Towards a Desirable Data Collection Tool for Studying Long-Term PIM

Jesse David Dinneen
School of Information Studies
McGill University
Montreal, QC H2T 2L3, Canada
jesse.dinneen@mail.mcgill.ca

Fabian Odoni
Swiss Institute for Information
Research
University of Applied Sciences
(HTW Chur)
Chur, CH-7004, Switzerland
fabian.odoni@htwchur.ch

Charles-Antoine Julien
School of Information Studies
McGill University
Montreal, QC H2T 2L3, Canada
charles.julien@mcgill.ca

Abstract

Long-term personal information management (PIM) scenarios entail unique challenges not only for those practicing PIM but also for those attempting to study it; overcoming these challenges will require very good data collection tools. In this position paper we note some important traits that such tools should have, thereby producing a kind of wishlist, and comment on the strengths and limitations of existing PIM methods that result from the absence of such tools. In accordance with my current research, most examples are drawn from studies of users' file management behaviour. We then briefly outline the tool we have built to treat several of these limitations.

Author Keywords

personal information management; file management; methodology; data collection

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous; See [http://acm.org/about/class/1998/]: for full list of ACM classifiers. This section is required.

Introduction

This is a position paper for the CHI '16 workshop 'For Richer, for Poorer, in Sickness or in Health... The Long-Term Management of Personal Information'. Our work is concerned with digital file management (FM): the not-well-understood activity performed by many computer users every day as they perform the broader task of personal information management (PIM). The workshop topic 'long-term management of personal information' obviously applies to FM, as there are challenges for performing and studying long-term FM; perspectives and insights on these difficulties likely also abstract to broader PIM contexts. Below, we note some observations made while reviewing PIM literature with regards to the pros and cons of various broad approaches and specific methods for studying PIM, especially as they apply across time and as a result of limited data collection tools. The result is a 'wishlist' of desirable traits for data collection tools that would facilitate studying both short- and long-term PIM. It is certainly not exhaustive, and we look forward to discussing it further at the workshop.

Desirable tool traits

Practical to recruit and administer

Manual methods of collecting data, such as the 'guided tour' of the desktop [2] and reviewing video recordings of users' behaviour [8] are time consuming to carry out, limiting the amount of data that can be collected. They also prohibit participation because they require that participants be willing to let another person see deeply personal digital contents and organization (or lack thereof) [2]. Perhaps as a result, nearly two-thirds (19) of the FM studies to date (31) feature fewer than 50 participants, while others have sampled from a single corporation [1] or university project [18]. Extending these methods across time is essential to studying long-term PIM, but will increase the amount of time required to implement them and give potential participants more reason to become missed participants. Desirable tools, therefore, should partially automate the collection process without losing the capacity to record qualitative data, encourage participation by respecting privacy, and encourage ongoing participation by offering some value to participants.

Collects data accurately and consistently

Two common approaches to studying PIM face challenges within single experiments and across time. First, one may ask participants for their opinions and reflections, e.g., about the challenge of coordinating files across multiple devices [12] or their preference for searching for files [25, 4], reveals users' experience and perceptions of PIM activities. Sometimes, however, participants' reflections do not accurately represent their PIM behaviour [5], and this effect may be exaggerated across time. Second, one may make observations of participant's PIM behaviour, e.g., directly, during some prompted behavior like file retrieval [6], or after the fact, by using software to examine the file space they have created [14]. With this approach, however, it is rather unclear which observations to make, and perhaps as a result most studies have report very few measures – typically fewer than ten, and never more than seventeen [14]. These measures are also typically inconsistent across studies, e.g., one reports maximum folder depth [20] while another reports average file depth [13]. By collating all the measures used and inferring other possible ones, we count over 50 measures that can be taken of FM storage, structure, and semantics; standardization and prioritization of these measures is obviously needed, but this is evidently not possible with the current tools.

Portable across technology

Digital PIM happens across platforms, frameworks, computers, spatial locations, occupational contetxs, and so on. For example, it has been suggested that operating system has an effect on FM [2, 7], but no single tool runs on all three major operating systems. It is understandable that tools are designed for only one platform, but this doesn't maximize re-use and makes it harder to study the role of technological differences in PIM. This will become harder still as the number of platforms that users can interact with grows; desirable tools, therefore, should be able to be implemented uniformly in various places while collecting data that can be abstracted across platform differences.

