**Synthesis and Cyberinfrastructure for SBE Research**[[1]](#footnote-1)

White Paper in Response to:

SBE 2020: Future Research in the Social, Behavioral & Economic Sciences

Keith W. Kintigh,[[2]](#footnote-2) Francis P. McManamon[[3]](#footnote-3) and Katherine A. Spielmann[[4]](#footnote-4)

Arizona State University

**Abstract.** Insights from the social and behavioral sciences, as well as from economics, are increasingly recognized as essential components of solutions to an enormous range of problems faced by the United States. There is an increasing need for social scientists to integrate their understandings with those of life and physical scientists, mathematicians, engineers, and policy makers to achieve transformative understandings of the ways in which we can use diverse and complex data to address problems of coupled social and natural systems. NSF has invested heavily in cyberinfrastructure in the natural sciences. For social science to deliver effective, policy-relevant knowledge, substantial investments in social science information infrastructures are essential. Many cyberinfrastructure challenges faced by social science are qualitatively different than those of the natural sciences, because of the diversity and complexity of our data. While we can leverage cyberinfrastructure investments in other domains, we cannot just piggyback on them. We need sustained investments in social science synthesis and cyberinfrastructure, and social science-driven research on the integration of social and natural science data in the investigations of coupled social and natural systems. The problems are too important and the potential payoffs too large to delay.

Insights from the social and behavioral sciences, as well as from economics, are increasingly being recognized as essential components of solutions to an enormous range of problems faced by the United States and the World. Social and economic research is not only essential for problems that are fundamentally social, such as ethnic conflict or religious fundamentalism, but also for problems that are widely seen to be susceptible to technological solutions—from global warming to energy independence—because technical solutions have social and economic demands and consequences. While the importance of social, behavioral, and economic insights is not news to the SBE community, there is an increasing demand for social scientists to integrate their understandings with those of life and physical scientists, mathematicians, engineers, and policy makers in order to achieve transformative understandings of the ways in which we can use diverse and complex data to address problems of coupled social and natural systems.

For social science to achieve its potential to contribute rich scientific understandings of coupled human and natural systems, there is a pressing need for investment in social science information infrastructures that will allow us to archive, access, integrate, and synthesize diverse social data and then engage with relevant natural scientists and natural science data to effectively address real-world issues. This presents a number of interrelated challenges to SBE researchers. In this white paper, we focus on the critical importance of major investments in information infrastructures for social sciences. We address the relationships of these infrastructures to basic and applied social science research and to our ability to apply rich social science data in transdisciplinary problem solving.

NSF has invested heavily in cyberinfrastructure in the natural sciences though the development of data centers, synthesis centers, and integrative initiatives such as the [Long Term Ecological Research Network](http://www.lternet.edu/) (LTER), the [National Center for Environmental Analysis and Synthesis](http://www.nceas.ucsb.edu/) (NCEAS), and the [Data Observation Network for Earth](https://www.dataone.org/) (DataONE). Comparable investments in social science research and cyberinfrastructure are needed to transform social scientific research in ways that will enable us to deliver the fundamental knowledge that is desperately needed. This task must fall, in large part, to NSF with leadership from SBE, because no other institution has the resources, the mandate, *or* the foresight to lead this effort. Without NSF-SBE support and validation, scholarship will not achieve the large-scale and in-depth, comparative and integrative research that an effective cyberinfrastructure enables.

The challenges for SBE are no less difficult than they are in the natural sciences, but many of them are of a different nature. SBE cyberinfrastructure challenges lie less in moving petabytes of data or utilizing tens of teraFLOPS of processing, but derive more from the inherent complexity of much of our data. *Indeed, a fundamental and pressing challenge for SBE is understanding the ways in which we can use complex and diverse data to address problems of coupled social and natural systems.*

How to deal most effectively with this complexity is not yet clear. However, it is reasonably clear that the solution is not simply a matter of directing more resources into the application of existing technologies to social science domains. Instead, we urgently need both fundamental research on social science cyberinfrastructure problems and investment social science cyberinfrastructures to explore the productivity, in practice, of different approaches and in different disciplines. The diversity of the social sciences and the inherent difficulties of these challenges suggest that a number of quite substantial investments in cyberinfrastructure will be needed, and that some will be more successful in transforming data to useful knowledge than others. It is crucial that different avenues be pursued to find the most effective routes forward.

