12-8-2010

Data Matters: An Analysis of Data Practices in IS Research

Wietske Van Osch
University of Amsterdam, w.vanosch@uva.nl

Michel Avital
University of Amsterdam, michel@avital.net

芽|Sprouts Working Papers on Information Systems ISSN 1535-6078

Data Matters: An Analysis of Data Practices in IS Research

Wietske Van Osch
University of Amsterdam, Netherlands

Michel Avital
University of Amsterdam, Netherlands

Abstract
Based on content analyses of 653 articles published between 2004 and 2008 in five key journals in the IS field, we demonstrate that the IS discipline exhibits data poverty characteristics as evident by the prevailing data sharing and reuse practices. As a result, the scope and scale of IS research is inhibited and the likelihood of generating far-reaching salient results is low. Finally, we discuss the implications to the IS discipline and draw recommendations for shifting the related practices from a vicious circle of data poverty to a virtuous circle of data richness.

Keywords: data sharing, data reuse, data repositories, IS research agenda,IS discipline

Permanent URL: http://sprouts.aisnet.org/10-119

Reference: Van Osch, W., Avital, M. (2010). "Data Matters: An Analysis of Data Practices in IS Research ," Proceedings > Proceedings of SIGPrag Workshop . *Sprouts: Working Papers on Information Systems*, 10(119). http://sprouts.aisnet.org/10-119

Data Matters: An Analysis of Data Practices in Information Systems Research

Wietske van Osch
University of Amsterdam
The Netherlands
w.vanosch@uva.nl

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

Abstract

Based on content analyses of 653 articles published between 2004 and 2008 in five key journals in the IS field, we demonstrate that the IS discipline exhibits data poverty characteristics as evident by the prevailing data sharing and reuse practices. As a result, the scope and scale of IS research is inhibited and the likelihood of generating far-reaching salient results is reduced. Finally, we discuss the implications for the IS discipline and derive recommendations for shifting the related practices from a vicious circle of data poverty to a virtuous circle of data richness.

Introduction: Background and Motivation

In the past few years, we have witnessed an emerging attention to data as a critical resource of the scientific enterprise and the need to reassess related practices and policies in response to the digital revolution, which altered fundamentally the ways data is created, distributed, and used. Whereas some academic disciplines (e.g., astronomy, oceanography, biology, high energy physics) have taken advantage of the new opportunities through the execution of mega-scale charting projects[1], it has been argued that the IS discipline has not kept up with the changes in spite of its natural affinity with the digital revolution (Avital et al., 2007).

Subsequently, it was suggested that the current practices of producing, maintaining and using data assets in the IS discipline result in economic deficiency, research ineffectiveness, erosion of social legitimacy, and missed opportunities. Moreover, it was proposed that systematic approaches to data collection and sharing practices, 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 the IS discipline as a thriving scholarly field of study (Avital et al., 2007; Lyytinen, 2009.)

The previous attempts to address the underlying issues of data-related practices and policies (see: Avital et al., 2007; Lyytinen, 2009) have argued eloquently about the merits of data rich and the pitfalls of data-poor scholarly disciplines. Furthermore, it was submitted without any substantive evidence that the IS discipline is a data-poor field with inadequate data preservation and reuse practices. This paper aims to close this gap by offering some preliminary empirical evidence about the current state of the IS discipline with respect to data-related practices based on longitudinal analysis of leading IS journals. Hence, the scope of this paper is limited to collecting and presenting rather than interpreting aggregated information about data integrity, sharing and reuse in order to prompt further grounded discussion about data richness and poverty in the IS field.

The main objective of the following analysis is to provide an evidence-based balance sheet that represents prevailing data-related research practices in the IS discipline. The analysis builds on the conclusions of the National Academy of Sciences Committee on Ensuring the Utility of Research Data in a Digital Age (2009). The committee stipulated

[1] For example, mapping the human genome (http://genomics.energy.gov), the sky (http://www.sdss.org), or the ocean (http://www.oceanservice.noaa.gov/welcome.html).

three broad areas of concern with respect to research data in the digital age, as follows: ensuring the integrity of research data, ensuring access to research data, and promoting stewardship of research data. Subsequently, we focused our analysis on these facets of research data in the IS discipline and operationalize it as follows:

- ***Integrity***, which refers to data and instrument quality
- ***Access***, which refers to data and instrument sharing
- ***Stewardship***, which refers to data and instrument reuse

Granted, these evaluation criteria are largely based on the positivistic view of data and do not apply well to some other views, such as the critical or design science views of research. However, given that approximately 75% of the papers in the IS discipline are based on a positivist view (Mingers, 2003; Nandhakumar and Jones, 1997; Orlikowski and Baroudi, 1991), we can appropriately use these criteria for assessing the majority of papers included in our data set.

