Cultivating Innovative Data Science Communities

Jason Owen-Smith, Institute for Research on Innovation and Science, University of Michigan
Cultivating Innovative Data Science Communities

Jason Owen-Smith
Executive Director, Institute for Research on Innovation & Science
Professor, Sociology
Research Professor, Institute for Social Research
University of Michigan
jdos@umich.edu
iris.isr.umich.edu

DRAFT – 16 June 2021

Introduction

This conference highlights the immense promise and substantial difficulties that accompany efforts to systematically understand, explain and improve the value of science. We seek to address questions whose answers simultaneously represent both dramatic advances in fundamental knowledge in multiple fields and immediately relevant, practical contributions to some of today’s most pressing policy issues. Federal agencies, policy makers, academic leaders and researchers alike recognize that addressing these kinds of questions with rigor, clarity and timeliness requires a new approach.

That approach is anchored on data, its responsible, reliable, and secure integration and its safe, effective use for expert analyses and their lucid communication. The technical challenges are significant, but admirably addressed by other speakers. This essay attends to the social and organizational issues raised by work to link, protect, document and use many different types of data that are distributed across a plethora of organizations operating under varied, sometimes contradictory legal, proprietary, and ethical restrictions.

Consider a relatively simple question we might wish to answer: what is the return on federal investments in academic research and development (R&D)? ROI should be easy to calculate, right? To dramatically oversimplify things, billions of dollars of grants are made every year by more than dozen federal agencies to hundreds of universities where they support innovative research conducted by hundreds of thousands of people who organize their work in a wide variety of ways.

Grants, by themselves, don’t do anything. They enable research work. That work produces new knowledge while training people to the levels of expertise necessary to effectively apply it. Most of those highly skilled people leave the academy for employment across all sectors of the economy. They carry with them the knowledge and skills they gained during their grant-funded work. Much of that knowledge is codified in print or concretized in technology. Much more remains tacit, embodied in the people who produced it. Some codified knowledge is protected with patents, which grant a degree of legal protection that can enable it to be sold. Much more finds its way into the public domain in the form of articles, dissertations, books, and
other publications.(5) Often, both things are true.(6) Private sector access to tacit knowledge largely comes through employment.(7, 8)

At a bare minimum, making a start on this question requires that we know what specific grants have been made, in what amounts, to what institutions for what kinds of work. We also need to know how they were spent, what inputs were purchased and, critically, which particular people were employed and trained in what kinds of teams. Then we need to know what those teams produced, what role specific researchers played in that production, where those people eventually found other work, what they earned for themselves and what value they produced for their employers. The list could go on and on, but this should be sufficient to make the point.

The point is that answering even a relatively simple question in this domain requires many types of data (grant obligations and details, granular research expenditures, publication and patent data, dissertation information, transcripts, employment and wages, and detailed business data to name just a few) housed in many different locations (federal science agencies, university administrative systems, publishers and repositories, federal statistical agencies, state workforce and higher education agencies, firms) under many different sets of restrictions (federal and state laws, ethical protections, proprietary limits on use).

Pulling all those data together and using them to effectively and safely address the requires expertise from many fields (information science, domain science, data producers, social scientists). Translating the findings that result to the people and organizations who need to understand and use them requires an entirely different mix of skills housed in yet another set of organizations. In short, realizing the promise of this meeting requires new, or in some cases very old, forms of social organization even for relatively simple questions that we know, at least in general terms, how to answer. The unknown unknowns are much harder but potentially even more valuable. Addressing both requires means to integrate, develop and sustain innovative communities that can nimbly produce and communicate cumulative, replicable, validated knowledge built from far-flung, disparate component parts. This requires collective learning and shared access to individual resources that participants have strong positive and negative incentives to keep to themselves. The community must meaningfully include at least data producers and researchers, and probably also the communicators who use their results if it is to generate the value we envision. Sometimes these groups are the same. More often they not.

The days when complete, documented, self-contained datasets are handed off from a single, authoritative source to an anxious but largely passive user community are over. Rather than a system of more or less benevolent and centralized data monopolies that respond (or don’t) to the needs of distinct, well-defined user groups by fiat, we now face the challenge of building and sustaining a federated system comprised of many co-equal data providers and heterogenous, fluid groups of data users who face different, sometimes contradictory restrictions and requirements while seeking to achieve disparate and sometimes orthogonal goals. Matters are complicated because the skills and capabilities necessary to responsible, effective integration and use of disparate, dynamic, and continually changing datasets are often not housed any single
organization. As a result, the research community must often be directly and intimately involved in the work of data production and integration on an ongoing basis. That integration has great potential to inform and improve the work of both groups, if the right synergies can be cultivated.

