Title
2nd Knowledge Infrastructure Workshop – question responses

Author
Palmer, Carole

Publication Date
2020-02-26

2nd Knowledge Infrastructure Workshop – question responses

Carole Palmer
Information School
University of Washington

What are the most urgent research questions to address about KI? Why?

The questions that rise to the top for me relate to data infrastructures for researchers and the public.

Metastructures. On the functional front of data infrastructure, we need more sophisticated metastructures that support strategic data access and reuse. Limited progress has been made on foundational problems of "scatter" and interoperability across large distributed information systems. Directories such as *Re3data* and *Fairsharing* provide access to thousands of repositories, each with hundreds or thousands of datasets. They are useful, especially for discerning the landscape of options and density of particular kinds of data, and perhaps seeding certain kinds of data-driven inquiry. They are far from optimized, however, for cross-disciplinary inquiry or an array of other important purposes, such as identifying data from key events or locations or addressing sparse data problems. Interesting tensions are at play. For example, journal requirements that treat datasets as discrete supplements to published papers reinforce centrifugal scatter and siloing. At the same time, these papers are the richest source of contextual metadata, guarding against loss of provenance and intellectual history as data are continually subsetted, integrated, and repurposed.

Public KI. How can KI function as infrastructure for knowing, for the public? While some research data centers state that they hold data "for everyone" (National Snow and Ice Data Center) and data of interest to the "general public" (IPCC Data Distribution Centre), high levels of intermediation would be required for the public to benefit from these vast and complex resources. What interpretive functions would make evidence, and its significance, understandable or knowable? The transparency of the "who", "how", and "why" of evidence seems fundamental, but how would such documentation need to differ for researchers, journalists, policymakers, and citizens?

Relatedly, KI for the public is taking form in the open government data movement. A research blueprint is needed to achieve more than access to ad hoc collections or

dumps of unusable data across the emerging panoply of city, state, and federal open data portals. Data literacy needs to be examined to show how government transparency and accountability, local evidence-based decision making, and informed civic engagement can be achieved for groups beyond civic tech activists.

Politicization. Risks from the politicization and undermining of science and expertise cut across KI for research and the public. What is the path forward for proactive safeguarding of data and information vital to scientific and social progress?

Long-term KI strategy. KI is costly with limited short-term payoffs. How do we assess progress and impacts across KI to improve where and how we invest? How do we ramp up KI for existential crises in climate change and democracy? What can we learn from comparison with Europe where much KI is more centrally supported and managed?

Identify a KI whose survival is under threat.

The Qualitative Data Repository (QDR)

The observations offered here were developed in consultation with Nic Weber, the lead technical director of QDR.

QDR is hosted by the Center for Qualitative and Multi-Method Inquiry (CQMI) at the Maxwell School of Citizenship and Public Affairs. The initiative has been led by political science scholars, funded primarily through 3 grants from NSF/SES/SBE beginning in 2015. More specialized research on tools, protocols, and interoperability of qualitative and sensitive data has been funded by the Robert Wood Johnson and the Sloan foundations. Development has progressed in close consultation with leaders of long-standing social science data repositories in the US and UK, as well as qualitative scholarly communities.

Disclaimer--I have served as a co-technical director and advisor since 2015.

a. What led to these threats? Over what time frame?

The prioritization of investment in quantitative over qualitative research is an overarching factor. Many of the other major threats are also familiar and typical of

many KI initiatives—dependency on grant funding, misalignment with academic priorities, and scarce expertise. The observations below are all ripe for further empirical study.

The short cycles and competitive nature of sponsored research programs encourage incremental innovation, but often at the expense of stability in development, responsiveness to a broad user base, and progress on institutional sustainability. Tool development is much easier to fund but also pulls attention, effort, and expertise away from building robust core operations and services.

Advantages gained from being part of an academic center of excellence--in this case CQMI are offset by the misalignment with university mission, structure, and incentives that work against applied R&D and continuity of service.

Professionals with technical experience with infrastructure and unstructured data are very difficult to recruit and retain, and demand is growing. Need for sophisticated security expertise is acute, as seen with the recent ransom hack of GESIS data catalogue at the Leibniz Institute for the Social Sciences. Competing with industry for talent is a crisis for universities, but also for city governments experiencing ransomware threats and other cyberattacks.

b. What actions or changes in circumstances might lead to its survival?

Recognition of the value of qualitative inquiry and a business case for investment are key (see part c, below). In addition, institutionalization, collaboration, and professionalization are needed to achieve sustainability over the long term.

Hybrid organizations within universities can integrate important resources and expertise from academic research centers, research computing, and library data services to create centers of excellence for research products--data, code, protocols, documentation, and other related outputs.

Community supported platforms hold considerable promise. QDR's transition to Dataverse has resulted in the kind of win-win relationship an open source model should foster. Uptake of an institutional membership business model has also been stronger than expected. However, fee structures are unlikely to ever fully cover R&D innovation, the level of curation required for sensitive data, and competitive compensation for technical personnel. Investment in professionalization of KI work

will be essential, with a focus on sociotechnical education for public sector technical and intermediary roles.

Consortia are also essential but take a special kind of institution building; and, more consortial activity is not necessarily better. In the social sciences, DataPass has been important for building cohesion and a shared catalog. The recent Dataverse Global Community Consortium is particularly promising for technical development. And, while there is value to be gained from coattail momentum with quantitative data services, keeping the interests of qualitative research as a priority takes continual commitment and proactive engagement.

c. What will be gained or lost, by whom, if this KI fails to survive?

Stakes are high for the survival of rich research resources but also for the continued promotion of the value of qualitative inquiry and attention to ethical research practices for the sharing and reuse of sensitive data.

The discourse surrounding QDR activities helps balance attention on privacy solutions, for example, by extending discussions beyond differential privacy and other computational approaches that require distant, buffered, or obscured use of evidence. It expands on notions of quantitative reproducibility and replication as gold standards to include techniques that support verification of validity, as seen with QDR's work with American Political Science Association on data publication requirements. Unique KI contributions are fostered through design for specific needs of qualitative scholarly communities. For example, QDR's work with integrative socio-environmental researchers will make progress on data sharing approaches that reduce degradation of fragile locations and stress on over-studied groups, two unintended consequences of research on high profile events and localized environmental and health phenomena. Work with native American researchers is determining the curatorial support needed for data sharing consistent with principles for indigenous data governance.

How do KI spread information? Misinformation? Alone and in combination with other infrastructures?

This is an important question for knowledge transfer within and across scholarly communities, but also with educators, journalists, and the public. In addition to the filter bubble problem, risks are high for the veracity of knowledge in the dynamic

and volatile digital environment. We need strategies for sustaining the intellectual foundations and authenticity of bodies of evidence and interpretation for claims in science, the popular press, and social media. Techniques such as QDR's "annotation for transparent inquiry" are designed to capture the process and logic of traditional qualitative social science. They can benefit other fields, such as computational social science research on web interactions, where there is not yet a tradition of interpretive documentation and archiving of the many interpretive acts that require transparency, such as categorizing data, algorithm selection, etc.