An analysis of the published papers in five key IS journals between 2004 and 2008, suggested that the IS discipline indeed demonstrates data poverty characteristics as evident by the prevailing low rates of data sharing and reuse, which in turn erode the overall data integrity and potential impact of the discipline. It is of interest to members of the Special Interest Group on Pragmatist IS research[2], that the findings can provide a basis for a better understanding of the prevailing data practices in the IS discipline and for further development of data-related policies by journal editors, funding agencies, and other policymakers who aim to facilitate and foster scholarly exchange.

In the remainder of this paper, we describe the research approach, provide a detailed account of the findings, and discuss some of the implications in the context of research practices and a virtuous cycle of data richness.

Research Approach

Overall, we aimed to provide an evidence-based indication of the current state of the IS discipline with respect to data-related practices. In an attempt to balance breadth of analysis, temporal span, and resource constraints, we analyzed a subset of the IS literature that comprised a set of published articles over five years, from 2004 to 2008, in five key IS journals. The selected journals appear at the top of the various journal-ranking lists and publish primarily mainstream research, which seemed to be authored by a representative set of IS scholars. The selected journals were as follows: Information Systems Journal (ISJ), Information Systems Research (ISR), Journal of the Association for Information Systems (JAIS), Journal of Management Information Systems (JMIS), and MIS Quarterly (MISQ).

Building on the National Academy of Science's three dimensions of concern with respect to research data, namely integrity, access, and stewardship, we developed a coding scheme that focuses on data and instruments general qualities, sharing, and reuse respectively (see Appendix 3). Data coding was conducted by pairs of two independent coders. An initial interrater reliability estimate of 86% percentage agreement based on a subset of articles[3] provided an initial assessment of the coding scheme validity and coding process reliability.

From the total of 653 published papers, 444 (68%) used data to substantiate their knowledge claims, and the rest primarily encompassed conceptual papers and papers proposing test models for future research. The number of empirical papers out of the total published papers varies substantially per journal (see Appendix 1). For instance, ISR has the highest percentage of empirical papers (78.4%) and JAIS has the highest percentage of conceptual papers; therefore about half of all papers (49.2%) are empirical. If we analyze the data by year (see Appendix 2), we can see that there is a gradual increase in the proportion of empirical papers from 57.8% in 2004 to 76.4% in 2008 (with the exception of 2007, where it went down to 62.4%). For the remainder of the analysis of data practices, we used only the 444 empirical papers.

[2] http://www.sigprag.org/index.html

[3] Coding was conducted in two steps. For each journal, published papers of 2008 were coded independently by two coders, who then negotiated and reconciled their respective differences. This calibration phase yielded an average interrater reliability of 86%. Then, the remainder of the papers was split between the coders.

Findings

Integrity: Qualities of Data and Instruments

The first set of findings pertains to data integrity, as represented by data type and data complexity with respect to number of observations, number of data collection points (cross-sectional vs. longitudinal) and unit of analysis (Avital, 2000; Avital *et al.* 2007; Lyytinen, 2009). Assessing the overall quality of the field's repository of research data and instruments in terms of intrinsic – let alone external – validity is an ambitious undertaking that is outside the scope this study. However, in line with the objectives of this study, we opt to examine the type and complexity of the overall data in the IS field as an indication for the available data resources and the means to its production. Furthermore, data type and complexity also provide an indication for the kind of phenomena under investigation, the extent of the possible knowledge claims, and an opportunity to assess the alignment between the scope of overall research outputs and the intended scope of the body of knowledge aspired to.

