5-16-2023

DIGITAL TRACE DATA RESEARCH IN INFORMATION SYSTEMS: OPPORTUNITIES AND CHALLENGES

Bastian Wurm
LMU Munich School of Management, bastian.wurm@lmu.de

Oliver Müller
University of Paderborn, oliver.mueller@upb.de

Shaila Miranda
University of Oklahoma, shailamiranda@ou.edu

Yash Raj Shrestha
University of Lausanne, yashraj.shrestha@unil.ch

Michael Wessel
Copenhagen Business School, wessel@cbs.dk

See next page for additional authors

Authors

Bastian Wurm, Oliver Müller, Shaila Miranda, Yash Raj Shrestha, Michael Wessel, and Monica Chiarini Tremblay

DIGITAL TRACE DATA RESEARCH IN INFORMATION SYSTEMS: OPPORTUNITIES AND CHALLENGES

Panel

Bastian Wurm, LMU Munich School of Management, Germany, bastian.wurm@lmu.de

Michael Wessel, Copenhagen Business School, Denmark, wessel@cbs.dk

Monica Chiarini Tremblay, William & Mary, USA, monica.tremblay@mason.wm.edu

Michel Avital, Copenhagen Business School, Denmark, michel@avital.net

Philipp Hukal, Copenhagen Business School, Denmark, hukal@cbs.dk

Iris Junglas, College of Charleston, USA, junglasia@cofc.edu

Abstract

Digital trace data research is an emerging paradigm in Information Systems (IS). Whether for theory development or theory testing, IS scholars increasingly draw on data that are generated as actors use information technology. Because they are 'digital' in nature, these data are particularly suitable for computational analysis, i.e. analysis with the aid of algorithms. In turn, this opens up new possibilities for data analysis, such as process mining, text mining, and network analysis. At the same time, the increasing use of digital trace data for research purposes also raises questions and potential issues that the research community needs to address. For example, one key question is what constitutes a valid contribution to the body of knowledge and how digital trace data research influences our collective identity as a field? In this panel, we will discuss opportunities and challenges associated with digital trace data research. Reflecting on the panelists' and the audience's experience, we will point to strategies to mitigate common pitfalls and outline promising research avenues.

Keywords: Digital Trace Data, Computational Social Science, Computational Theory Development, Research Methods.

1 Introduction

Digital trace data are created when actors use information technology. Compared to other types of research data, such as survey data or interviews, digital trace data offer high levels of granularity and often cover a longer period (Howison, Wiggins and Crowston, 2011). Digital trace data typically cover information on who performed which action at what exact point in time, but can contain additional information, such as expressions of opinions in the case of Twitter data. Because digital trace data often share the properties of big data (volume, variety, and velocity), some researchers also refer to research that uses this type of data as big data research (Grover, Lindberg, Benbasat and Lyytinen, 2020).

An increasing number of studies demonstrate the potential of digital trace data for information systems research. For example, Avital et al. (2023) use trace data collected from the enterprise social media platform Yammer to study the social fabric of organizations. Godoy-Descazeaux et al. (2023) use YouTube data to reveal an assemblage of metaphors used to animate and make sense of quantum computing. Schirrmacher et al. (2021) conduct a netnography of DAO's communication channels to show how tokens shape work practices in fluid organizations. Another example is the study by Lindberg, Berente, Gaskin and Lyytinen (2016) that capitalizes on trace data from GitHub about the Rubinius

open-source project to investigate how open-source software developers coordinate work. All of these studies explore phenomena that are characterized by the entanglement of humans and digital technologies (e.g., Zammuto et al. 2007), and provide insights about the use of information technology that would be difficult or impossible to surface through other research methods.

The potential of digital trace data for information systems and its neighboring disciplines has prompted several editorials and methodological articles that address, for example, how human and machine capabilities can be combined for pattern recognition (Lindberg, 2020) and how patterns may contribute to computational theory construction (Miranda et al., 2022). Despite these efforts, however, there remain many option questions, especially for newcomers to this research genre. For example, how should one design and kick off a research project that aims to use digital trace data? What are the particularities with respect to data collection and data analysis? And how can we move from patterns observed in the data to theory? More generally, digital trace data research also calls us to reflect on our identity as a field. How can we build research programs that capitalize on digital trace data (Grisold et al., 2022) without neglecting the cumulative tradition (Baiyere, Berente and Avital, 2023)? Should IS scholars solely apply algorithms that computer science develops or should we aim to develop algorithms ourselves?

