infrastructure (NFDI) in Germany. The RfII has been effective in convincing German stakeholders that in facing the ever-growing amount and complexity of research data a common concept in designing future scientific information infrastructures that are sustainable, user friendly across disciplines, and built on the benefits and experiences of existing organizational and technical structures is crucial. Meanwhile, future approaches should avoid the pitfall of adopting any solution that is not founded on the practices of scientific communities themselves. Thus, in addition to the necessary shift from the existing mostly project-based funding model to a sustainable and reliable funding scheme, the fundamental challenge is to successfully design and implement the user-driven development and operation of research data infrastructures. The attendees discussed this challenge from three different perspectives: 1) the role of policy actors and funders, 2) the role of infrastructure providers, and 3) the role of users and research communities.


## THE ROLE OF POLICY ACTORS AND FUNDERS

The first discussion was held among a panel of twelve participants and was moderated by the U.S. National Science Foundation Program Director, Amy Walton. A set of key questions was provided to the group to stimulate their dialogue, starting with how policy actors and funders can incentivize user involvement. The group initially sought to determine the meaning behind the word “user”. Group members mentioned the importance of not limiting the definition merely to

scientists but to include the research community as a whole in order to maintain a sustainable information infrastructure. In terms of incentives, the group pointed out the effectiveness of using metrics to assess existing infrastructures. The first idea was that there needs to be a key measure of success. Policy actors, particularly when faced with a constrained funding environment, should have a clear way of evaluating services. Through the use of metrics, the actors would be better able to decide which activities to continue. A key component on the topic of metrics was the notion that metrics should be derived from and clear to all stakeholders. Metrics in themselves are not uncommon, but transparent metrics that clearly state how they are measuring success, and that have been determined through input from all stakeholders, may be more helpful as a yardstick. Additionally, the group countered the question by asking whether users should be incentivized *not* to create infrastructures on their own. There was a wide consensus among the members that it is more effective to build on existing infrastructures to ensure sustained and long-term capability.

The second topic was centered on concrete actions funders can take to foster user involvement in infrastructure development. The group began its analysis, once again, by asking a fundamental question: "what is infrastructure?" The discussion highlighted that infrastructure is multidimensional as it can serve both small research communities and well as very large ones. Therefore, it is important to ask *what* is being developed in addition to *why* it is being developed when building infrastructures. Furthermore, the group argued that fostering user involvement should not only be done on a national level, but also on an institutional level. It emphasized the need for an institutionally-based reward system that could potentially accelerate the adoption process of common data management practices and of RDM services among researchers. The participants also discussed how long the process of project development can be, which can particularly be the case in IT, where necessarily long lead-in times exist, whereby academics and research teams may have a preference for quick solutions. The group concluded that speeding up the evaluation, testing, and utilization of project proposals warrants wider participation from a broader community.

The final discussion point was on how funding policy can build functional interdisciplinary information infrastructures and avoid the "siloing" of scientific disciplines. The participants delved into the topic by first asking: "what is the value of interdisciplinary work?" They considered whether the motive is merely to have a count of diverse disciplines or rather to develop systems that are interoperable and allow people genuinely to answer their research questions by combining tools, data and services from various disciplines. In the process of infrastructure project evaluation, especially those elements or benefits should be considered which only an

interdisciplinary approach could add to a regular disciplinary approach. Interdisciplinarity can unleash its potential when a different community evaluates, reviews, or comments on one's infrastructure. Through such reviews, collaboration across disciplines could be triggered, leading to new technologies, tools, and best practices to target increasingly complex and multidimensional research questions in the future.

## THE ROLE OF INFRASTRUCTURE PROVIDERS

This perspective was discussed by two sub-groups: the results of the first group were presented by Laurel L. Haak, the Executive Director of ORCID, and those from the second group by Ralph Müller-Pfefferkorn from the Technische Universität Dresden. The initial point of discussion dealt with the best practices and practical problems associated with user involvement in infrastructure development from the provider's perspective. The first group began by clearly defining, on the one hand, users as the researchers, scientists, that is, people with data, as well as interested citizens; and, on the other, infrastructure as the data platforms and data repositories utilized by such groups. The team discussed how the lack of a secure login system in any infrastructure can create a general unwillingness on the part of users to input data. To address this issue, this group recommended the adoption of an authenticated login system built into data repositories to help identify users, increase collaboration, and help respond to their needs. Such a system could also help identify who the data creators and curators were, to enable citation and potentially also encourage more researchers to share their data. The group highlighted that this system would be particularly useful for those scientists who frequently have to manage sensitive data and who are thus in need of stronger management protocols regarding access to data.

