Journal of the Association for Information Systems

JAIS

Editor Note

Data matters in IS theory building*

Kalle Lyytinen
EIC of JAIS
Case Western Reserve University
kalle@case.edu

* I am indebted to Michel Avital, Varun Grover, John King, Frank Land, Steve Sawyer, Carol Saunders and Susan Winter for useful discussions and critical comments on the editorial. I am also indebted to all participants of ICIS 2007 panel on Data Rich and Poor disciplines (Avital et al. 2007), and Senior Editors of JAIS who provided early comments and criticism on the JAIS data policies.

Volume 10, Issue 10, pp. 715-720, October 2009

Data matters in IS theory building

The *Journal of the Association for Information System (JAIS)* has been since its inception a theory-driven outlet that aspires to generate plausible, credible, generalizable, and illuminating understanding about information technology (and related socio-technical systems) in human enterprise. Therefore, our focus has been on promoting strong theory and associated inquiries in the information systems field as witnessed by our theory development workshops, recent editorial notes (e.g., http://aisel.aisnet.org/jais/vol9/iss2/5/ or http://aisel.aisnet.org/jais/vol9/iss8/21/) and several JAIS best paper awards (e.g., http://aisel.aisnet.org/jais/vol9/iss10/5/). These emphasize primarily new theoretical models and original theory building. All this would suggest that we are neither a data-driven journal nor interested in empirics of our field. Yet, nothing could be further from the truth. JAIS, while emphasizing theory generation as our mission, recognizes the central role of data in this endeavor. This is witnessed in our publication profile, which involves a large number of rigorous empirical studies, our interest in engaging in construct and instrument development, and the care we place in guarding rigorous data collection and analysis during the review process. To underline the criticality of data in theory building within our field, our senior editor team has recently instituted new data policies for JAIS (see http://aisel.aisnet.org/jais/policies.html) that we expect to guide authors on how data matters need to be addressed in prospective submissions to JAIS. We hope that these policies can be seen as a first step in clarifying our expectations and views on data matters.

In particular, these policies state guidelines for: 1) access to data for editors and reviewers during the review process, as well as some requirements for included data sets in submissions involving quantitative or qualitative analysis; 2) conditions under which the same data set can be used across a number of papers and submissions; 3) ethical standards associated with data collection, storage, and distribution; and 4) data liability in that JAIS bears no legal responsibility of the correctness of the data and related conclusions. We regard these policies to be instrumental in clarifying and unifying our stance toward the treatment of data and data sets in theory-driven submissions. It also shows that we do not take data matters lightly, and signals that we value data as a journal (JAIS). We also believe that these policies are important for development of the field. We hope that prospective authors will examine and recognize these guidelines before submitting their manuscripts to JAIS.

Next, I locate these data policies in a broader disciplinary context, and explain why I view these policies to be important in shaping our field and improving its prospects. To this end, I will examine more broadly the role of data, data policies, and associated institutional practices in establishing the IS community as an institutional force.

Since Keen's (1980) call for building a cumulative tradition in Information System (IS) research, the IS field has been seen to fall exceedingly short in providing an environment conducive to Kuhnian cumulative knowledge building. In most cases, the deficiency has been attributed to a chronic lack of unique core theories (Lyytinen and King 2006, Benbasat and Weber 1996, Benbasat and Zmud 2003), creating a heightened anxiety about the legitimacy and identity of the field (Lyytinen and King 2006, King and Lyytinen 2004). Subsequent discourse has centered on whether and why theory is (not) critical in defining and legitimizing the field. Though many have argued either against the need for "core" theory (Robey 1996, Klein and Hirschheim 2006), or the necessity of having theory at all (Lyytinen and King 2004), the discussion has, for some time now, been shunted into polarized positions about the need of few, some, many diverse, or no theories. Only a few alternatives focus interactions among the field's institutional strength -- legitimacy in providing strong and salient understanding of the IT rich world -- and its broader antecedents. DeSanctis' (2003) discussion of the IS discipline as a different sort of knowledge community where "maintaining the success of the community will require greater focus on internal matters than on external legitimacy" (DeSanctis 2003 p. 362); and Lyytinen and King's (2004) suggestion that "the IS field must replace the notion of the fixed core with a metaphor of a center of activity (or life form) that builds identity and legitimacy for the field" (Lyytinen and King 2004, p. 221) are among the few.

