Association for Information Systems
AIS Electronic Library (AISeL)

ICIS 2007 Proceedings

International Conference on Information Systems (ICIS)

December 2007

Data Rich and Data Poor Scholarship: Where Does IS Research Stand?

Michel Avital
University of Amsterdam

Steve Sawyer
Penn State

Kenneth Kraemer
University of California, Irvine

V. Sambamurthy
Michigan State University

Kalle Lyytinen
Case Western Reserve University

See next page for additional authors

Authors
Michel Avital, Steve Sawyer, Kenneth Kraemer, V. Sambamurthy, Kalle Lyytinen, and C. Suzanne Iacono

DATA RICH AND DATA POOR SCHOLARSHIP: WHERE DOES IS RESEARCH STAND?

Michel Avital
University of Amsterdam
The Netherlands
avital@uva.nl

Kalle Lyytinen
Case Western Reserve University
USA
kalle@case.edu

Suzanne Iacono
National Science Foundation
USA
siacono@nsf.gov

Kenneth L. Kraemer
University of California, Irvine
USA
kkraemer@uci.edu

Vallabh Sambamurthy
Michigan State University
USA
sambamurthy@bus.msu.edu

Steve Sawyer
Pennsylvania State University
USA
sawyer@ist.psu.edu

Abstract

So far, the discourse around the nature of the IS discipline has focused on the centrality of theory in defining and legitimating the field. In contrast, we re-position this debate and focus on the centrality of data. Specifically, in this panel we discuss how systematic approaches to data sharing practices, improved data collection instrumentation as well as increased access to (and uses of) large institutionally managed corpora of data can play a critical role in the evolution and shaping of IS as a scholarly field of study. Subsequently, we explore the current position and status of data in IS research, and ask: how does it affect the prevailing research practices and the legitimacy of the field? And how, if at all, we should address the situation? We submit that the IS discipline as a whole has been a data poor field with inadequate data preservation and reuse practices, and with relatively less advanced data collection instrumentation. Overall, we argue that the practices of producing, maintaining and using data assets in data poor fields as the IS discipline result in economic deficiency, research ineffectiveness, and missed opportunities. Furthermore, we aim to highlight some emerging data enrichment opportunities and encourage more IS researchers to think about data enrichment in the prevailing programs of research around data. The IS research community's current lack of attention towards developing of large-scale cumulative data on IT-related subject matters begs the question: can we afford staying a data poor field?

Keywords: Large-scale datasets, data sharing, data repositories, IS research agenda, IS discipline

Panel Objective

Overall, we argue that the current practices of producing, maintaining and using data assets in data poor fields as the IS discipline result in economic deficiency, research ineffectiveness, and missed opportunities:

- *Economic deficiency* – Too often, we're busy reinventing the wheel. Reuse of instruments and capitalizing on existing data is not prevalent. Limited data availability and lack of refined instruments force IS researchers to spend too much of their resources on instrument development and raw data collection.
- *Research ineffectiveness* – Regardless of methodology, honing instruments over time and in multiple contexts results in more accurate and more reliable instruments. Accumulating datasets over time and from multiple sources results in richer and cleaner datasets. Currently, the availability of instruments is limited for few specific areas (e.g. technology acceptance), and even when available, they are poorly documented and rarely one can find further evidence that attests to their validity and reliability. Moreover, when multiple datasets and instruments are available, they are often incompatible. Consequently, we are nearly always forced to work with crude instruments and small-n or incomplete datasets.
- *Missed opportunities* – As noted, large and robust datasets that span across time and populations provide research opportunities that are not possible with small-scale single snapshot datasets. No such datasets are widely recognized as being central for the IS academic community at large. The few exceptions that are available are proprietary, expensive, not subject to public scrutiny, and practically out of reach for most due to high access cost.

The panel will explore a position arguing that the IS field will benefit from developing norms, institutional incentives and operational vehicles that promote building of large and robust datasets, improve data transparency and enforce institutional mechanisms that enable data preservation and reuse. Based on lessons from other disciplines and with a desire to strengthen our cumulative knowledge, the panel will explore two key issues related to the data ecology of the IS field: the potential and implications of intense data collection, and revamping the infrastructure and policies that facilitate preservation and sharing of data of all sorts. Furthermore, we will discuss building up data infrastructure for storage and dissemination of research data, and what kind of mega scale data projects the IS community should consider to champion. In all, the guiding question is how can we transform the IS field from a data poor to a data rich discipline – maybe through embracing open data community practices or changing the policies associated with current data monopolies.