From the 444 empirical papers, the majority is quantitative (77.5%) with qualitative papers accounting for 57.9% of all empirical papers. The overlap implies that 35.4% of the papers use some sort of combination of quantitative and qualitative. However, only 18.9% of all papers were coded as using deliberately a mixed method approach to data collection. There are a few exceptions, as follows: ISJ (44.8% quantitative and 100% qualitative methods) and JAIS (63.4% quantitative and 74.6% qualitative methods). Assessing the data by year rather than journal shows that 2007 is the only year in which there were slightly fewer quantitative than qualitative papers (65.4% versus 66.6%), the difference being negligible. With respect to the specific utilization of methods, the record shows that surveys (41.2%) followed by interviews (26.4%) account for the majority of methods used.

One interesting finding is the fact that from the 444 empirical papers, only 407 (91.7%) mentioned the number of observations *N*, with ISJ having the highest percentage of papers mentioning *N* (94.0%) and JMIS the lowest percentage of papers reporting *N* (88.3%). This item was coded not only by searching for explicit references to *N*, but also by analyzing tables or figures that might help to find or calculate the total *N*. In articles using interviews, only 65.8% mentioned the total number of interviews conducted and an even smaller number of papers (38.5%) mentioned the duration of the interviews. Out of the articles mentioning *N*, the median[4] number of observations is 121. However, the differences between journals are quite substantial, with ISJ having a median *N* of 21—which can be attributed to the fact that the majority of papers adopted qualitative methods. Only (55.2%)[5] and almost half of all papers (49.3%) used interviews—ISR and JMIS having a median *N* of 192 and 188 respectively, due to the high percentage of quantitative data papers in these journals (81.0% and 89.3% respectively).

The majority of all empirical papers were cross-sectional (84.5%), hence only a small percentage of papers (15.5%) were based on longitudinal data, with JAIS having the highest percentage of longitudinal studies, namely 22.2%, and JMIS the lowest percentage, namely 10.7%, which is likely related to the large share of papers relying on survey data (40.4%). Analyzing the results by year rather than by journal indicated that 2004 had the lowest percentage of longitudinal studies, namely 10.4%, and 2005 had the highest percentage, namely 23.3%.

With respect to the unit of analysis, more than half of all empirical papers (59.7%) used individuals as a unit of analysis, followed by organizations (25.9%), by groups (12.2%) and other (2.2%). The only exception was ISJ in which nearly half of all empirical papers (46.3) used organizations as the unit of analysis. Interestingly, ISR had more papers on groups (13.2%) than on organizations (11.0%), yet the individual represented the most dominant unit of analysis (76.9%). All other journals followed the general pattern.

In summary, with respect to data type and data complexity the IS discipline displays several characteristics of data poverty. First of all, the overreliance of the field on survey-based and interview-based research is poorly suited for integrated theory development and confirmatory research with strong, generalizable results (Lyytinen, 2009). Second, the field is still dominated by studies using cross-sectional data (i.e. single snapshot sampling) and individuals as the unit of analysis (Avital, 2000). This is despite the fact that most of the complicated process

[4] The median was selected to be the most appropriate summary statistic as the mean was too sensitive to extreme cases (such as the high *N* in simulation-based data papers).
[5] And another 44.8% using qualitative methods to supplement, precede or follow quantitative methods.

phenomena under investigation require longitudinal data, and that the IS discipline is increasingly analyzing topics that require a group-level, organizational-level or multi-level focus.

Access: Data and Instrument Sharing

The second set of findings pertains to access as represented by data and instrument sharing.

With respect to data sharing, we analyzed whether quantitative papers provided access to the dataset in any usable form, for example as a covariance matrix, whether the papers provided a detailed description of the data, and whether they provided a link to the original data. Overall, 82% of all quantitative papers provided a covariance/correlation matrix, with all (100.0%) quantitative data papers in ISJ and MISQ providing a covariance/correlation matrix and a mere 55.4% of quantitative papers in JMIS providing such a data summary matrix.

We examined whether the papers provided a detailed description of the data, and looked for any elaborated description of properties or characteristics of the dataset. Less than half (48.6%) of all empirical papers provided detailed information about data, with the exception of JMIS where 69.9% provided detailed information about the data, and MISQ where only 22.5% of all papers provided detailed information about the data.

A minor share (4.3%) of all empirical papers provided a link to the original data, with the exception of JAIS in which 9.5% of all papers provided a link to the original dataset. In ISR and MISQ only 1.1% and 1.7% respectively of all papers provide a link to the original dataset.