DeSanctis' call requires us to make a foray into alternative strategies to strengthen our field as a knowledge community with institutional strength. One critical element here is the impact of what we know about the subject matter seen through the lens of data sets at our disposal. In this regard, I

posit that much of the browbeating in our field overstates the importance of theory relative to the central role of data in developing and extending conceptual insight and strong knowledge claims. Simply, solely to emphasize theory is to mistake its symbiotic relationship to data. Calls for increased attention to theory have to be seen, in fact, in many situations as cries for better and more data. Theory is hollow without sustained and institutionally strong data (practices) to back up and warrant theoretical claims. In all research fields, theory evolution and data change are closely knit, and in most cases new theory follows from changes in data practices to the extent that in most cases good data precedes good theory. In the IS field this may come as a surprise to many, as in most cases the role of data is relegated only to the latter part of the research cycle (as an indicator or falsification) Many of the greatest scientists are famous, not because they invented new theory, but because they invented new data practices. The list is long: Copernicus, Curie, Newton, Milgram, Roentgen, Rubbia, just to name a few. One reason for this is that the presence of discomforting, but credible data is what leads scholars to search for or even imagine new theories. In short, most paradigms are born with and die with data. Thus, it is imperative to our field to recognize that data enter and shape all phases of the theory cycle and often serve as an important trigger and means to improve, or re-direct theorizing.

By data I denote anything that can be brought to bear in support (as evidence) while making knowledge claims (Toulmin 1958). Data can be numbers, words, images, or even (attributes of) artifacts that are entered as instances or aggregates to make knowledge claims about the world. Another premise is that data cannot be separated from the socio-technical networks that are instrumental in their production and legitimation (Latour 1987). Therefore, what counts as data is a social and technical construction and evolves over time. In research, data are only valuable for their use, i.e., as an element of arguments where knowledge claims are made and certain things about the world get settled[1] (Latour 1987, Toulmin 1958). Social construction determines that the meaning of data gets constantly negotiated within disciplinary discourse, but these negotiations are not a free for all, as they can only be framed when the data and its delivery mechanisms are effectively introduced as part of the argument. Conversely, the meaning of data evolves as they are either admitted, excluded, or modified in the discourse in support of varying knowledge claims. For example, x-rays were regarded as noise until, well, they weren't!

This diverse and socio-technical nature of data stresses the importance of tracing the evolution of more diversified and new, legitimate forms of making effective claims with data within all disciplinary fields and the related material and institutional process of generating increasingly broad and complex data sets. This evolution is accompanied by creating, enacting, and enforcing shared and institutionalized data policies and data analysis and use practices as integral elements of strong scholarly fields. It is, therefore, no surprise that many disciplines as diverse as astronomy, high energy physics, geology, history, archeology, and ocean sciences (to name a few) are known for their strong traditions of building and sharing cumulative data sets and institutionalized practices (Knorr-Cetina 1999, Gallison 1997). These fields afford their respective communities a wealth of data through publicly accessible, normalized datasets. The success of physicists and economists, for example, makes clear that building large-scale datasets and treating them as a community asset can have strong cumulative positive effects on the sophistication of research methodologies and the community's ability to make sense of and discover the world. In addition, building on the increased digitization of data, and the availability of low cost storage and transmission, many of these fields are moving quickly to become ultra data rich (e.g., astronomy, oceanography, biology, high energy physics). They are now executing mega-scale data charting and marking projects, such as mapping the sky,[2] the ocean,[3] and the human genome,[4] which yields petabytes or even zettabytes of data. Finally, stakes that relate to data have risen recently, as data is always imbued with a purpose. Data have become one of the epicenters of politics in fields as diverse as medicine, climate change, and poverty studies. For example, some members of the anti-regulation movement in the U.S. Congress

[1] Whether or not data exist without use is a deeper philosophical question and beyond our interests in this editorial.
[2] http://www.sdss.org
[3] http://www.oceanservice.noaa.gov/welcome.html
[4] http://genomics.energy.gov

have promoted legislation requiring that data be kept for all research that might impact regulatory decisions. This was, of course, the result of scientific studies producing data showing harm from various industrial activities that the Environmental Protection Agency (EPA), or other organizations then stepped in to regulate.

In this editorial, I submit to the view that the IS discipline is a data poor field with inadequate data preservation and reuse practices, and with relatively little advanced data instrumentation. Therefore, we need data policies that foster data re-use and preservation and give adequate attention to data. Though the field's data poverty and wealth vary significantly by individual researcher, I am concerned with the relative collective (or disciplinary or community) wealth in comparison to other disciplines, and the potential for such wealth in the IS discipline. By data richness, I denote a community's collective and heedful attention, commitment, and care to data and its quality, availability, and growth. When there is data richness, large corpora of data are viewed as a common asset, which deserves institutional response, care, and policies that extend beyond any one individual or small group. As an institutional asset, data repositories, methods, instruments, education, and use are commonly shared, enforced, and available. This has strong implications for a discipline's division of labour and incentives, as scholars who are specialized in instrumentation, data collection, and storage are viewed as important members of the community. Data poverty, likewise, is defined as the lack of collective data assets and attention to data matters, where data are part of private and personal scholarly efforts. Methods, by necessity, are shared, as they are the primary means to guarantee a public review to ensure that conclusions from "private" data are valid. But typically, this only takes place among scholars who already have the data. Data poverty results in little cross-method pooling of data across communities, and the range of methods in a field remains small (Avital 2000). Knowledge claims often become a mode of disbelief and are hard to discount, since data to curtail and falsify claims are not widely available: they remain private and are not well understood in the community. In the end only few know the methods and data, creating a negative reinforcing cycle.