In light of these questions, this panel aims to continue and extend the discussion around the use of digital trace data and its role in information systems research. It brings together experts from data analytics, computational theory development, and machine learning, each providing a distinct perspective on the topic. In conversation with the audience, we want to discuss opportunities and challenges that arise from digital trace data research.

2 Questions and Issues in Digital Trace Data Research

In this section, we outline open questions and issues in digital trace data research, which will serve as a basis for the discussion at the panel. The discussion will be organized along three considerations that most empirical research projects have to take into account: 1) project set-up, 2) methodological choices, and 3) contributions to the body of knowledge. While in research projects these considerations are not made in isolation from one another, they help us structure the discussion on this complex topic.

2.1 Set-up of digital trace data research projects

At the start of their research project, involved scholars should reflect on whether digital trace data will likely lead to a more thorough understanding of a phenomenon than the application of other, more established research methods. For the application of many computational techniques, there are no guidelines available. This often leads to re-iteration and the probing of different techniques to decide on which one an analysis should be based. While most qualitative and quantitative techniques can neither be executed using a simple cookbook approach, in many cases there will be articles, textbooks, or experienced co-authors that can provide guidance. In this respect, digital trace data research projects are often more emergent and definitive guidelines and advice are scarce.

Now, if a scholar or a team thereof decides to pursue a research project on the basis of digital trace data, how should they go about when setting up a research team? Digital trace data projects require diverse competencies (Lazer et al., 2020). For projects in the IS domain, a research team will require at least three competencies. First, and different from other research projects, digital trace data research projects require strong technical competence. At least one person in the research team needs to be able to extract trace data via APIs, scrape data from webpages, and process data in such a way that they can be used for actual research. Second, at least one person needs to have methodological expertise, i..e. expertise of the specific computational technique(s) that should be applied. This does not only include know-how of the technical setup, but also knowledge on the advantages, disadvantages, and overall reliability of the respective technique. Third, the project team will need theoretical and domain expertise to make sense of the analyses and representations generated with the help of computational techniques. Of course, the exact competencies and their specificities will vary due to project characteristics.

Consequently, we should ask: What kind of team and what competencies are needed? Furthermore, what are best practices for managing such projects and involved competencies?

2.2 Data quality and methodological choices

Howison et al. (2011) emphasize that digital trace data are found, rather than produced. That is, these data "are a by-product of activities rather than produced by a designed research instrument" (Howison et al., 2011, p. 769). In other words, researchers typically have no influence on *what* data are recorded as well as *how* these data are recorded and stored. Researchers may only find out after an initial round of data extraction that the data do not capture a phenomenon they are interested in or that the data have quality issues. Thus, often only after data extraction one can assess whether and how the data can be used for research purposes.

Alternatively, a good starting point for an initial digital trace data research project can be open datasets. The number of open datasets is continuously increasing and they are becoming increasingly diverse in terms of coverage. Many governmental institutions and cities have open data initiatives in place and some research communities regularly publish datasets that scholars can use. In the Business Process Management community, for example, there is a yearly competition around a dataset that covers the performance of one or multiple business processes (e.g., van Dongen, 2017). We have to bear in mind, however, that only because data are publicly available, they are not necessarily correct (Alsudais, 2021). A potential solution to data quality issues could be that every dataset (public or private) has to undergo systematic data quality checks before it can be used for research. But how should we design, incentivize, and control such data quality checks?

Even once data quality is ensured, extracted data usually needs to be further processed for scholars to analyze it. For deductive research, different types of unstructured data (text, pictures, location, etc.) need to be transformed into variables. Researchers then need to reflect on whether their data are representative of the phenomenon in question and whether the data accurately capture what the research team wants to measure (Xu, Zhang and Zhou, 2020). One also needs to consider that some statistical approaches need to be tailored when applied to digital trace data (Qiao and Huang, 2021) and that due to large sample sizes variables tend to be significant even though effect sizes are comparably small (Lin, Lucas and Shmueli, 2013).