**Conceptual tools for innovative data communities.**

The problem is social and organizational. A conceptual toolkit drawn from sociological, organizational, and network theory can help us solve it. That solution isn’t going to be sufficient but is going to be necessary. Technical and legal challenges remain and must also be confronted. But here I focus on the social aspects, briefly introduce two key concepts -- network forms of organization and generalized exchange – highlight a few of their features and identify some points of possible failure. We can then outline the essential components of an innovative learning community built around integrated data and its effective use, which I illustrate with examples drawn from the Institute for Innovation and Science (IRIS), an organization I co-founded with Julia Lane and Bruce Weinberg, which we purpose built to address just these challenges.

Generative communities are fickle things. Their existence implies a degree of trust and affinity among their members, often anchored on some productive form of interdependence enforced by reputational or other social means. For voluntary communities to work, membership must provide something of value to participants that they could not realize solely with their own skills and resources. That value must offer sufficient inducement to defer some of their individual interests and some of their autonomy in favor of the community’s health. It must also lead them to stick and stay in the face of the inevitable problems and conflicts that accompany any complex, collaborative endeavor. Communities are defined in part by a willingness to exercise voice in the face of those challenges rather than exit.(9) These features are particularly hard to develop when communities must draw from diverse groups whose cultures, priorities and constraints are at least somewhat alien to one another. They are harder to sustain when participants and their interests are at least somewhat inimical. I submit that both things are true of the community we might hope to build.

Several questions follow: (1) what kind of relationships create and sustain the positive social and affective features of community? (2) what kind of organizational arrangements can support and sustain such relationships? And (3) what dangers must we guard against in cultivating both?

**Question 1. What kind of relationships?** Voluntary communities that one must choose to join and work to help maintain at some cost to oneself are forged from interdependent goals and value. At the heart of both things is a peculiar social relationship, exchange. Some types of exchange relationships, for instance between strangers in market contexts, are relatively socially thin. Buying an apple from a stranger at a farmer’s market while visiting a distant town requires a lot of formal institutional infrastructure (such as enforceable property rights) (10) but relatively little social embeddedness (11). I don’t have to have any kind of relationship or any degree of commonality or fellow-feeling with the person who sells me my snack. Other types of exchange, like gift giving, or sharing of valuable and costly to transfer private information are more socially
complicated (12, 13) and require deeper social supports. The problems are compounded when exchanges carry significant risk or might be perceived as illegitimate.(14)

Again, I submit that the kinds of resource exchanges - of access to restricted, complicated data; of expertise and skill – necessary the community we envision have much of this character. Likewise, the conditions under which relationships will work – under time pressure; when parties collaborators might also be rivals pursuing orthogonal or unrelated ends; when there are high scientific and practical stakes – can introduce a substantial degree of risk. Finally, real and important concerns about privacy and security as well as binding legal restrictions raise the possibility that at least some, potentially powerful, observers might consider the sorts of exchanges we hope to facilitate to be illegitimate.

Under these conditions, we could trust to altruism. That has possibilities, but I’m enough of a realist (and hang out with enough economists) not to trust it too much. Instead, we should focus on reciprocity, the familiar tit-for-tat of game theory. Reciprocal exchanges are those where one party gives to another with the expectation that similarly valuable resources will be returned. When the shadow of the future is long, that can be sufficient to support the emergence of relatively complex cooperative systems, but it also runs the risk of a downward spiral of defection.(15)

There are two general types of reciprocity. Direct reciprocation happens when one enters an exchange relationship expecting that the specific partner with whom one is exchanging valuable resources will also be the one who gives back. A second form, indirect or generalized exchange, is more diffuse and takes place when one gives to a partner with the expectation of return from a person or organization who was not party to the original exchange. This requires first that there be multiple parties (a community if you will) engaged in similar forms of valuable, reciprocal exchange who are aware of and engaged with one another (16). Generalized exchange is a common form of gift giving among kinship groups.(17) I might offer something of value to a cousin with the expectation that eventually an uncle will return the favor, even if that cousin never does.