With respect to instrument sharing, we investigated whether papers provided detailed information about the instrument used, such as reliability and validity, as well as detailed descriptions of how instruments were developed, tested, and applied. A little over half (55.9%) of all papers provided detailed information about the instrument used, with the exception of JMIS in which 71.8% of the papers provided detailed information regarding the instrument used. We also examined whether or not papers provided the instrument used (e.g. an appendix including survey or a web link to the survey). Only a quarter (25.2%) of all papers provided the instrument, with the exception in this category of MISQ where only 14.2% of all papers provided the instrument.

In summary, with respect to data and instrument sharing too, the IS discipline displays several characteristics of data poverty. First, only a small portion of the published papers provide detailed information about their data, let alone providing access to the original data. It is apparent that the IS discipline lacks a normative infrastructure to data sharing and the systematic approach to support it. Second, given that only a quarter of all papers provided the instruments used, opportunities for reusing and subsequently revalidating or refining the instruments in the IS discipline are limited. Hence, we argue that the gatekeepers, who promote sharing of findings through peer-reviewed research papers without paying attention to the simultaneous sharing of datasets and the instruments used to collect it, also contribute to the data poverty in the IS field.

Stewardship: Data and Instrument Reuse

The third and last set of findings pertains to stewardship as represented by data and instrument reuse.

Data reuse was assessed by looking at the number of papers that used self-collected data and the number of papers that used secondary data sources. A majority (81.3%) of the empirical articles were based on self-collected data. Only a fifth of all papers (19.4%) used secondary data as the primary data source. This pattern applied to all journals, although for JAIS it was more equally distributed (58.7% vs. 41.3%). One of the most alarming findings was that from all 86 papers using secondary data, only 53 (61.6%) mentioned the data source used.

Instrument reuse was assessed by examining the number of papers that used or adapted existing instruments for data collection. We found that 34% of all papers used instruments that were developed in previous work in their original form, and another 29.7% of all papers adapt existing instruments to their specific research purpose or context. On average, approximately a third of all papers (36.3%) relied on self-developed instruments for data analysis. The exception is JAIS, in which 71.4% of all papers used instruments that were self-developed. The latter is likely to be related to the large percentage of qualitative papers in this journal.

In summary, with respect to data and instrument reuse, the IS discipline displays again characteristics of data poverty. Even when possible and methodologically encouraged, reuse of instruments and capitalizing on existing data is not prevalent. Hence, the lack of instrument availability forces IS researchers to re-craft instruments again and again rather than to build on an accumulated pool of instruments. Moreover, progress is inhibited if the IS discipline does not accumulate datasets over time and from multiple sources, as well as if IS researchers do not bother to refine data collection instruments over time and in multiple contexts. Developing richer and cleaner datasets as well as more accurate and reliable instruments require a concentrated community effort. Hence, lack of stewardship undermines the position and status of the IS field.

From Data Poverty to Data Richness

In the previous section we have demonstrated that concerns about data poverty in the IS discipline as raised by Avital *et al.* (2007) and Lyytinen (2009) are valid. The IS discipline indeed displays characteristics of data poverty, at least with respect to data integrity, access, and stewardship. In all, this paper provides three key contributions. First, by empirically assessing the three dimensions of concern with respect to data practices, we were able to evaluate and validate prior claims about data poverty in the IS field. Second, by empirically demonstrating characteristics of data poverty in the IS discipline, we hope to promote a focus on data practices and data policies as crucial foundations for establishing the IS community as an institutional force. Third, in what follows, we also suggest an overall process model that portrays the interrelations between the three suggested dimensions of data. All in all, we would like to emphasize that the main objective and contribution of this paper is not the discussion of the relationships among these dimensions of data, but rather the set of preliminary observations and suggestions that can open up further dialogue in the IS field. Hence, overall, this paper aims to provide the data and evidence base for future evaluations and further discussions of the role of data in the IS field.

The report of the National Academy of Science's Committee on Ensuring the Utility of Research Data in a Digital Age (2009) stipulated three broad areas of concern and a set of directives with respect to research data and related work practices. Building on our study, we would like to stress the interrelationships among these three dimensions—data integrity, access, and stewardship.