In terms of inductive and abductive research, computational techniques support researchers' capacity to detect patterns in large datasets (Lindberg, 2020). For example, Vaast, Safadi, Lapointe and Negoita (2017) iterate between computational and qualitative analysis of Twitter data to identify connective action episodes and understand how actors use social media for collective engagement. While the advantage of combining human and machine capabilities for data analysis is apparent, it is less clear how to achieve data-method fit, i.e., how to select a suitable computational technique given a dataset. Often data can be transformed, such that it can be analyzed with different algorithms. Thus, for a certain dataset, there might be multiple computational techniques that can be employed. How should one choose a respective method then?

Relating thereto, computer science constantly develops new techniques that can offer new insights by exploring data from another perspective. The more novel the method, the more difficult might be its application due to missing guidelines. It also might be more difficult to publish a piece that builds on a new technique, since reviewers do not know how to evaluate the respective work. Given these circumstances, what might be a suitable strategy for a team of authors fond of a particular method? One way forward might be a first methodological article that outlines the advantages and disadvantages of a computational method. Such an article needs to demonstrate how the respective technique allows other researchers to see the phenomenon from a different perspective and how this may extend or alter the existing knowledge on the phenomenon. For example, several authors have proposed that process mining can be used to examine and theorize about how organizational processes change over time, extending and possibly altering current knowledge on business process management and routine dynamics (Grisold, Wurm, Mendling and vom Brocke, 2020; Pentland, Vaast and Wolf, 2021).

2.3 Contributions to the body of knowledge

At the end of the day, each research project needs to make a contribution to the body of knowledge for it to be publishable. Yet, there is some controversy around what constitutes such a contribution, especially with respect to digital trace data research. For instance, Agarwal and Dhar (2014, p. 447) argue that "it is entirely possible that the contribution of a study lies primarily in the uniqueness of the data set and the rigor of the empirical methods used to analyze the data." In this respect, the ideas of *patterns* (Miranda et al., 2022) and *situated explanations* (Grisold et al., 2022) have gained increasing attention as they are more phenomenon-oriented and at a lower level of abstraction than mid-range or even grand theory. Indeed, papers that make use of large sets of digital traces tend to be less abstract and focus more on methodological aspects than on theory (Grover et al., 2020). Against this development, Baiyere et al. (2023) caution that IS should avoid clickbait research and not lose out of sight the cumulative tradition of the field.

Finally, a key question is whether the IS field should embrace other forms of contributions, such as the curation of open datasets and the development of novel algorithms (Grisold et al., 2022). As datasets and algorithms are particularly important for digital trace data research, IS needs to ensure that relevant datasets are publicly accessible and that algorithms produce reliable results that, in turn, can be used for theory development. Certainly, this would imply changes to editorial practices (Grisold et al., 2022), but neglecting these contributions leaves us a petitioner to computer science and related disciplines and will, in the long run, diminish our ability to carry out impactful digital trace data research projects.

3 Panel Organization

Bastian Wurm will serve as the panel moderator. He will steer the debate alongside the above-mentioned questions and issues. He will facilitate interaction with the audience by opening the debate for questions and remarks. Table 1 outlines the tentative schedule of the panel.

Part	Time Budget
1 Introduction	15 minutes
2 Panel discussion	30 minutes
3 Round table discussions	30 minutes
4 Joint reflection	15 minutes

Table 1. Tentative panel schedule.

The panel will be comprised of four parts. In the first part, the moderator and the panelist will provide a brief introduction to the panel topic. Each panelist will be given the opportunity to share their unique angle on digital trace data research. This will be followed by a moderated discussion among the panelists. Third, the panelists will facilitate round-table discussions with the audience to engage with and gain their perspective. After a summary of the round table discussions, we will together reflect on the panel and the key learnings from the discussions.