This kind of “pay it forward” ethos is subject to real dangers of free-riding or opportunistic exit and thus requires a degree of trust that typically only emerges within a system of ongoing, successful relationships; an exchange network. However, when it works generalized exchange generates and sustains social solidarity, which encompasses most of the features of generative communities that opened this section (18). Kinship clearly won’t do the trick here. Instead, I suggest an organizational scaffold.

Question 2. What kinds of organizational arrangements?

Network forms of organization exist at a midpoint between markets - the socially thin but institutionally thick contexts for generally short-term, transactional, and arms-length exchanges like buying that apple in a far-away farmers market – and hierarchies - the formal bureaucratic
structures that characterize large organizations. It seems unlikely that market coordination will result in the kind of community we need. Even if it were possible, internalizing the wide range of data, skills and expertise necessary to our task in a large bureaucracy would create a dangerous monopoly and might increase security risks by concentrating all the data in a single, presumably fallible organization. We also know that generalized exchange and the benefits it could bring requires a distributed network of interdependent players.

Network forms of organization are more nimble, “lighter on their feet,” than bureaucracies and much better at the transfer of complicated, highly tacit information necessary to the sorts of substantive work we’d want an innovative data community to accomplish than markets (19, 20). They are generally stable enough to create a long shadow of the future and to foster the kinds of trust, informal, reputation-based social control, and forbearance that are necessary in the absence of perfect contracts (21). Network forms of organization are common and generally effective in precisely the kinds of diverse, skilled, and creative communities we seek to develop. They are often found in industrial districts, regional technology communities, creative industries, and open-source software projects where rivals on one project may be collaborators on another and where disparate, sometimes conflicting interests and goals are commonplace (22–26).

Network forms are particularly common and effective “when the knowledge base . . . is both complex and expanding and the sources of expertise are widely dispersed” (27). Dispersed, rapidly changing knowledge and complex interdependencies among component parts of innovations characterize this type of governance. Those are also exactly the features of a data science community that seeks to understand, explain and improve the value of science (28).

The seed crystals for successful networks have particular characteristics. They are often themselves organizations who anchor communities, convene partners, and make significant, substantive contributions to common projects. Network anchors help to set the technical directions and the tone for their communities (29, 30). Though they lack the authority or power necessary to compel participation or dictate behavior, they exert significant influence by serving as neutral brokers and meeting grounds for participants. They are not disinterested because they work best when they are active and credible participants in collective work and thus can play important roles in fostering new partnerships, engaging new participants, aiding in collaborative problem solving, and helping to build system-wide capabilities (31, 32). Their ability to do all these things is aided by transparency and by the ability to help translate across the different languages and goals of their many partners. In doing so, they help provide a basis for the emergence of generalized exchange in a community. IRIS, which seeks to play just this kind of role in a consortium of major research universities and a growing, interdisciplinary research community was designed and operates according to these principles. One or more similar organizations could play this role for the community we seek to develop.

**Question 3. What are the dangers?**
Networks, like markets and hierarchies can and do fail. To put it starkly, networks fail “when exchange partners either screw each other or screw up”(28). There are instructive differences between screwing your collaborators and screwing up. The former is most often a function of opportunism, “self-interest seeking with guile”(33). The problems of defection, free-riding, self-dealing, and strategic exit from relationships or communities that are dangers of both direct and generalized reciprocal exchange are all examples. While it’s often hard to distinguish them in practice, or in the heat of the moment, screwing up is less malignant but no less dangerous. Partners screw up when they lack the competencies or knowledge necessary to successfully pursue complex and difficult collaborative projects. Networks screw up when they stop seeking new information from outside their current orbit and calcify, closing themselves off and becoming unable to identify and adapt to new situations or needs. A state of affairs that renders them incapable of the kind of flexibility and responsiveness that help ensure membership in the community is valuable. The problem here isn’t guile, it’s ignorance or a simple competency shortfall. These, along with various forms of opportunism are the primary dangers for network forms of organization that seek to foster generalized exchange in support of vibrant, innovative learning communities.

*Cultivating community at IRIS*

These abstract concerns have concrete practical implications for cultivating data science communities. Several of those bear greater scrutiny and can be illustrated with concrete examples from IRIS’ work in this domain. While mindfully designing network anchors to address these challenges is essentially, the long-term success of any community depends on all participants taking at least some responsibility as well.