The group went on to address whether research data literacy and culture should be established independently by users or rather by infrastructure providers. The group stated that scientists are often reluctant to use infrastructures due to a lack of knowledge on how to optimally deploy the platforms on offer. The team proposed that training and curricula for researchers should be an intrinsic component of infrastructure development. The example of the German Federation for Biological Data – GFBio was brought up to respond to the question of how users can be integrated into the development models of infrastructure providers. This particular platform simultaneously provides data from several collections and allows individual researchers to share their data. The group emphasized that implementing data sharing as a critical aspect of the profession can act to motivate researchers to actually practice data management and data re-

use. Furthermore the group underlined the benefits of establishing research data management as a profession – to create a cadre of professionals amongst researchers themselves who can act as liaisons between the users of the tools and the owners of such tools. In terms of data management, it also became clear that there is potential benefit in terms of user engagement in infrastructures where there is systematic adoption of creating actual data management plans (DMPs). It is common to have a lack of a verification process even though data management practices are well established. To address this, the group mentioned the potential benefit in having a data management plan scheme that would not only cultivate harmonization across the user community but would also allow for real follow-up on actual data practices on the part of research infrastructure providers, to be able to refine services.

The other sub-group commenced its dialogue on current practices and practical problems by acknowledging the need for collaboration within the international community. The group emphasized that research communities have no borders and are in need of international integration. One problem that exists, however, is the difficulty of identifying the demands and requirements of the users in any disciplinary community. The group proposed that this obstacle can be overcome through the completion of a stakeholder analysis. Representative users have to be found, so that providers and the research community in question can together identify user types, typical workflows, data types, and so on. Possible partners for the providers in this analysis could be the professional societies/organizations of the user communities. The group emphasized that providers should take into account not only large communities with large data sets but also the small, and often forgotten, long tail communities.

The team further explored the ways in which users and their requirements can be integrated into the development processes of infrastructures. The development model should be user-centric from the outset. In order to involve users from the start, it is essential to develop a cyclic or iterative process that enables better communication, interaction, and coordination between users and the providers. Additionally, it would be beneficial to evaluate and learn from projects that have failed in the past. The implementation of a user advisory board was also mentioned as a method to ensure the representation of a community and its requirements in the development process.

Another part of the discussion was centered on the dynamically changing user demands for services and technologies. The group noted that the needs of individual researchers do not change that quickly but instead move in a longer-term, relatively predictable arc. However, there are two types of changes that lead to a need to adapt services and technology: (a) An extension of the

services e.g. due to more instruments of the same type or larger experiments that provide more data but do not change the research workflows significantly. Such an extension can be satisfied with a scaling of the services and hardware. (b) Innovations in research workflows caused by new research infrastructures (e.g. new instruments) or new methods (e.g. machine learning for data analysis). This results in a need to adapt the information infrastructures or even develop new ones. Besides the group-specific changes, providers also need to take into account the individual researcher needs. Users tend to think about data management only in terms of their individual work. Their focus is on research and the related research workflows – not on data management. They are less willing to use an infrastructure when there is no clear incentive or advantage to do so. Providers can establish this incentive factor by tailoring their platforms to the individual needs of the users in terms of methodology.

On a final note, the group touched on the topics of research data management literacy and research data culture. Even if infrastructures or tools can help to "mask" necessary literacy, basic knowledge is still needed (e.g. to know what metadata are and why they are required). As in the case of the first group, it was acknowledged that information infrastructure providers should train researchers in the actual usage of their infrastructures. The team further highlighted the need for more expertise on cultural change in research data handling. For many research communities, the globalization of information infrastructures is somewhat of a cultural shift. The group recommended that cultural experts from the fields of psychology and sociology be included in order to accompany and support users in this process.