Background and Rationale

Since Keen (1980) called for building a cumulative tradition in IS research, it has been often argued that the IS field falls short of providing an environment conducive to cumulative knowledge building based on Kuhnian normal science. In most cases, the deficiency has been attributed to a chronic lack of unique core theories within the IS field (e.g., Benbasat and Weber 1996). Though many have argued decisively against this view (e.g., Lyytinen and King 2004, Klein and Hirschheim 2006, Robey 1996), so far, the discourse has focused on the centrality of theory in defining and legitimating the field. In contrast, we engage this debate and focus on the centrality of data. Specifically, in this panel we will discuss how more systematic approach to data sharing practices and improved data collection instrumentation as well as increased access to (and use of) large institutionally managed corpora of data play a critical role in the evolution and shaping of IS as a scholarly field of study. Subsequently, we will explore the position and status of data in IS research and ask how does it affect the prevailing work practices and the legitimacy of the field? And how, if at all, we should address the situation?

We submit that the IS discipline is a data poor field with inadequate data preservation and reuse practices, and with relatively less advanced data collection instrumentation (Sawyer 2007). Consequently, data poverty has inhibited cumulative theory development and has limited both the scope and scale of typical research projects. Therefore, for example, studies with an individual as the unit of analysis and a single snapshot data collection (Avital 2000) largely dominate our research landscape in spite of its organizational orientation and the longitudinal nature of the underlying phenomena. Another artifact of the data poverty has been the over reliance on case studies, which are useful for exploratory research, revelatory research and theory generation, but incapable of long-term and integrated theory development let alone confirmatory and strong generalizable results.

Data poverty in the IS discipline largely stems from the prevailing norm that promotes only sharing of findings through publication of peer-reviewed research papers. This norm inhibits largely sharing of datasets and the instruments used to collect it. Yet, the preservation, sharing and reuse of data of all sorts are critical for continuous development of any scientific field. This has been demonstrated by disciplines such as astronomy, physics, geology, history, archeology and ocean sciences (to name a few), which have traditions of building up and sharing cumulative datasets. The work of physicists and economists makes clear that building large-scale datasets and treating them as a community asset has a strong cumulative positive effect on the sophistication of research methodologies in use and the community's ability to make sense and discover the world.

The IS research community's current lack of attention towards developing of large-scale cumulative data on IT-related subject matters may have contributed to the emergence of alternative data brokers that responded to the market needs. Capitalizing on a growing demand from commercial enterprises and government agencies, and through systematic data collection over time, several large consultancies (e.g. IDC, Gartner) have become the de facto brokers controlling IT-related data and their analyses. However, with hefty use charges, limited access to raw data, and often unknown (and thus suspect) data and instrumentation quality, such services are difficult to apply for groundbreaking research and cumulative knowledge building. This begs the question: **can we afford staying a data poor field?**

In contrast, data rich fields such as labor economics or health sciences afford their respective communities a wealth of data through publicly accessible normalized datasets. Building on the digitization of data and availability of low cost storage and transmission, many fields are moving quickly to become highly or ultra data rich (e.g., astronomy, oceanography, biology, high energy physics) through the execution of mega-scale charting projects, such as mapping the sky[1], the ocean[2] or the human genome[3] which result in petabytes or even zettabytes of data. With the ever-growing digital footprint available through the internet traffic, telecommunication records and similar information infrastructures, much of the IT-related data have become much easier to gather and share. Should we, as a community of scholars, be interested similarly in generating large-scale open access datasets of cumulative information about IT use behaviors and structures?

In spite of our contention that IS discipline is a data poor field, some promising opportunities and approaches toward data enrichment have emerged recently. First, with the growing salience of IT for the economic performance of firms, public data sources are beginning to yield data on phenomena of interest to IS researchers (e.g., IT-enabled outcomes in event studies and CIO compensation). Second, with the growing digitization of business models and business processes, IS researchers are finding opportunities for data mining and creation of "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) are beginning to offer opportunities for harvesting data on issues of interest to IS researchers. Nonetheless, awareness of these opportunities and their use in research is limited to a relatively small subset of the IS research community. Further effort to diffuse awareness of the available data enrichment opportunities and to develop new data sources is required.

Data Ecology Archetypes

All scientific enterprises are built on evidence, which often labeled as data. The nature of a scientific field reflects to a large degree its data ecology, which is a result of the kind of data explored and the way data is generated, refined, replenished, distributed, organized and generalized.