The Institute for Research on Innovation and Science (IRIS) is an IRB approved data repository housed at the University of Michigan, which anchors a consortium of major research universities and maintains productive partnerships with federal statistical and science agencies, foundations, higher education advocacy groups, and other stakeholders in or adjacent to the science policy domain. University members of the IRIS network share transaction level administrative data on the direct cost expenditures of federal and non-federal sponsored project grants. IRIS uses these data along with information derived from more than fifty other sources to construct and document an integrated research data release, the Universities Measuring the Effects of Research on Competitiveness, Innovation and Science (UMETRICS) dataset. Bruce Weinberg (cite piece for conference), one of the IRIS co-founders, presents many more details about the UMETRICS data and its uses. For now, it’s only important to know that university administrative data form the kernel of a growing data mosaic that, because of the central role of academic research in the development and value of science, can be extended through responsible linkage to include many types of information at a high degree of granularity.

IRIS uses UMETRICS data for several purposes under protections established in MOUs and other agreements executed with university members and diverse other partners. Three of those
uses are particularly important background for the examples and discussion to follow. First, IRIS makes research data, documentation, and support available to a growing interdisciplinary research community of more than 350 people from nearly 150 institutions through a secure virtual data enclave maintained at Michigan, through the Federal Research Data Center (FSRDC) system and, soon, through the Coleridge Initiative’s Administrative Data Research facility (ADRF). Developing and supporting a vibrant research community represents one use for the data we produce and protect.

Second, IRIS uses UMETRICS, the results of its own and community research, and the skills we have built in working with the data to design, develop and share valuable data products with our university members. We routinely produce nine different interactive data products for a variety of use cases - ranging from government relations and communication to faculty and research development and the improvement of education. Universities thus receive direct and tangible benefits, that would be very difficult to realize on their own. We continually update and pilot new products in collaboration with our university partners as research findings create new knowledge and as their needs shift. For instance, rapid turnaround reporting and benchmarking on the research effects of COVID-19 aided many of our universities in tracking and responding to the effects of shutdowns.

Third, we use data aggregated across all participating campuses, which currently represent about 41% of all academic R&D spending in the nation, to develop and produce public facing data products used by policymakers, advocacy groups, science agencies, and other stakeholders to forward their own goals and missions in science policy, science communication, federal and state outreach. IRIS is increasingly turning these-privacy disclosed aggregate products toward communicating the value of science to various publics.

At its base, then, IRIS is a network anchor that makes responsible, secure use of a wide variety of open access and restricted data from many sources to support fundamental and policy relevant research and provide valuable benefits to all members of its community. This is the context in which we work to mitigate the dangers of opportunism and competency shortfalls. As I noted, those two things are sometimes difficult to distinguish in practice, so most of IRIS’ work addresses aspects of both.

Addressing the dangers.

Minimizing opportunism requires formal and informal arrangements designed to promote fairness and transparency in procedures while also supporting inclusiveness and accessibility in operations and data access. IRIS strives to accomplish this by making all requirements and materials for data access public along with easily accessible and clear documentation, ready and equitably distributed administrative and research support, and encouragement for data and expertise sharing within groups (e.g. researchers and data producers) and across them.
A key part of this work is ensuring not only that access and support are transparent and fair, but also that there are mechanisms to allow all members of the community to receive credit for work they do that contributes to the common good. With credit, we believe, comes credibility, for IRIS and more importantly for individual researchers. A major source of value for research community members is recognition for completing quality work that is documented and freely shared. A major means to limit opportunism is to facilitate means for data users and data providers to establish positive (or negative) reputations for themselves, which can help guide the collaboration decisions of their future partners. Both of these mechanisms benefit from transparency and reinforce procedural fairness. They are augmented by mechanisms such as research support for documenting and archiving code and measures as well as replication datasets. Collaborative tools (such as secure git repositories within the IRIS virtual data enclave and means to discovery, disseminate and cite user-generated work (such as IRIS-minted DOIs) aid in both these processes while helping to make certain that the network IRIS anchors does not “screw up” by becoming closed to new ideas, new techniques and new analysts.

