Towards Concise Preservation by Managed Forgetting: Research Issues and Case Study

Nattiya Kanhabua
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
kanhabua@L3S.de

Claudia Niederée
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
niederee@L3S.de

Wolf Siberski
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
siberski@L3S.de

ABSTRACT

In human memory, forgetting plays a crucial role for focusing on important things and neglecting irrelevant details. In digital memories, the idea of systematic forgetting has found little attention, so far. At first glance, forgetting seems to contradict the purpose of archival and preservation. However, we are currently facing a tremendous growth in volumes of digital content. Thus, it becomes ever more important to focus, while forgetting irrelevant details, redundancies and noise. This holds true for better organizing the information space as well as in preservation management for making and revisiting decisions on what to keep. Therefore, we propose the introduction of the concept of *managed forgetting* as part of a joint information management and preservation management process in digital memories. Managed forgetting models resource selection as a function of attention and significance dynamics. Based on dynamic, multidimensional information value assessment it identifies information objects, e.g., documents or images of decreasing importance and/or topicality and triggers *forgetting actions*. Those actions include a variety of options, namely, aggregation and summarization, revised search and ranking behavior, elimination of redundancy, and finally, also deletion. In this paper, we present our vision for managed forgetting, discuss the challenges as well as our first ideas for its introduction, and present a case study for its motivation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information filtering

General Terms

Human Factors, Measurement

Keywords

Digital Preservation; Dynamic Information Value Assessment; Time-aware Information Access; Managed Forgetting

1. INTRODUCTION

While preservation of digital content is now well established in memory institutions, such as, national libraries and archives, it is still in its infancy in most other organizations, and even more so for personal content. This is unsatisfying for two reasons: 1) with the growing volumes of and reliance on digital content there is a clear need for better long-term storage solutions in the organizational and in the personal context than the currently used backup strategies and 2) advanced and mature preservation technology is meanwhile available, also due to the intensive research and development work in this area in the recent years. For example, a variety of preservation platforms have been developed, such as, the SCAPE platform [25], which focuses on scalability or the platform developed in the PROTAGE project [11], which relies on a smart multi-agent architecture.

There are several obstacles for the wider adoption of preservation technology in organizational and personal information management: There is a considerable gap between active information use and preservation activities. Active information use refers to dealing with information objects for everyday private or professional activities, typically supported by some information management environment, such as, a content management system in an organization or a desktop environment in the context of personal information management. In addition, especially in personal information management, there is typically little awareness for preservation. Although the need for personal preservation has been recognized in theory [12, 14], this did not propagate to more practical settings and solutions yet. This is further aggravated by the fact that no benefits are seen for moving from more or less systematic backup to systematic preservation.

For improving preservation support in organizations, there is considerable research work underway as for example in the project ENSURE[1]. Lately, this also includes work on the preservation of business workflows [15]. In practical settings, systematic backups have become part of daily routine within organizations, at least with respect to a short- to mid-term perspective. However, the readiness to invest into preservation is low, if not enforced by legal regulations. Finally, establishing effective preservation and concise and usable archives still requires a lot of manual work for selecting content that is relevant for preservation and for keeping the archives accessible and meaningful in the long run, thus entailing expenses much larger than just the storage costs.

In this paper, we propose the introduction of the novel concept of *managed forgetting* as part of a joint information and preservation management process, in order to overcome some of the above obstacles. This concept is inspired by the important role of forgetting in the human brain, where forgetting enables us to focus on the things that are relevant instead of drowning in details by remembering everything. The idea of managed forgetting is to systematically deal with information that progressively ceases in impor-

[1] `http://ensure-fp7-plone.fe.up.pt/site/`

tance and becomes redundant. At first glance, forgetting seems to contradict the idea of preservation, which is about keeping things, not about throwing them away. However, if no special actions are taken for long-term preservation, we already face a rather random digital forgetting process in the digital world today. This is triggered, e.g., by changing hardware, hard-disk crashes, or changes in employment. Furthermore, on a more global level there is a growing understanding that *forgetting* has to be considered as an alternative to the dominating keep it all paradigms, especially for information about individuals available in the Web [16].

We aim to replace such random forgetting processes with managed forgetting. In particular, we envision an idea of *gradual forgetting*, where complete digital forgetting is just the extreme and a wide range of different levels of condensation for preservation is foreseen. This concept is expected to both help in preservation decisions (also taking into account constraints for digital forgetting, e.g., legal regulations) and to create direct benefits for active information use by helping to keep the active information spaces more focused.

