# Sustainability of Bits, not just Atoms

**Manas Tungare**
Google, Inc.
manas@tungare.name

**Manuel A. Pérez-Quiñones**
Dept. of Computer Science, Virginia Tech.
perez@cs.vt.edu

**Pardha S. Pyla**
Bloomberg Finance, L.P.
pardha@pyla.name

**Ben Hanrahan**
Dept. of Computer Science, Virginia Tech.
bhanraha@cs.vt.edu

**Uma Murthy**
Dept. of Computer Science, Virginia Tech.
umurthy@cs.vt.edu

**Ricardo Quintana-Castillo**
Dept. of Computer Science, Virginia Tech.
rqc@cs.vt.edu


Submitted to CHI 2010 Workshop: Examining Appropriation, Re-use, and Maintenance for Sustainability.

## Abstract

In this paper, we discuss sustainability as it applies to digital artifacts and personal information. We continually create and/or receive new information items in the form of emails, files, photos, media, etc., but once these artifacts enter our information ecosystems, they stay permanently and are rarely deleted even if their intrinsic value is no longer the same as earlier. This impacts information seeking tasks negatively, as users must now learn to navigate a larger corpus of information, and leads to information overload. We describe the technological causes of information overload in the context of existing finding, filing, and refiling practices and information heirlooms. We conclude with an example of a solution that can address this challenge.

## Sustainability in Computing

Most current discussions around sustainability in computing center either around the environmental impact of computing artifacts and their reuse [4] or around interaction design that encourages sustainability in their use [6]. Another discussion on sustainability focuses on the ability of future generations to understand, interpret, and make sense out of the digital artifacts created in the past. This includes preserving and maintaining access to digital artifacts, including the related issues of proprietary formats, evolution of hardware, encoding standards, and transformation of content into more accessible formats. [5]. While these may be the dominant concerns of sus-

tainability in computing, they are by no means the only ones.

## Information Sustainability

In this paper, we eschew these oft-discussed themes, and instead take on the subtler issue of sustainability of the information management practices of users. We argue that this is crucial, given the alarming rate at which information is created, transmitted, consumed, and archived today. Computing in general, and personal computing in particular, has made it extremely easy and affordable to create and store large amounts of data. Because of the availability of low-cost, always-available storage solutions (including local disks and remote cloud storage), the need for curating one's information (represented in their data) has been declining over the last decade. Users choose to save everything, archive everything, and have little to no incentive for pruning their information collections.

However, while storage systems have scaled up to meet the space demands of increasingly larger information collections, our attentional resources and information-seeking tools and strategies have not seen a corresponding increase in capabilities, leading to information overload [7].

In this paper, we discuss several related problems in this domain, and analyze a few issues framed within terminology from the Personal Information Management literature. While we discuss a few solutions towards the end, this paper is not particularly about proposing solutions—it is intended to initiate a discussion about them.

### Analog Information and Physical Constraints

Prior to the advent of digital technologies, the constraints on information archives were predominantly physical, since only a limited number of files could be held in a storage location. Since space was a scarce resource, archivists coped with the problem by implementing appropriate retention policies. Historical transactional data were expected to have limited lifetimes. Personal collections of memorabilia were limited by the number of shoeboxes that could fit in an attic.

### The Challenges Of Managing Personal Information

While the power of computers made it possible for us to structure, analyze, and consume much of business data, often along several dimensions, that is unfortunately not the case with personal information. Personal information, because it is situated in the personal context of use, poses unique challenges. All the processing power made available by today's powerful computers cannot effectively be used to perform such tasks as *'locating that email with the contact information of the person you met at the meeting two weeks ago'* or *'the set of photos of your daughter taken around sunset at the picnic last July'* unless the original data items (contacts, or emails, or photos) were appropriately filed or tagged at the time they were created/received. The end result is that our archives increase in size over time, while our tools and strategies have yet to adapt to the current deluge of information.

## Information Landfills

The cost of magnetic disk storage space has dropped by about forty-five percent *per year* since 1989 [9] while our attentional resources have remained constant. Thus today, human attention is a more costly resource than storage space. A consequence of this is that more and more users tend to skip filing or managing their archives, and instead rely on automatic or manual archiving to move their information from the foreground to the background. Doing so makes it 'go away' from their primary view in the

short term, which is good, but does not solve the ultimate problem—that of retrieval after a longer interval.