So what are the problems that need to be addressed?

1. Synthesis. We are, in many ways drowning in data but struggling to take advantage of it. We desperately need to foster synthetic research that transcends the spatial and temporal scales of individual research projects, both within and across traditional disciplinary boundaries. This synthetic research will require that we integrate not only recently collected datasets but also legacy datasets—whatever their problems—that cannot possibly be reproduced. NSF synthesis centers have pioneered approaches to synthesis. Leveraging these insights, we need a next generation of synthesis efforts *driven by social scientists* that address the challenges of understanding coupled human and natural systems. We believe that initiatives coming out of SBE that engage socio-environmental problems will prove complementary to the Directorate for Biological Sciences’ new Environmental Synthesis Center.
2. Accessibility. Enormous amounts of social scientific data have been collected and exist in various digital, as well as paper, forms. However, our research workflows (plus confidentiality and other concerns) generally don’t result in the publication of the primary data. We need to capture these data and to acquire adequate technical and semantic metadata to keep these data usable into the future. Achieving effective discovery and access requires domain knowledge to insure responsiveness to particular disciplinary research demands. Such accessibility also requires commitment to deposit newly collected data in appropriate repositories (as encouraged by the May announcement that Data Management plans will be required in NSF proposals) and an investment in acquiring (and sometimes digitizing) targeted legacy datasets.
3. Metadata. As we try to understand what would constitute adequate semantic metadata—the information needed for a sophisticated third party to make sensible scientific use of the data—we encounter, head on, the enormous diversity of the kinds of data that are collected by social, behavioral, and economic scientists. A one-size-fits-all approach to social science metadata is certainly a nonstarter—but we need research, jointly executed by social and information scientists, into the most effective representations for different classes of social science metadata. For example, the highly abstract [Ecological Metadata Language](http://knb.ecoinformatics.org/software/eml/) (EML) could be applied or adapted to many sorts of social science observations. The [CIDOC Conceptual Reference Model](http://www.cidoc-crm.org/) is a quite abstract ISO standard for documenting cultural heritage resources. In contrast, [the Digital Archaeological Record](http://tdar.org/) (tDAR)[[5]](#footnote-5) has employed an approach to metadata more directly linked to specific kinds of observations made by archaeologists—an approach that has facilitated high quality data integration.
4. Metadata Standards and Interoperability. While we believe that adequate metadata standards for many sorts of social, behavioral, and economic data have not yet been developed, some existing standards (e.g. FGDC for geospatial data) may serve perfectly well for some data elements. Others, such as Dublin Core are sufficiently general that mappings can be readily achieved from more detailed schemes that are responsive to the intensive metadata demands of more targeted applications. We need to identify representations that are up to the data integration challenges and that, at the same time, can be mapped into standards that are interoperable with different social science data sources and with relevant counterparts in the life and physical sciences.
5. Integration. Synthesis demands data integration. Understanding fundamental processes that operate in coupled social and environmental systems will require investigations at multiple spatial scales (including very large ones) and over long time periods. This in turn will require integrating data collected by diverse researchers using different recording protocols. As we explore strategies for encoding metadata, we argue that we will not know what strategies are, in fact, adequate until we use them to integrate datasets. Indeed the effectiveness of a metadata encoding strategy should measured by its ability to enable data integration—not by some free-standing expert evaluation of adequacy. This is important because two different metadata representations that in some abstract sense contain the same information may differ greatly in their susceptibility to data integration. Data integration itself represents enormous challenges both because of the complexity of the data and because of the uncertain degree to which they are conditioned by the context of their collection and the underlying uncertainties about what precisely is relevant in a given context. These considerations complicate and constrain data integration because the relevant scientific communities properly insist that only some sorts of operations on the component data sets are acceptable in an integration process if the resulting integrated data set is to retain the quality necessary to support a rigorous (e.g., publishable) argument. For example, data mash-ups in which different pieces of information are pulled from available web resources without serious review may be useful exploratory tools but often will not yield rigorous results.
6. Visualization. Coupled social and natural systems are notoriously complex. Their investigation will not only require sophisticated modelling, it will also demand innovative visualizations of empirical and model results. For example, significant progress, funded by SBE’s Decision Making Under Uncertainty program, has been made by the Decision Center for a Desert City in collaboration with Arizona State University’s Decision Theater. However, investment in a diversity of approaches is needed.
7. Preservation. Believing in the long-term value of data implies that we provide not only for discovery and access to these data but for their preservation. Preservation is insured through the acquisition of detailed semantic and technical metadata, data migration, and highly reliable backup. Concerns about preservation also highlight the need to acquire valuable data that are critically endangered by media degradation, software obsolescence, and inadequate metadata. More broadly, we need to be concerned about the sustainability of established formal or informal data centers, recognizing that the sustainability has technical, financial, and social dimensions.
8. Social Dimensions. Finally, NSF should encourage a new paradigm in which integrative and synthetic research are highly valued and should incentivize a scientific culture in which sharing of data becomes an accepted norm of professional behavior. A recent study by UC Berkeley’s Center for Studies in Higher Education[[6]](#footnote-6) had the troubling finding that junior faculty in archaeology, economics, history, and political science are following the same paths of research focus and professional publication as established faculty. They are not making use of the existing variety and range of digital data in their research or in the dissemination of their scholarship. The investments outlined above should be designed to particularly engage junior scholars in these modes of inquiry.