To provide a simple organizing framework to represent a field's approach to data, we identify four archetypes of data generation and distribution in a space comprising two primary dimensions: data complexity and data availability (Figure 1). *Data complexity* on the Y-axis refers to the properties of a dataset ranging from a few attributes that were collected in a single snapshot from a small number of data points to many attributes that have a temporal dimension and were collected from a large number of data points. *Data availability* on the X-axis refers to the data preservation and possible reuse ranging from token availability and normative data and instruments hoarding to open access and normative sharing. Subsequently, four distinct archetypes of data ecology emerge:

[1] http://www.sdss.org
[2] http://www.oceanservice.noaa.gov/welcome.html
[3] http://genomics.energy.gov

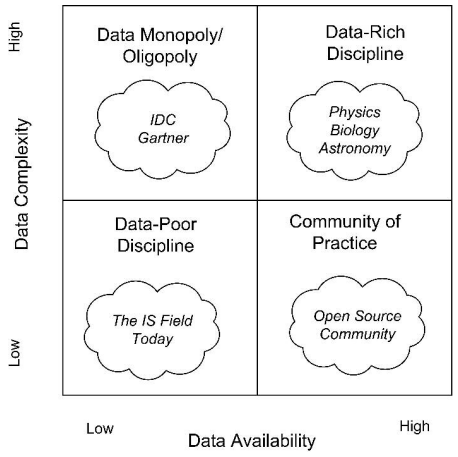

Figure 1. Archetypes and examples of data generation and distribution

- *Data-poor field* – populated with individuals or small groups that collect their own data. Datasets are often small due to limited resources and incompatible with similar data due to propriety instrumentation. Data availability is limited. (e.g., the IS field today)
- *Data monopoly/oligopoly* – populated with large companies or agencies that collect and analyze systematically large datasets for resale or other for-profit activities. (e.g., ITU, IDC, Gartner, OECD, US Census)
- *Community of practice* – populated with individuals or small groups that collect their data and share it voluntarily with their respective community. Datasets are often small but with good compatibility with similar data due to transparency and exchange of data collection instruments and cumulative build-up. (e.g., F/LOSS, NASA SEL, Open Courseware)
- *Data-rich field* – populated with teams that systematically collect large datasets over time as a joint effort in predefined well-funded projects. Datasets are often huge and complex, and made available to all interested stakeholders often supported by institutional endowment or policies. (e.g., Genetics championing the Human Genome Project)

Perspectives on Data Rich Scholarship

Specifically, the panelists will address the underlying topic from five perspectives:

Community of Practice Approach

Sharing data publicly should become part of the normative evidence required to support a published article (i.e., making data available must be part of publishing). Taking this idea to the next level would be developing data preservation and sharing infrastructure in the form of a centralized repositories. Subject to certain quality control mechanisms similar to those established in the open source community, such repositories should allow accumulating data both by adding normalized data points to an existing set and by extending datasets with new type of data.

Economical Approach

Economies of scale and economies of scope can yield higher returns on investments in data. Consequently, we need to recognize the criticality of data and make its production a bona fide specialization, as opposed to merely a general research skill expected from all. For example, building on the prevailing practice in Physics, we can create ecology of symbiotic relationships between distinct groups of data collectors and data analysts/theorists.

Data Mining Approach

The plethora of digital data that can be captured combined with the increasing processing capabilities and accuracy of data analysis create new opportunities for advancing IS research. Building on our extensive experience in consumer-oriented databases, pattern recognition, data visualization, modeling, and data warehousing, data mining and knowledge discovery tools and techniques can be harnessed for knowledge building and theory development.

Political Approach

Information technology is at the heart of the current societal change and it should be a public interest to collect and maintain more accurate and better records of its use, primary and secondary effects, returns on investment, growth and so forth. Building institutions that harvest and maintain grand datasets similar to the one generated in the genome project will help not only to develop and extend the knowledgebase of the field but also to attract the financial resources and human capital that can lift it to the next level. Furthermore, collecting large datasets and generating valuable findings will enhance IS research visibility and hopefully its disciplinary legitimization.

Constructionist/Critical Approach

The devil's advocate position argues that data by itself does not matter—only its interpretation matters. Any collected repository implies an inherent interpretation that is embedded in the data structures and what eventually is included and excluded in the dataset. Therefore, given the a priori set of paradigmatic assumptions and implied theories that are embedded in any dataset, overly controlled centralized data repositories may actually inhibit innovation and overall advancement of knowledge. Moreover, building on the critical theory position, we can argue that data repositories serve the dominant culture and are used by power elites to control radical change and inhibit non-conformist innovation that is not desirable by those who benefit from the status quo.