Persistent across time
As single-shot experiments take place during one particular time, they necessarily cannot capture changes in behaviour across time. Longitudinal studies ask a lot of the researchers and participants, however, and perhaps as a result, only three FM studies [9, 23, 13] have collected data from a participant more than once. Broader PIM studies have also made longitudinal observations, for example of how people return to Web-based information [22, 10] or how project collections change over time [11]; observing the management, refinding, and reuse of such information once downloaded from the Web and beyond the lifespan of a project will require data collection tools that integrate longitudinal FM and Web-based data. Logging software is one such option for collecting data across time [21, 13], but requires careful implementation and packaging as persistent logging software may be a hard sell to potential participants when comapred with discrete instances of data collection. It is essential, therefore, that future data collection tools support be either easily re-implementable or persistent but unimposing in their collection of longitudinal data.

Re-usable
Previous FM studies have feature a wide variety of contexts; for example, where one study examines the retrieval of recently used files seen during a controlled experiment [7], another examines the folder structures created by students in a proprietary, online environment during a class assignment [17]. It is perhaps for these reasons that we have never heard of any PIM study reusing the data collection tools of another. But doing so, if possible, would facilitate implementing new studies, and therefore, desirable data collection tools should be flexible enough to be used and re-used in varied contexts while collecting similar data across studies and time. Ideally, these tools and any source code would also be shared with the research community.

Can be used to study individual differences
Individual differences play a role in PIM behaviour [16] and will likely be necessary to understanding long-term PIM. As the tools used to study this have ranged from computerized tasks [26] to automated scripts [24] and even f-MRI [3], the need for a multitude of tools is clear. However, many psychological constructs can be measured through reliable and easily distributed instruments like questionnaires [15, 19]. It would be desirable, therefore, for any tool that collects PIM data to facilitate the close integration or collection of individual difference data, for example by offering a modular method for implementing and interchanging questionnaires and prompts for demographic attributes. Ideally, it would also have provisions for tracking these changes over time.

Towards a better tool

We have recently built a program specifically to exhibit many of the traits outlined above. Greater details of our software and its first use in a pilot study will appear soon in a recently submitted publication, but we are happy to discuss and demonstrate it at the workshop. We also encourage the workshop participants to try the software for themselves by participating in our study of file management behaviour[1]. In short, our tool:

- Allows anonymous, remote, and asynchronous data

[1] http://dinneen.research.mcgill.ca

collection of users' FM behaviour along many measures of the file-system

- Supports including different instruments (e.g., questionnaires) within the user interface
- Automatically detects technological variables, has space for inputting demographic attributes
- Provides value to participants by telling them about their collection
- Collects data that enable comparisons across studies and facilitate identifying a standardized set of measures for future studies
- Does not use persistent logging but can collect longitudinal data via multiple collection instances
- Runs and looks native in Windows, Mac OS X, and GNU/Linux
- Is open-source and meant to be re-used[2]

Conclusion

Performing and researching FM and PIM in long-term contexts is a substantial challenge. The existing FM tools have limited the approaches taken to studying FM and the results that have followed. Though one cannot expect a single data collection tool to fit every approach or avoid every limitation, improved tools are essential to collecting data that will, in turn, enable the development of nuanced models, frameworks, and theories of PIM. The next generation of research tools therefore need to be relatively easy to implement, administer, and re-use, accurate and consistent, portable and persistent, reusable, and modular enough to support investigating determining factors like individual differences.

[2] https://github.com/jddinneen/cardinal

References

[1] Nitin Agrawal, William J Bolosky, John R Douceur, and Jacob R Lorch. 2007. A five-year study of file-system metadata. *ACM Transactions on Storage (TOS)* 3, 3 (2007), 9.

[2] Deborah Barreau. 1995. Context as a factor in personal information management systems. *Journal of the American Society for Information Science* 46, 5 (1995), 327–339.

[3] Yael Benn, Ofer Bergman, Liv Glazer, Paris Arent, Iain D Wilkinson, Rosemary Varley, and Steve Whittaker. 2015. Navigating through digital folders uses the same brain structures as real world navigation. *Scientific Reports* 5 (2015), 14719.