## THE ROLE OF USERS AND RESEARCH COMMUNITIES

This subject was also discussed among two different groups. On behalf of the first group, CLARIN representative Darja Fišer began by analyzing how users can best get organized to shape the service portfolios they need. Research communities would benefit from a balance of top-down and bottom-up approaches through which solutions are not enforced on users while, at the same time, not expecting them to develop best practices on their own. From a top-down perspective, providers should offer training and support on how to deposit data into the infrastructure and work with the services and tools available. They should offer insights on how users make use of the infrastructure which would help to align services with the users' needs, especially in early stages of infrastructure development. In terms of a bottom-up approach, three successful models have been singled out: ambassadors, data champions, and collaborative research. Individual researchers who are enthusiastic about

infrastructures can add an invaluable contribution to the process by acting as ambassadors in their community and recruit more users. Additionally, sufficient resources should be allocated to data champions who can visit universities and research communities to advocate for, showcase, and teach how to use the infrastructure, tools, and data sets that have been collected. Lastly, collaborative research networks who already recognize the benefits of data sharing can act as a catalyzing group for other researchers who need to match interest with more concrete knowledge.

The group discussion progressed towards whether the definition of data quality criteria can be a means to promote user participation in service development. The team agreed that there is a continuum in data quality and that it is important to demonstrate to researchers why they should adhere to certain standards. This demonstration can be done practically by listening to problem-solving practitioners and involving researchers equipped with the knowledge of developing and refining best practices in their field. The team also highlighted the need to invest in building consensus in the research community. They explained that building consensus should go beyond national borders and the smaller, homogeneous communities; international and heterogeneous research communities should be promoted so that researchers are cognizant of further possibilities. Consensus-building can be achieved not only by promoting the advocates and early adopters of research infrastructures but also listening to the more skeptical and critical researchers.

Lastly, the dialogue considered the "chicken and egg" issue of whether users should independently develop the research data culture or if this should result from the services offered by the infrastructure providers. The group redirected this question towards the "carrot and stick" metaphor. Researchers should not only be asked to create reproducible science and share their data, they should also be rewarded for doing so. In terms of interaction with users, this should be a continuous process throughout infrastructure development, as opposed to merely at the initial stages, so that infrastructure providers can get continuous feedback and can constantly improve the infrastructure based on user feedback. To conclude, the group emphasized a key takeaway message from their discussion by stating that "once you've convinced the users, they will stay."

On behalf of the second 'Users and communities' group, Frank Oliver Glöckner from the Max Planck Institute for Marine Microbiology and Jacobs University Bremen, debated the best approach to create inroads into user communities. The group, which was mostly constituted of infrastructure actors, commented on the lack of self-organization among scientists and their

(perceived) unwillingness to bring their views to bear on the development of research data infrastructures. In the discussion two main reasons for this overall environment were articulated, in order to find a means of realizing the enormous potential of user communities' participation here:

1. Scientists are reluctant since there are currently no incentives for proper research data management, such as greater reputation impacts or scientific credits for their career.

2. Potential users are often frustrated by the high entry barriers involved, the low technological status of tools and the lack of support scientific infrastructures provide. They rather use commercial alternatives such as Google and Dropbox that provide innovative features and are easy to use.

In terms of participation in the development of services supporting researchers, the group indicated that users are often simply unaware of the existence of appropriate research infrastructure. For instance, libraries may hold extensive resources on research data management (RDM) but users may not know that these exist. Compounding this, users might even mistrust their local infrastructures in terms of quality of service provided and long-term stability. As a recommendation, cooperative structures should be established between the users and the infrastructures to address the need for stronger communication and interaction.

The discussion also touched on the role of institution in fostering a research data culture. In general, a research data management friendly environment would help, but the group emphasized that it is rather the culture in their respective field of science, not so much the institution, that steers researchers. Therefore, actions taken to improve this situation should focus on establishing science ambassadors who can showcase the importance and potential benefits of research data management to their own communities.

The final point of the discussion was centered on how best to proceed in integrating users in the operational models of infrastructure providers, namely the proposed process for the National Research Data Infrastructure (NFDI) in Germany. The concept to date has relied on the notion that scientific communities should be directly involved in building service portfolios. Unfortunately, not every community is represented by a particular scientific or scholarly society; and even those with representation tend to give voice to multiple views, given the dynamics of research practice. To address this issue, the NFDI consortia proposals should clearly state the existing procedures and mechanisms in community building and user engagement. The group further suggested that the proposals could be evaluated by users with a standing community. Ultimately, such an evaluation process will allow for concrete, rather than mostly conceptual, conclusions to be reached regarding the types of infrastructures needed.