Discussion Format

Aligned with the conference theme--"Diversity in IS research and practice," the panelists will offer different perspectives on data and focuses on the state of IS research and practice. Each panelist will be limited to a ten minute opening statement so the majority of the time for the panel will be for discussion with the audience.

- Michel Avital - Panel overview, rationale, aspirations and possible strategies for data enrichment
- Vallabh Sambamurthy -Shared data and instruments repositories to complement publishing in IT topics
- Kenneth L. Kraemer - Projects involving large-scale cross-sectional and longitudinal databases
- Suzanne Iacono – Government and NSF role in projects involving the development of large-scale databases
- Steve Sawyer - Social and organizational informatics aspects of data repositories
- Q&A, open discussion, audience comments and suggestions. We seek to engage the audience in the conversation and to allow as many exchanges as possible within the allotted time.
- Kalle Lyytinen - summarizes the emerging themes and concludes the session with a discussion of its implications and potential for scholarly endeavor in the IS field.

At minimum, we hope that the panel will stimulate new thinking about the role of data beyond its immediate application and its long-term implication for the discipline. We hope to transform the discussion in the ICIS panel into a sustainable discourse that will serve the community of Information Systems researchers at large.

Participants

Michel Avital is an Associate Professor of Information Management at the University of Amsterdam. Building on positive modalities of inquiry, his research focuses on information and organization with an emphasis on the social aspects of information technologies. He has an interest in generative design, collaborative systems development methodologies, knowledge sharing in heterogeneous environments, and unconventional research methods and methodologies.

Suzanne Iacono is a Senior Science Advisor in the Directorate for Computer and Information Science and Engineering (CISE) at the National Science Foundation (NSF) and Division Director (Acting) for the Division of Computer and Network Systems (CNS). Previously, she was the Division Director (Acting) for the Division of Information and Intelligent Systems in CISE, the head of the Information Technology Research (ITR) Program and Program Director for Digital Society and Technologies in CISE. She also has interagency duties and serves various IT-related committees. Prior to coming to NSF, she held a faculty position at Boston University, was a Visiting Scholar at the Sloan School, MIT, and was a Research Associate at the Public Policy Research Office at the University of California, Irvine. Over the years, she has written journal articles, book chapters and conference papers on Social Informatics, an area of interdisciplinary research and education that integrates aspects of computer and social sciences. Suzi received her PhD from the University of Arizona.

Kenneth L. Kraemer is Professor of Information Systems and Director of the Center for Research on IT and Organizations, at the Paul Merage School of Business, University of California, Irvine. His research interests include the social implications of IT, national policies for IT production and use, and the contributions of IT to productivity and economic development. His recent book is *Globalization of E-Commerce* (Cambridge University Press, 2006). He is engaged new work on the offshoring of knowledge work and who captures the value from innovation radical and incremental innovations. Kraemer has been engaged in four multinational projects involving large-scale cross-sectional and longitudinal databases.

Kalle Lyytinen is Iris S. Wolstein Professor of Information Systems in Case Western Reserve University. He is the Chief Editor of JAIS and has served on the editorial boards of several leading IS journals including MISQ, ISR, EJIS, JSIS and many others. He has published over 70 articles and edited or written ten books. His research interests include critical theory, information system theories, system design, computer supported cooperative work, and diffusion of complex technologies. He is particularly interested in large-scale data and research infrastructures, standards and evolution of disciplines based on their research instrumentation.

Vallabh Sambamurthy is the Eli Broad Professor of Information Technology and the Executive Director of the Center for Leadership of the Digital Enterprise at the Eli Broad College of Business at Michigan State University. He is the editor-in-chief of ISR and serves on the editorial board of Management Science. He has served on the editorial boards of numerous journals, including MISQ, JSIS and IEEE Transactions on Engineering Management. He has researched issues related to the impacts of leadership and institutional forces on organizational IT assimilation, and the capabilities and factors associated with strategic leverage of IT.

Steve Sawyer is a founding member and an associate professor at the Pennsylvania State University's College of Information Sciences and Technology. Steve holds affiliate appointments in the department of Management and Organization; the department of Labor Studies and Employer Relations; and the program in Science, Technology and Society. Steve does social and organizational informatics research with a particular focus on people working together using information and communication technologies.

References

Avital, M. "Dealing with Time in Social Inquiry: A Tension between Method and Lived Experience," *Organization Science*, (11: 6), 2000, pp. 665-673.

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

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.

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.

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.

Sawyer, S. "Data Rich Fields, Data Poor Fields and Information Systems," A colloquium presentation at Case Western Reserve University, 30 March, 2007.