Subject to the prevailing norms and work practices, the highly interrelated dimensions can invoke either a vicious circle that reinforces data poverty or a virtuous circle that reinforces data richness, as illustrated in Figure 1 below. Specifically, high quality data provides the basis of data integrity and for successful publishing, which in turn provides an opportunity for data sharing via publically accessible channels. The availability of trustworthy data encourages reuse and overall stewardship with respect to data. This is the virtuous circle that promotes data rich disciplines. In contrast, low quality data erodes data integrity and reduces the likelihood of publishing, which in turn diminishes the possibility of data sharing via publically accessible channels. The lack of trustworthy data inhibits reuse and encourages self-reliance that undermines stewardship with respect to data. This is the vicious circle that upholds data poor disciplines. The preceding description holds in a similar fashion for instruments as well.

Given the high interrelatedness of these dimensions, it is necessary to address all three simultaneously and in a coordinated fashion. Addressing only one or two dimensions is unlikely to foster the desired ecology of data richness for the IS discipline. Furthermore, we would like to acknowledge the two-way nature of the relationships between the three dimensions. In other words, for example, the extent of data sharing and availability has a positive effect on data integrity, which in turn has a positive effect the extent of data reuse, and so on.

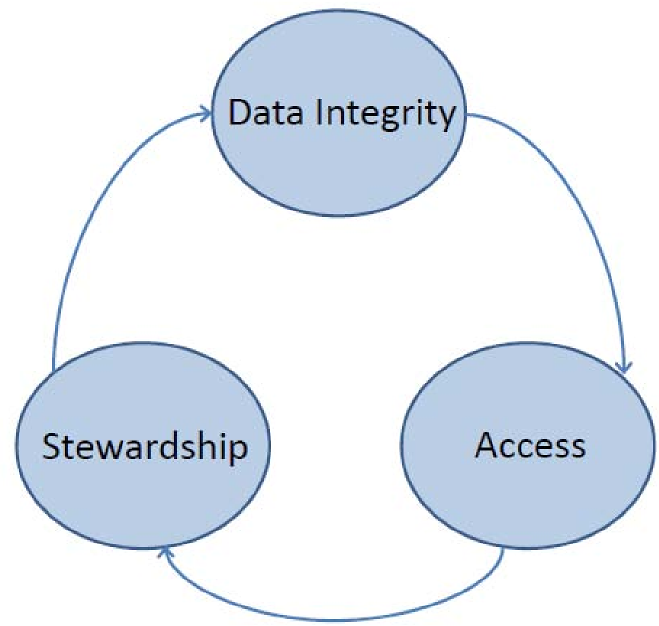

Figure 1: A Virtuous Circle of Practices in a Data Rich Discipline

Challenges and Future Research

For this preliminary analysis of data practices in the IS field, we relied on the criteria for assessing data richness and poverty as developed in the report of the National Academy of Science's Committee on Ensuring the Utility of Research Data in a Digital Age (2009). These criteria are based on a positivistic view of data which is appropriate only for the majority of papers in the IS discipline (Mingers, 2003; Nandhakumar and Jones, 1997; Orlikowski and Baroudi, 1991), but not for all the prevailing brands of research. Future attempts to assess data in the IS field should look for additional and/or alternative criteria that can be used for studies that are more interpretivist, critical or design-oriented.

Conclusion

Building on a longitudinal analysis of papers in five key IS journals, this study reinforced the prevailing concerns that the IS discipline demonstrates characteristics of data poverty as evident from poor research data-related practices and a vague normative infrastructure, which result in subpar integrity, sharing, and reuse of data and instruments. The findings also suggested that the highly interrelated dimensions of data—integrity, access, and stewardship—can invoke either a vicious circle that upholds data poverty or a virtuous circle that promotes data richness.

This ad hoc study provided a clear indication about the state of the IS discipline with regard to data-related practices. However, moving onward from diagnosis to prescription requires additional studies to develop explicit quality criteria, reporting requirements in terms of format and depth, preservation procedures, sharing channels and markets, and similar mechanisms and policies.

All in all, we hope that this paper will contribute to the awareness of the IS community of the fundamental role of data with respect to its scholarly impact, legitimacy, overall viability and future prospects.

Acknowledgment

We thank Robert de Boer, Arnold Commandeur, Steven de Hertogh, Sander Huizinga, Ileana Maris, Eric Onderdelinden and Yang Song wholeheartedly for their dedication and help with coding the dataset.

References

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

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

Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age; National Academy of Sciences. 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*, The National Academies Press, Washington, DC.