Running throughout the whole of this essay is the theme of value for participants. If we are not to rely solely on altruism or on pre-existing bonds of affinity or kinship (which could, themselves contribute to network closure and calcification), then the pursuit of individual interests and goals that can only be met with inputs from others becomes a primary driver of participation and spur toward generalized exchange. We directly ensure value to data providers, which builds trust that facilitates the creation of streamlined data protections that, in turn, put access to data within reach of a wider community and increase the range of types and sources of data that are available. As the quality, scale, and scope of data resources increase, so too does the value for all members of the community. Perhaps unsurprisingly, shared, community-driven data resources and vibrant generalized exchange networks have positive externalities, reinforcing a virtuous cycle of credit and engagement while further diminishing the likelihood of opportunism by increasing the amount one stands to lose when bad behavior is identified.

Taking advantage of the many sources of value that are ‘baked into’ joining and fully participating in the IRIS network also requires attention to costs. The fundamental cost of complex collaborations that involve data sharing and integration are as much transactional as they are technical or financial. Indeed, the challenge individual researchers would face in overcoming the frictions necessary to get many data access permission with a wide range of providers would be unsurmountable for most. By ensuring that IRIS directly returns valuable products to data providers and building trust through transparency, collaboration and responsiveness, we are able (with the help of well-crafted agreements) to minimize the transaction costs of data access and collaboration for all. IRIS negotiates dozens of agreements with provisions that allow us to make the full data set we produce available to researchers under a single agreement with no additional university approvals.

Likewise, non-university data providers can, with appropriate approvals from IRIS’ elected Board of Directors, access data and linking assets from many sources with a single agreement. Establishing the value of the dataset as a whole for aggregate reporting on matters of pressing
concern to non-university partners increases the likelihood that future data-sharing agreements will augment value for everyone. The benefits of participation compound to limit the likelihood of opportunistic behavior by lowering the cost participation through streamlined data access and increasing its value through growing scale and scope. Here too, reciprocity becomes an essential, and positive force for community.

Transparency, fairness, accessibility, value, and efficiency may seem far removed from the day-to-day details of developing research community. But they are essential accelerants for generalized exchange and retardants for opportunism.

The primary means we use to decrease the dangers associated with screwing up flow from and reinforce these principles too. If one primary danger for data science communities is competency shortfalls among participants, it becomes essential to meet data users and data producers where they are. Building research community is thus inseparable from building collaborative, technical and analytic capacity. Training opportunities, public convenings, clear documentation and support for mechanisms to allow learning within and across constituent groups increase accessibility and the flow of new people and ideas into the network. They also reduce the forms of ignorance can lead to unintentional screw-ups. IRIS routinely offers a 9 day long virtual workshop series aimed at providing novice research users of large-scale restricted data with the tools they need to get started. Sustained collaborative work and substantial technical support in those settings, jumpstarts expertise building and generalized exchange. Other courses, which IRIS contributes to, but that are run by our partners the National Center for Science and Engineering Statistics (NCSES) and the Coleridge Initiative, those course aim to engage analysts and technical staff at federal and state agencies as well as universities. This serves not only to increase capacities for collaboration, but also to further integrate disparate but interdependent portions of the community and to help members with different backgrounds and expertise learn from and communicate with each other. The macroscopic results of all that work refresh the network, expand the grounds for generalized exchange, and decrease the likelihood of both screw-ups and screw-ings.

Discussion

The examples could continue, but more important than any single activity is the general that is, I believe, essential to the success of any innovative data community. Without denying the essential importance of technical skills and tools, or the concrete formal institutional architecture of contracts and agreements, or the analytical muscle of expert researchers, I argue that a fundamental challenge for creating and sustaining innovative data communities is social and organizational. The features - interdependence, trust, affinity, exchange, responsiveness, inclusivity – that sociological and organizational theories suggest are required for successful community building, pose both clear opportunities and pressing dangers. Addressing them effectively is matter of mindful and integrated organizational design no less than analytic and substantive success is a matter of mindful and integrated data construction. But what this means is that there’s no magic bullet, and community development can neither be limited to one set of
relevant stakeholders, nor treated as an isolated or one-off intervention. Instead getting community right in the sense of seeding a vibrant and open network form of organizational governance that can support and expand the community and substantive value of generalized exchange requires that close attention be paid to the fundamental structures of social life and relationships that must necessarily underpin any lasting and successful community.
Works Cited


2. J. Goldschlag, Nathan; Jarmin, Ron; Zolas, Nikolas; Lane, in *NBER CRIW volume on The Measurement and Diffusion of Innovation*, J. Corrado, Carol; Sichel, Dan; Miranda, Ed. (NBER, Boston, MA, 2018).


22. S. O’Mahony, F. Ferraro, The emergence of governance in an open source community.