For social science to deliver transformative, policy-relevant knowledge about the operation of coupled social and natural systems will require substantial investments in cyberinfrastructure. Many of the cyberinfrastructure challenges faced by social science are qualitatively different than those of the life and physical sciences, largely because of the diversity and complexity of the data. While we can leverage cyberinfrastructure investments in these other disciplines, we cannot just piggyback on them. Important lessons have been learned and technologies developed; however, alone they will not permit research on coupled social and natural systems to be as effective as it *must* become. We need sustained investments in social science synthesis and cyberinfrastructure, and in social science-driven research on how social and natural science data can be most effectively integrated in investigations of coupled social and natural systems. The problems are too important and the potential payoffs too large to delay.

1. This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA. [↑](#footnote-ref-1)
2. Keith W. Kintigh (kintigh@asu.edu) Associate Director and Professor, School of Human Evolution & Social Change, Arizona State University, Tempe, AZ 85287-2402. [↑](#footnote-ref-2)
3. Francis P. McManamon (fpm@digitalantiquity.org), Executive Director, Center for Digital Antiquity, Arizona State University, Tempe, AZ 85287-2402 [↑](#footnote-ref-3)
4. Katherine A. Spielmann (kate.spielmann@asu.edu), Professor, School of Human Evolution & Social Change, Arizona State University, Tempe, AZ 85287-2402. [↑](#footnote-ref-4)
5. McManamon, Francis P. and Keith W. Kintigh (2010) Digital Antiquity: Transforming Archaeological Data into Knowledge. *The SAA Archaeological Record* 10(2):37-40. <http://digital.ipcprintservices.com/publication/?i=34758&pre=1&p=39> [↑](#footnote-ref-5)
6. Harley, Diane, Sophia Krzys Acord, Shannon Lawrence, C. Judson King (2010) *Final Report: Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines* Center for the Studies in Higher Education, CSHE 1.10. University of California, Berkeley. <http://cshe.berkeley.edu/publications/publications.php?id=351> [↑](#footnote-ref-6)