Lyytinen, K. 2009. "Data matters in IS theory building," *Journal of the Association for Information Systems*, (10:10), pp. 715-720.

Mingers, J. 2003. "The Paucity of Multimethod Research: A Review of the Information Systems Literature," *Information Systems Journal*, 13: 233-249.

Nandhakumar, J. & Jones, M. 1997. "Too close for comfort? Distance and engagement in interpretive information systems research," *Information Systems Journal*, 7, 109–131.

Orlikowski, W. & Baroudi, J. 1991. "Studying information technology in organizations: research approaches and assumptions," *Information Systems Research*, 2, 1–28.

Appendices

Appendix 1: Synopsis of summary data tables by journal

	Pool		Data Type		Data Complexity				
Journal	**Total Papers**	**Data Papers**	**Quant**	**Qual**	**N**	**longitudinal**	**ind**	**group**	**org**
ISJ	95	67 (70.5)	30 (44.8)	67 (100.0)	63 (94.0)	8 (11.9)	21 (31.3)	10 (14.9)	31 (46.3)
ISR	116	91 (78.4)	94 (81.0)	30 (33.0)	83 (91.2)	19 (20.9)	70 (76.9)	12 (13.2)	10 (11.0)
JAIS	128	63 (49.2)	40 (63.4)	47 (74.6)	58 (92.1)	14 (22.2)	38 (60.3)	7 (11.1)	14 (22.2)
JMIS	152	103 (67.8)	92 (89.3)	38 (36.9)	91 (88.3)	11 (10.7)	61 (59.2)	12 (11.7)	26 (25.2)
MISQ	162	120 (74.1)	88 (73.3)	75 (62.5)	112 (93.3)	17 (14.2)	75 (62.5)	13 (10.8)	34 (28.3)
Total	**653**	**444 (68.0)**	**344 (77.5)**	**257 (57.9)**	**407 (91.7)**	**69 (15.5)**	**265 (59.7)**	**54 (12.2)**	**115 (25.9)**

	Data Reuse		Instrument Reuse		Data Sharing		Instrument sharing	
Journal	**Self Collected**	**Secondary**	**"as is"**	**adapted**	**General cov/cor**	**details**	**Instrument**	**Link to instrument**
ISJ	52 (77.6)	16 (23.9)	23 (34.3)	24 (35.8)	30 (100.0)	32 (47.8)	37 (55.2)	17 (25.4)
ISR	74 (81.3)	21 (23.1)	23 (25.3)	41 (45.1)	77 (81.9)	44 (48.4)	51 (56.0)	26 (28.6)
JAIS	37 (58.7)	26 (41.3)	7 (11.1)	11 (17.5)	36 (90.0)	41 (65.1)	27 (42.9)	18 (28.6)
JMIS	91 (88.3)	9 (8.7)	44 (42.7)	26 (25.5)	51 (55.4)	72 (69.9)	74 (71.8)	34 (33.0)
MISQ	107 (89.2)	14 (11.7)	54 (45.0)	30 (25.0)	88 (100.0)	27 (22.5)	59 (49.2)	17 (14.2)
Total	**361 (81.3)**	**86 (19.4)**	**151 (34.0)**	**132 (29.7)**	**282[6] (82.0)**	**216 (48.6)**	**248 (55.9)**	**112 (25.2)**

[6] Note that this percentage is calculated by the ratio of papers presenting covariance/correlation matrices (282) out of total amount of quantitative papers (so 344) instead of total amount of empirical papers (444)

Appendix 2: Synopsis of summary data tables by year

	Pool		Data Type		Data Complexity				
Year	**Total Papers**	**Data Papers**	**Quant**	**Qual**	**N**	**longitudinal**	**ind**	**group**	**org**
2004	116	67 (57.8)	58 (86.5)	44 (65.7)	60 (89.6)	7 (10.4)	34 (50.7)	9 (13.4)	19 (28.4)
2005	92	60 (65.2)	46 (76.7)	34 (56.7)	55 (91.7)	14 (23.3)	35 (58.3)	6 (10.0)	14 (23.3)
2006	159	116 (73.0)	93 (80.2)	54 (46.6)	111 (95.7)	14 (12.1)	91 (78.4)	9 (7.8)	35 (30.2)
2007	125	78 (62.4)	51 (65.4)	52 (66.6)	65 (83.3)	11 (14.1)	40 (51.3)	9 (11.5)	22 (28.2)
2008	161	123 (76.4)	96 (78.0)	63 (51.2)	116 (94.3)	23 (18.7)	65 (45.8)	21 (17.1)	25 (20.3)
Total	**653**	**444 (68.0)**	**344 (77.5)**	**257 (57.9)**	**407 (91.7)**	**69 (15.5)**	**265 (59.7)**	**54 (12.2)**	**115 (25.9)**