I also argue that data poverty — rather than theory poverty — has created barriers for the development of a unique and strong IS discipline. It has limited the scope and scale of IS research projects and degraded the likelihood of reaching strong results with salience. For example, studies with an individual as the unit of analysis and a single snapshot data collection dominate our research in spite of the discipline's increasingly organizational orientation and the longitudinal nature of the underlying phenomena. Second, the field has an overreliance on singular and short-lived case studies, which are useful for exploratory and revelatory research but poorly suited for integrated theory development, let alone for confirmatory research with strong, generalizable results. Third, though significant progress has been made over the last decade, especially in instrumentation and instrument validation, we lack digital repositories of validated primary data that could truly contribute to cumulative tradition. Finally, the irony of data poverty is that, though our field is dedicated to the study of the design, development, and uses of information systems, it has largely failed to recognize the shift in the research enterprise to harvest and use digital data as one of the world's largest information systems design, development, and use projects. As a consequence, our field is missing the opportunity to capitalize on the shift, as we are not actively studying and designing our own digital research information systems. Though others are making their way to our new laboratory, we are not even banging on the door! One reason for this is that IS scholars — as members of a data poor socio-technical field — cannot put themselves into the shoes of data rich "big" natural sciences. IS scholars cannot think of going to public funding agencies and asking for support in building and maintaining new data sets beyond those already available. Data about IT strategy alignment just is not as exciting as data helping to find Higgs boson. This makes it difficult for social science-oriented IS researchers to conceive how they could create, maintain, and benefit from large shared data resources.[5]

[5] There are some successful examples how to build shared data sets, shared data policies and instrumentation in social sciences over time. For example, NSF's Sociology Division has been funding since 1972 large surveys like the General Social Survey which could serve as a model of the development and management of a data source of interest to IS researchers. This data base was developed, guided and managed by the Sociology field (although less than half of the users are from Sociology) and has seen phenomenal use over time with 9,000

Data poverty in the IS field stems largely from its tradition-bound internal norms and research practices. These are largely inherited from the individualistic research practices valued in business schools that promote sharing research findings only through publication of peer-reviewed papers, but ignore the synergistic effects of sharing of datasets and the instruments at earlier stages of research, during the review process, or post hoc. This attention deficit has also contributed to the emergence of alternative data brokers on IT-related data matters in response to the market's need to understand this pervasive and critical socio-economic phenomenon. Capitalizing on the growing demands and through systematic data collection over time, some large consultancies (e.g., IDC,[6] Gartner[7]) have become the de facto data brokers controlling IT-related data and their analyses. With hefty use charges, limited access to raw data, and often unknown (and, thus, suspect) data and instrumentation, such services, however, are difficult to apply for groundbreaking research and cumulative knowledge building.

Yet, some promising opportunities for data enrichment have recently emerged. First, there are some resources in ISworld that offer means for instrument sharing (e.g., http://www.misq.org/discovery/surveys98/surveys.html) and also for data sharing (e.g., the AIS working paper repository SPROUTS, which offers electronic means to share data sets: http://sprouts.aisnet.org).[8] Second, with the growing salience of IT, public data sources are beginning to yield data on phenomena of interest to IS researchers, which can provide a starting point for shared data analysis practices (e.g., IT-enabled outcomes in event studies and CIO compensation). Third, with the growing digitization of business processes, IS researchers are finding new data mining opportunities for pattern generation and discovery of anomalies. They are also creating shareable "archival data sets" (e.g., research on price dispersion, pricing of digital goods, or social networking). Finally, institutional data providers such as the Census Bureau, Bureau of Economic Analysis (BEA), and Bureau of Labor Statistics (BLS) in the U.S. or the international Organization for Economic Co-operation and Development (OECD: http://www.oecd.org/statsportal/) offer new opportunities for harvesting data. Finally, the ever-growing digital footprint is available through Internet traffic, telecommunication records, (especially through mobile phones and GPS, etc.), and other information infrastructures providing sensor data. Digital data sets are growing around a broad range of scientific and entrepreneurial initiatives including: NSF cyber-infrastructure,[9] EU E-Science,[10] Amazon,[11] Google,[12] Linden Labs,[13] and E-bay.[14] Unfortunately, awareness of these opportunities and their use is currently limited to a relatively small subset of the IS research community.