4 Panelists

We carefully selected panelists to represent various perspectives on digital trace data research. While some of the panelists are designated experts for specific computational techniques, such as machine learning or text mining, the panel also comprises scholars with experience in the area of computational theory development. All panelists confirm that we will attend the conference and serve on the panel should the panel be accepted. Below we provide the short bio for each panelist.

Bastian Wurm is a post-doctoral researcher and research group leader at the Institute for Digital Management and New Media at LMU Munich School of Management. His group investigates various topics that relate to Process & Algorithmic Management. Before joining LMU Munich in 2022, Bastian

worked as a research and teaching associate at the Vienna University of Economics and Business (WU Vienna). Bastian's work is published in journals such as *Information Sciences*, the *Journal of Information Technology*, the *Journal of Strategic Information Systems,* and *Communications of the Association for Information Systems*. His dissertation entitled "Organizational Complexity: Insights from Digital Trace Data Research" was awarded the Stephan Koren-Award for outstanding dissertations by WU Vienna.

Michael Wessel is an Associate Professor at the Department of Digitalization, Copenhagen Business School. He holds a PhD in Information Systems from Technical University of Darmstadt, Germany. His research resides at the intersection of information systems, digital entrepreneurship, and digital innovation. Methodologically, he prefers to work empirically and blends traditional experimental and quantitative research methods with computational approaches such as machine learning and text mining based on digital trace data. His work has been published in leading IS and entrepreneurship outlets such as *Journal of Management Information Systems, Journal of Information Technology, Information Systems Journal, and Decision Support Systems as well as Journal of Business Venturing and Entrepreneurship Theory and Practice.*

Monica Chiarini Tremblay is the Dorman Family Professor of Business at the Raymond A. Mason School of Business, William and Mary. Her research focuses on business analytics, particularly in healthcare, and design science research. She is currently working on several projects examining the role of digital technologies in delivering social justice and methods for transparent AI. Her publications appear in *MIS Quarterly*, *Journal of the AIS*, *Journal of American Medical Informatics*, *Decision Sciences*, *Decision Support Systems*, *European Journal of Information Systems*, *ACM Journal of Data and Information Quality*, and *Communications of the AIS*. She has been the principal investigator on several federal, state, and private grants in Health Information Technology. She was a study session member for the Health Information Technology section of Agency Healthcare Research Quality (National Institute of Health).

Michel Avital is Professor of Digitalization at Copenhagen Business School. Michel is an advocate of openness and an avid proponent of cross-boundaries exchange and collaboration. His research focuses on the relationships between digital innovation ecosystems and organizational practices. He studies how emergent technologies are developed, applied, managed and used to transform and shape organizations. He has published more than 100 articles on topics such as blockchain technology, the future of work, sharing economy, open data, open design, generative design, creativity, and innovation. He is a senior editor and editorial board member of leading IS journals and serves in various organizing capacities at major international conferences on digital technology and organization studies. Michel is a recipient of the AIS Fellow Award. Further information: http://avital.net

Philipp Hukal is an Assistant Professor at the Department of Digitalization at Copenhagen Business School. His research examines digitally-enabled innovation within and across organizations covering topics such as digital platforms, open source software development, and digital entrepreneurship. He holds a PhD in Information Systems and Management from Warwick Business School, as well as an MSc in Management, Information Systems, and Innovation from the London School of Economics and Political Science. Prior to this, Philipp has worked in analyst roles in the tech sector.

Iris Junglas is the Noah T. Leask Distinguished Professor of Information Management and Innovation in the Supply Chain and Information Management Department at the College of Charleston. She holds a Ph.D. from the University of Georgia, as well as a Bachelor's and Master's degree in Computer Science from the University of Koblenz, Germany. She has published more than 50 refereed journal articles in the field of Information Systems, including outlets, such as the *European Journal of Information Systems, Journal of the Association of Information Systems, Information Systems Journal, Journal of Strategic Information Systems, Management Information Systems Quarterly* and *Management Information Systems Quarterly Executive*. She is a Senior Editor for the *European Journal of Information Systems* and the Editor-in-Chief for *Management Information Systems Quarterly Executive*.

References

Agarwal, R. and V. Dhar. (2014). "Big data, data science, and analytics: The opportunity and challenge for IS research." *Information Systems Research*, *25*(3), 443–448.