	Data Reuse		Instrument Reuse		Data Sharing		Instrument sharing	
Year	**Self Collected**	**Secondary**	**"as is"**	**adapted**	**General cov/cor**	**details**	**Instrument details**	**Link to instrument**
2004	61 (91.0)	5 (7.5)	15 (22.4)	20 (29.9)	30 (51.7)	53 (79.1)	50 (74.6)	29 (43.3)
2005	52 (86.7)	10 (16.7)	28 (46.7)	16 (26.7)	40 (87.0)	35 (58.3)	46 (76.7)	20 (33.3)
2006	76 (65.5)	38 (32.8)	43 (37.1)	23 (19.8)	65 (69.9)	59 (50.9)	56 (48.3)	27 (23.3)
2007	65 (83.3)	14 (17.9)	22 (28.2)	24 (30.8)	51 (100.1)	18 (23.1)	26 (33.3)	10 (12.8)
2008	107 (87.0)	19 (15.4)	43 (35.0)	49 (39.8)	96 (100.0)	51 (41.5)	70 (56.9)	26 (21.1)
Total	**361 (81.3)**	**86 (19.4)**	**151 (34.0)**	**132 (29.7)**	**282[7] (82.0)**	**216 (48.6)**	**248 (55.9)**	**112 (25.2)**

[7] Note that this percentage is calculated by the ratio of papers presenting covariance/correlation matrices (282) out of total amount of quantitative papers (so 344) instead of total amount of empirical papers (444)

Appendix 3: Code Book

Data Ecology Study Code Book

The following coding scheme is designed with the aim of gaining insight concerning data and the reuse of instruments as well as sharing, irrespective of research method.

Please code your assigned papers and store the work file in our share directory. We will run this exercise in two phases:

(I) Designated pairs will code individually the same subset of 2008 papers, and then compare results and reconcile their coding. (See Tasks Worksheet)

(II) Then, each person will code individually a subset of papers as assigned. (See Tasks Worksheet)

General Code Description

1. General Quality
   - 1A. Paper Characteristics
   - 1B. Data Type
   - 1C. Data Complexity
2. Reuse
   - 2A. Data reuse
   - 2B. Instrument reuse
3. Sharing
   - 3A. Data sharing
   - 3B. Instrument sharing

Detailed Code Description

1. General Quality (Data Type and Complexity)

1A. Paper characteristics

ID- Unique paper ID based the following format: JournalYearVolume-Issue (e.g., MISQ200704-1)

author – last name of first author

nauthor – number of authors

data – 1=yes, 0 =no (yes=data based paper; no=theory or other paper with no data)

→ if no data, then no further coding is required; move to the next paper

1B. Data Type *(mark multi-methods if applicable)*

survey (online or paper-based)- 1=yes, 0 =no

interview

Ethnography (observation)

Document analysis

database (mostly economics data)
quantitative other (specify in notes)
qualitative other (specify in notes)
Mixed sources - 1=yes, 0 =no (if yes, please specify all related)

1C. Data Complexity
N (number of observations)
individual (unit of analysis- 1=yes, 0 =no)
group
organizations
longitudinal (data collection over time, 1=yes, 0 =no)

2. Reuse (i.e. taking)

2A. Data Reuse
Self collected 1=yes, 0 =no
Secondary (reuse)
If secondary, source (name of source)

2B. Instrument Reuse
as is (used as is) 1=yes, 0 =no
adapted

3. Sharing (i.e. giving)

3A. Data Sharing
general cor/cov (provide general data relationship matrix) 1=yes, 0 =no
detailed (provide detailed data)
link (provide link to data)

3B. Instrument Sharing
Instrument (provide instrument, usually in Appendix)
link (provide link to instrument)

Remarks
Please note anything else unusual or unclear or any other additional remarks

芽|Sprouts
Working Papers on Information Systems | ISSN 1535-6078

芽|Sprouts