Many online storage providers, including email providers, now promise seemingly infinite inbox capacity. Some of them >explicitly encourage users not to delete their email, but instead archive it. Their policy is clearly, 'never delete anything'. For users who still attempt to file their information, the time required to manually create and maintain a filing scheme increases at least linearly with the number of messages needed to be filed.

Even when information is archived to make it go away from the active view, it still lingers, somewhere in an information landfill, and ends up polluting the information-seeking tasks performed by the user. Keyword searches are less effective because many more documents (emails, files, bookmarks, encountered information) match the user's query terms, and the onus of performing triage over these search results is ultimately upon the users.

## The Implications of Life-Logging

While in the previous section we talked solely about information that we already received, the availability of affordable storage solutions coupled with low-cost sensing technologies has spurred the development of life-logging. Projects such as SenseCam [12] and MyLifeBits [8] encourage users to wear multi-sensing devices that capture a life-log while a person goes about their everyday activities. Information collected from SenseCam has been used to study whether it supports remembering events from the past [13], but that seems like a self-fulfilling prophecy. It would be interesting to note if this data assists other tasks that users currently perform, including answering questions related to people they've met or events they have attended.

## The Marginal Utility of the Long Tail of Information

At some point, users need to make a tradeoff between keeping every last bit of information (and thus requiring a lot of time locating specific items when required) and keeping fewer information items, pruning their collections regularly, maintaining the ability to locate items faster, while being able to accept that a few items of future interest might potentially be deleted prematurely. As we keep more and more information, the long tail of it—information that is not accessed often, but still of non-zero interest—grows quickly. The question to ponder is whether the utility of this information outweighs the attentional costs required to maintain and preserve it.

## Information Heirlooms

By 2010, there has been at least one generation that interacted extensively with computing technologies. When, sadly, members of this generation pass away, their inheritance is no longer composed just of physical objects, but also digital artifacts. Photographs, documents, notable events, have all been captured digitally and passed along as heirlooms to their descendants. How can the decision be made about what objects to keep, and what to discard? Clearly, some objects are much more valuable than others, and the original creator of the information is no longer available to help make that decision.

At the other end of the spectrum, parents are capturing more and more data about their newborns [11], including such trivialities as number of diapers changed and number of minutes slept each night. Clearly, this is data that needs to be managed, pruned, and deleted after a while, lest it end up in an information landfill that outlasts the baby in question.

This issue of maintaining access to information as a souvenir extends to group memories as well: notable events—e.g., the Sep 11 attacks of 2001, the Virginia Tech shooting of April 16, 2007—generated large scale memorials in the form of photos, videos, emails, phone calls, text messages, web pages, blog posts, and written documents[1]. How can these memorabilia be preserved and curated for the future?

## Analysis from a PIM point of view

### Ephemeral, Working, and Archived Information

Barreau [2] described the distinction between ephemeral, working, and archived information early in the history of PIM. The pattern we see now does not deviate much from the one observed in 1995, except that the proportion of archived information (as a part of the entire corpus) has grown considerably since then. However, tools have not evolved to permit easy movement of a piece of information through these types.

A distinction can be made between data archived intentionally by users, versus data that was automatically archived. Perhaps systems can evolve to recognize this distinction, and prioritize explicitly-archived data over background-archived data.

### Sidestepping the Keeping Problem

Post-Valued Recall [16] refers to the interest a user may have in recalling information whose value is not recognized until some time after its initial retrieval. The Keeping Decision [10] in personal information management refers to the choice a user must make about her information items with incomplete knowledge about its value in the future. The non-choice that users make to keep everything they ever encountered effectively sidesteps the keeping problem.

But, clearly, as vast amounts of information of questionable importance continue to pollute users' personal information collections, the chances of finding the proverbial needle in the larger haystack diminish. Solutions such as better search capabilities attempt to reduce the time it might require to access older information, search is by no means the only technique users use to locate information [14]. Even if a theoretically perfect search engine were available, a significant factor in the success or failure of a search task is how well the query was formulated [1]. If the same query were issued to two corpuses, one with X items and the other with 100X items, it is more likely that more search results would be obtained from the larger corpus than from the smaller one. Thus, the result evaluation sub-task of the search process is expected to require a larger amount of cognitive effort for a larger corpus. In signal detection task terminology applied to the keeping problem [10], the likelihood of a false positive is higher in case of a larger corpus than it is with a smaller corpus.