Alsudais, A. (2021). "Incorrect data in the widely used Inside Airbnb dataset." *Decision Support Systems*, *141*(September 2020), 113453.

Avital, M., Jensen. T.B., and Dyrby, S. (2023) "The Social Fabric Framework: Steps to Eliciting the Social Making of Organizations in the Digital Age." *European Journal of Information Systems*, *32*(2), 127-153.

Baiyere, A., N. Berente and M. Avital. (2023). "On digital theorizing, clickbait research, and the cumulative tradition." *Journal of Information Technology*, *38*(1), 67–73.

Godoy-Descazeaux, I., Avital, M., and Gleasure, R. (2023). "Images of Quantum Computing: Taking Stock and Moving Forward." *Proceedings of the 31st European Conference on Information Systems (ECIS 2023)*, Kristiansand, Norway.

Grisold, T., W. Kremser, J. Mendling, J. Recker, J. vom Brocke and B. Wurm. (2022). "Keeping pace with the digital age: Envisioning information systems research as a platform." *Journal of Information Technology*, *38*(1), 60–66.

Grisold, T., B. Wurm, J. Mendling and J. vom Brocke. (2020). "Using Process Mining to Support Theorizing About Change in Organizations." In: *53rd Hawaiian International Conference on System Sciences (HICSS 2020)*.

Grover, V., A. Lindberg, I. Benbasat, K. Lyytinen. (2020). "The perils and promises of big data research in information systems." *Journal of the Association for Information Systems*, *21*(2), 268–291.

Howison, J., A. Wiggins and K. Crowston. (2011). "Validity Issues in the Use of Social Network Analysis with Digital Trace Data." *Journal of the Association for Information Systems*, *12*(12), 767–797.

Lazer, D. M. J., A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, … C. Wagner. (2020). "Computational social science: Obstacles and opportunities." *Science*, *369*(6507), 1060–1062.

Lin, M., H. C. Lucas and G. Shmueli. (2013). "Too big to fail: Large samples and the p-value problem." *Information Systems Research*, *24*(4), 906–917.

Lindberg, A. (2020). "Developing Theory through integrating Human & Machine Pattern Recognition." *Journal of the Association for Information Systems*, *21*(1), 90–116.

Lindberg, A., N. Berente, J. Gaskin and K. Lyytinen. (2016). "Coordinating Interdependencies in Online Communities: A Study of an Open Source Software Project." *Information Systems Research*, *27*(4), 751–772.

Miranda, S., N. Berente, S. Seidel, H. Safadi and A. Burton-Jones. (2022). "Computationally Intensive Theory Construction: A Primer for Authors and Reviewers." *MIS Quarterly*, *46*(2), i–xvi.

Pentland, B. T., E. Vaast and R. Wolf. (2021). "Theorizing process dynamics with directed graphs: A diachronic analysis of digital trace data." *MIS Quarterly*, *45*(2), 967–984.

Qiao, M. and K.-W. Huang. (2021). "Correcting misclassification bias in regression models with variables generated via data mining." *Information Systems Research*, *32*(2), 462–480.

Schirrmacher, N.B., Jensen, J. and Avital, M. (2021) "Token-Centric Work Practices in Fluid Organizations: The Cases of Yearn and MakerDAO." *Proceedings of the 42 International Conference on Information Systems (ICIS 2021)*, Austin, TX, USA.

Vaast, E., H. Safadi, L. Lapointe and B. Negoita. (2017). "Social media affordances for connective action - an examination of microblogging use during the gulf of mexico oil spill." *MIS Quarterly*, *41*(4), 1179–1205.

van Dongen, B. F. (2017). "BPI Challange 2017."

Xu, H., N. Zhang and L. Zhou. (2020). "Validity Concerns in Research Using Organic Data." *Journal of Management*, *46*(7), 1257–1274.

Zammuto, R. F., T. L. Griffith, A. Majchrzak, S. Faraj and D. J. Dougherty. (2007). "Information Technology and the Changing Fabric of Organization." *Organization Science*, *18*(5), 749–762.