[4] Ofer Bergman, Ruth Beyth-Marom, Rafi Nachmias, Noa Gradovitch, and Steve Whittaker. 2008. Improved search engines and navigation preference in personal information management. *ACM Transactions on Information Systems (TOIS)* 26, 4 (2008), 20.

[5] Ofer Bergman, Noa Gradovitch, Judit Bar-Ilan, and Ruth. Beyth-Marom. 2013. Tagging Personal Information: A Contrast between Attitudes and Behavior. In *ASIST 2013*. Montreal, Quebec, Canada.

[6] Ofer Bergman, Steve Whittaker, and Noa Falk. 2014. Shared files: The retrieval perspective. *Journal of the Association for Information Science and Technology* 65, 10 (2014), 1949–1963.

[7] Ofer Bergman, Steve Whittaker, Mark Sanderson, Rafi Nachmias, and Anand Ramamoorthy. 2010. The effect of folder structure on personal file navigation. *Journal of the American Society for Information Science and Technology* 61, 12 (2010), 2426–2441.

[8] Ofer Bergman, Steve Whittaker, Mark Sanderson, Rafi Nachmias, and Anand Ramamoorthy. 2012. How do we find personal files?: the effect of os, presentation & depth on file navigation. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. ACM, 2977–2980.

[9] Richard Boardman and M Angela Sasse. 2004. Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 583–590.

[10] Harry Bruce, William Jones, and Susan Dumais. 2004. Information behaviour that keeps found things found. *Information Research* 10, 1 (2004), paper 207.

[11] Harry Bruce, Abe Wenning, Elisabeth Jones, Julia Vinson, and William Jones. 2011. Seeking an ideal solution to the management of personal information collections. *Information Research* 16, 1 (2011), paper 462.

[12] Robert Capra. 2009. A survey of personal information management practices. In *PIM workshop at ASIS&T 2009*. Vancouver, British Columbia, Canada.

[13] Stephen Fitchett and Andy Cockburn. 2015. An empirical characterisation of file retrieval. *International Journal of Human-Computer Studies* (2015).

[14] Daniel J Gonçalves and Joaquim A Jorge. 2003. An empirical study of personal document spaces. In *Interactive Systems. Design, Specification, and Verification*. Springer, 46–60.

[15] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[16] Jacek Gwizdka and Mark Chignell. 2007. *Individual Differences*. University of Washington Press, 206–220.

[17] Sharon Hardof-Jaffe, Arnon Hershkovitz, Hama Abu-Kishk, Ofer Bergman, and Rafi Nachmias. 2009a. How Do Students Organize Personal Information Spaces?. *International Working Group on Educational Data Mining* (2009).

[18] Sharon Hardof-Jaffe, Arnon Hershkovitz, Hama Abu-Kishk, Ofer Bergman, and Rafi Nachmias. 2009b. Students' organization strategies of personal information space. *Journal of Digital Information* 10, 5 (2009).

[19] Mary Hegarty, Anthony E Richardson, Daniel R Montello, Kristin Lovelace, and Ilavanil Subbiah. 2002. Development of a self-report measure of environmental spatial ability. *Intelligence* 30, 5 (2002), 425–447.

[20] Sarah Henderson and Ananth Srinivasan. 2011. Filing, piling & structuring: strategies for personal document management. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. IEEE, 1–10.

[21] Carlos Jensen, Heather Lonsdale, Eleanor Wynn, Jill Cao, Michael Slater, and Thomas G Dietterich. 2010. The life and times of files and information: a study of desktop provenance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 767–776.

[22] William Jones, Harry Bruce, and Susan Dumais. 2003. How do people get back to information on the web? How can they do it better?. In *INTERACT*.

[23] William Jones, Abe Wenning, and Harry Bruce. 2014. How Do People Re-find Files, Emails and Web Pages? *iConference 2014 Proceedings* (2014).

[24] Charlotte Massey, Sean TenBrook, Chaconne Tatum, and Steve Whittaker. 2014. PIM and personality: what do our personal file systems say about us?. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 3695–3704.

[25] Jaime Teevan, Christine Alvarado, Mark S Ackerman, and David R Karger. 2004. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 415–422.

[26] Kim J Vicente, Brian C Hayes, and Robert C Williges. 1987. Assaying and isolating individual differences in searching a hierarchical file system. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 29, 3 (1987), 349–359.