## CONCLUSION

Although the participants analyzed the issue of user involvement from three different perspectives, there were clear-cut common points that arose from the separate but parallel discussions.

- Nearly every group recognized the **need to incentivize users** as a specific action to increase their motivation and participation within the realm of infrastructure development and operation. From the policy actor's point of view, incentivizing users involves establishing key measures of success whereas from the provider's perspective, it means integrating users by making data sharing a vital part of their profession, ideally a daily exercise and a key principle of their scientific and scholarly education and culture. The users and research communities themselves should advocate for a change in the academic reputation system that acknowledges research data management as an important contribution to the advancement of science and scholarship. As such, these claims are not new, however, being repeated here underlines the point that an operationally mature incentive system for research data sharing and re-use is still far from reality.

- The groups also acknowledged the **multidimensional nature of user involvement**. Infrastructure providers should consider that researchers come not only from large communities with large data sets but also from small, and often overlooked, long tail communities when tailoring to researchers needs. Unsurprisingly, the provision of research data infrastructures needs also to encompass a constant process of updating and innovating services, to keep pace with (a) the general, rather rapid, innovation cycles in information system technologies but also (b) with the slower evolutionary changes in research tools and methods. Thus, information infrastructures providers are challenged to design smart maintenance and updating strategies to create a bridge between the provision of reliable and sustainable services and offering services which are both near or at state of the art *and* up to specific tasks in current research.

- Another common point among the groups involved the significance of **incorporating the training of researchers in how to deploy infrastructures in their research operation**. It was highlighted that users should not only be trained in how to deposit data and how to work with particular services, but also in how to further develop the infrastructure itself following an open-door principle regarding innovations and extensions to services. Users can also frequently be unaware of the services available to them. Thus, providers should establish a clear communication strategy to showcase their infrastructures as well information and training on how to use them.

- All panels recognized the **importance of advocating greater research data management and encouraging researchers to use existing services**. In order to ensure sustained and long-term capacity it might be even helpful *not* to incentivize further micro research infrastructures but to emphatically reward the research community where it demonstrably re-uses data and shows how research has benefited from existing services.

- Furthermore, several groups supported the idea of promoting data champions among researchers within a specific scientific or scholarly community. These agents would act not only as early adopters but also as mediators or ambassadors for their communities and serve as liaisons and communication channels between providers and users. Meanwhile, the process of establishing sustainable research data infrastructure has to avoid privileging pioneers only and needs to be inclusive by also carefully listening to skeptical voices and by building on a wide consensus.

What remains uncertain, however, is how consensus-building among scientific users could best be organized regarding the services that would eventually most appropriately fit their needs. The same holds for the role of the data champions and the optimal way forward to recruit them and make them part of consensus-driven requirements gathering or engineering infrastructure development processes. In other words: Although participation and integration of users into infrastructure operations is a sensible strategy, the concepts on this seemed less common and even not yet clear.

RfII regards early user involvement as one of the guiding principles in the development of the NFDI services. A number of the panel suggestions align with existing recommendations for the NFDI and some ideas augment them, such as the proposal of identifying and supporting data champions as early movers and community ambassadors.

Building user communities that can voice their needs and collaborate actively with the research data and information infrastructure experts to shape common and accepted services for their research domains is vital. In this regard, the valuable impulses developed in the workshop should be explored further to build convincing concepts and development paths for research data infrastructures and the communities that will make them sustainable.

**Imprint**

The organizers (Lars Bernard, Barbara Ebert and Ilja Zeitlin) thank all workshop participants and contributors: Amy Walton, Rowena Davis, Michael Diepenbroek, Laurel L. Haak, Ralph Müller-Pfefferkorn, Tom Bakker, Darja Fišer, Harry Enke and Frank Oliver Glöckner. Special thanks go to Debbie Alfred and Mike Mertens.

The positions presented in the above report are not the official positions of the German Council for Scientific Information Infrastructures (RfII).

Rat für Informationsinfrastrukturen (RfII) – Head Office
Papendiek 16, 37073 Göttingen
Germany
Phone +49 0551-3920959
E-mail info@rfii.de
Web www.rfii.de