## Potential Solutions

This paper is not about solutions, but we discuss one particular solution to initiate the discussion.

For information that is archived automatically (without human interaction), it may be useful to have certain parts of it also expire automatically (i.e. discontinue to be available). Expiration need not equate instantaneous deletion; the basic idea is to move it out of the foreground and out of all information seeking activities performed over that corpus. Email, for example, can be annotated for expiration [15], or can be heuristically expired based on its

[1] http://scholar.lib.vt.edu/prevail/

content (e.g. a user could set a custom rule such that emails with the words 'lunch plans?' and fewer than a hundred bytes in size would be automatically set to expire within 12 hours.) An inverse exponentially decreasing priority value for information is potentially one more automatic strategy for continual expiration without user involvement.

Demoting older content while browsing or searching is a similar strategy that can assist in prioritizing recent content over older, automatically archived content [3]. Thus, while this content is still available, users must make a conscious decision to include archived content in their searches. Archives can also be stored offline to ensure better use of local storage.

## References


[1] A. Aula. Query formulation in web information search. In P. Isaias and N. Karmakar, editors, *Proceedings of IADIS International Conference WWW/Internet 2003*, pages 403–410. IADIS Press, 2003.

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

[3] O. Bergman, R. Nachmias, and R. Beyth-Marom. The use of subjective attributes in personal information management systems - initial results. In A. S. for Information Science and Technology, editors, *Proceedings of the American Society for Information Science and Technology*, volume 40, pages 509–510, School of Education, Tel-Aviv University, Tel-Aviv 69978, Israel; Department of Education and Psychology, The Open University of Israel, P.O. Box 39328, Tel-Aviv, 61392, Israel, 2003.

[4] E. Blevis. Sustainable interaction design: invention & disposal, renewal & reuse. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 503–512, New York, NY, USA, 2007. ACM.

[5] K. Bradley. Defining digital sustainability. *Library Trends*, 56(1):148–163, 2007.

[6] D. K. Busse, E. Blevis, C. Howard, B. Dalal, D. Fore, and L. Lee. Designing for a sustainable future. In *C&C '09: Proceeding of the seventh ACM conference on Creativity and cognition*, pages 493–494, New York, NY, USA, 2009. ACM.

[7] A. F. Farhoomand and D. H. Drury. Managerial information overload. *Communications of the Association for Computing Machinery (CACM)*, 45(10):127–131, 2002.

[8] J. Gemmell, G. Bell, R. Lueder, S. Drucker, and C. Wong. MyLifeBits: fulfilling the Memex vision. In *MULTIMEDIA '02: Proceedings of the Tenth ACM International Conference on Multimedia*, pages 235–238, New York, NY, USA, 2002. ACM.

[9] S. Gilheany. Archive builders' analysis newsletter for document management. *Archive Planning*, 5(5), 2001.

[10] W. P. Jones. Finders, keepers? the present and future perfect in support of personal information management. *First Monday*, 9(3), 2004.

[11] H. Leggett. Baby-by-number: Parents' new obsession with data. Wired Magazine,

http://www.wired.com/wiredscience/2009/12/baby-tracking/, Last Accessed: December 2009.

[12] Microsoft. Sensecam: Microsoft research. http://research.microsoft.com/sendev/projects/sensecam/, Last Accessed: October 2007.

[13] A. J. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother, and K. Wood. Do life-logging technologies support memory for the past?: An experimental study using SenseCam. In *CHI '07: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 81–90, New York, NY, USA, 2007. ACM Press.

[14] J. Teevan, C. Alvarado, M. S. Ackerman, and D. Karger. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 415–422, New York, NY, USA, 2004. ACM Press.

[15] M. Tungare and M. Pérez-Quiñones. "best if used by": Expiration dates for email. In *Proceedings of the CHI 2009 Workshop on Interacting with Temporal Data*, 2009.

[16] J. Wen. Post-valued recall web pages: User disorientation hits the big time. *IT & Society*, 1(3):184–194, 2003.