Given all this, we have institutionalized a first set of data policies in JAIS. I see this as a modest and early step on a long road toward stronger and more institutionalized data policies in our field.[i] I hope that by promulgating these policies, JAIS will increase awareness of new data enrichment opportunities, the urgent need to generate and integrate new and richer data sources, the urge to

---

published uses, 18,000 datasets downloaded, 19 million visits to the website between 1999 and 2003, 90 data extracts distributed with textbooks, and 250,000 students enrolled in courses that use the GSS (The General Social Survey (GSS): The Next Decade and Beyond, 2007). Although this survey is not without its detractors, its value to social science research and as a boundary object to focus conversations about sampling, survey methodology, content and international partnerships have helped the social scientists grapple with and solve many of the issues that the IS community would have to address to become a more data rich field (Personal communication with Susan Winter). Had we similar data sets on IT investment, use and capabilities and organizational use and governance over say 5000 organizations over last 30 years we could build better theories of IT impact, its evolutionary paths and so on.

[6] See e.g. http://www.idc.com/
[7] See e.g. http://www.gartner.com/technology/research.jsp
[8] Unfortunately there is very little indication that either of them have lead to increased data or instrument sharing or discourse around better data practices.
[9] See http://www.nsf.gov/dir/index.jsp?org=OCI
[10] See http://www.e-scidr.eu/
[11] See http://aws.amazon.com/publicdatasets/
[12] http://blog.wired.com/wiredscience/2008/12/googlescienceda.html
[13] http://www.whump.com/moreLikeThis/2009/02/02/data-visualization-in-second-life/
[14] See e.g. http://www.researchadvanced.com/

improve our data sharing and care, and the call for increased shared use and interpretation of data sets through our publication policies and standards. These are needed if we are to address the research challenges that come along with the pervasive digitizing of our external world and experience.

References

Avital, M., Iacono, S., Sambamurthy, V., Lyytinen, K., Kraemer, K.L., Sawyer, S. "Data rich and data poor scholarship: Where does IS research stand?," *Proceedings of Twenty Eighth International Conference on Information Systems*, Montreal, 2007

Benbasat, I. and Weber R. "Research Commentary: Rethinking "Diversity" in Information Systems Research," *Information Systems Research*, (7: 4), 1996, pp. 389-399.

DeSanctis, G "The Social Life of Information Systems Research A Response to Benbasat and Zmud's- Call for Returning to the IT Artifact", *Journal of the Association for Information Systems* Vol. 4 No. 7, pp. 360-376/December 2003

Galison, P. 1997. *Image and Logic: A Material Culture of Microphysics*., The University of Chicago Press, Chicago, IL.

Hirschheim, R. and Klein, H. "Crisis in the IS field? A critical reflection on the state of the discipline", in *Information Systems- The State of the Field*, King J. and Lyytinen, K. (Eds.), John Wiley & Sons, Chichester, England, 2006, pp. 71-146.

Keen, P.G.W. "MIS Research: Reference Disciplines and a Cumulative Tradition", in *Proceedings of the First International Conference on Information Systems,* McLean, E.R. (Ed.), Philadelphia, Pennsylvania, December 1980, pp. 9-18.

King, J., and K. Lyytinen (2004), "Reach and Grasp," *MIS Quarterly*, 28, 4, December, pp. 539-551.

King, J., and K. Lyytinen, eds. (2006), *Information Systems: The State of the Field*, Chichester: John Wiley & Sons

Knorr Cetina, K.. *Epistemic cultures. How the sciences make knowledge*. Cambridge, MA: Harvard University Press. (1999/3rd 2003)

Latour, B. 1987. *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press, Cambridge, MA.

Lyytinen, K. and King, J. "Nothing at the Center?: Academic Legitimacy in the Information Systems Field," *Journal of AIS,* (5:6), 2004.

Robey, D. "Research Commentary: Diversity in Information Systems Research: Threat, Promise and Responsibility," *Information Systems Research*, (7:4), 1996, pp. 400- 408.

Toulmin,S., *The uses of argument* , Cambridge University Press, Edelston, 1958

[i] We are aware that the proposed policies are just the first and humble step towards broader institutional practices and policies on data. Additional steps include decisions about the length of time that data would have to be made available for scrutiny, what form they would be kept in, whether they were raw or cleaned data (and how much cleaning), what metadata or para-data would be required and standardized, and whether computer code is also considered a form of data that must be provided upon request. These involve also questions about the costs of data retention i.e. who pays for storage and for access, IRB expectations regarding data re-use and data retention, and issues about anonymity for some forms of data such as photos