The rest of the paper is structured as follows: Section 2 describes the wider system context in which managed forgetting will be embedded in the ForgetIT project. Section 3 summarizes research challenges together with our first ideas for solving such challenges. Section 4 presents a case study in support of the motivation of managed forgetting. Finally, Section 5 concludes the paper with a description of the next steps towards realizing the concept of managed forgetting.

2. PROJECT AND SYSTEM CONTEXT

In our proposed approach, which will be implemented in the European project ForgetIT[2], the goal is to develop approaches and technologies for intelligent preservation management, which create a feasible and smooth path for preservation in the personal and organizational context and keeps the archived information concise, relevant and digestible by managed forgetting and contextualized remembering. For achieving its goal, the ForgetIT project will target: (a) enabling managed forgetting in information management and preservation management (cf. Section 3); (b) enabling contextualized remembering for keeping preserved content meaningful, useful and digestible through evolution-aware contextualization even when terminology, interpretation or context of use have changed considerably, and (c) closing the gap between information management and preservation management by introducing an approach for *synergetic preservation*. To validate the approach two application pilots will be built on top of the framework, one for preservation in the personal information management and the other in the organizational preservation management context.

2.1 Joint Model for Synergetic Preservation

For embedding the managed forgetting process we aim for an improved coupling between the information management system and the archival information system (AIS). This requires work on the conceptual level (preservation reference model extensions) as well as on the architectural level for system coupling. For institutional preservation organizations, reference models, such as, OAIS provide a solid foundation for the design and customization of preservation processes. Starting with ingest, OAIS describes very well how

[2]http://www.ForgetIT-project.eu

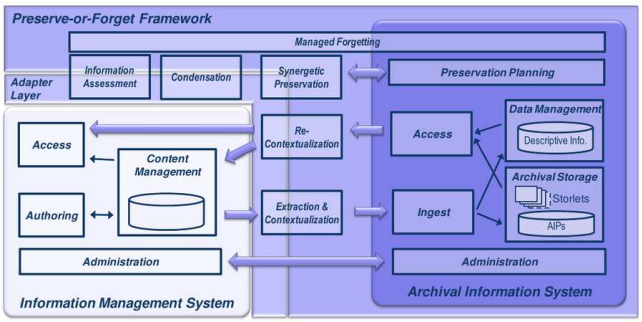

Figure 1: ForgetIT Architectural Approach.

content is transformed into self-contained archival packages and managed in the archival system. The part of the resource lifecycle which lies before ingest is, however, not included, although there is also work in the context of OAIS focusing on the pre-ingest phase such as the Producer-Archive Interface Methodology Abstract Standard (PAIMAS). Typically, this part of the resource cycle is described by information management workflows, covering tasks, roles, and resource states during the production process. To enable a tight connection between information management and preservation, these process models need to be coupled, to enable a seamless transfer of resources and their context information as well as to enable managed forgetting to be seamlessly applied.

It is planned to use OAIS as a starting point, and - taking into account other preservation process models as well (e.g., [23]) - and to develop a conceptual extension that covers the whole resource lifecycle. This reference model will treat issues such as when to create SIPs from resources of the active system for ingest by preservation storage, which context information from information management to preserve, or how to distribute responsibility for preservation tasks to information management roles.

2.2 Integration Architecture

In synergetic preservation, the roles of producer and consumer fall together. There may be other consumers, but one of the core consumers is the Information Management System. For the producer, preservation should be as transparent as possible; users which act as producers work in the active system and should not be forced to leave this environment for preserving their content. Consequently, the submission and access interfaces to the AIS should become part of the active system from the user's point of view. This poses an architectural challenge, because both information management as well as AIS come already with their own full-fledged software architectures. The aim is to achieve a tight integration without re-inventing a new integrated framework from scratch. The approach here is to use existing preservation architectures, and to realize the integration with an information system specific adaptation layer (see Figure 1). This layer connects system-specific content models, events, and processes to the corresponding generic preservation concepts implemented as part of what we call the Preserve-or-Forget framework, and the implementation of the managed forgetting process will be part of this framework.

A core factor for synergetic preservation is the smooth transition of content from information management storage to preservation storage. In addition, it is important to also support the reverse direction, i.e., to put the resources deliv-

ered by the preservation store back into active use. Depending on how far gone the information object is in its state of "inactivity", the object might be extracted into a format that is directly able to become ingested back into the active system, or to a format that is more platform independent and less likely to be directly ingestible in its original system. When re-activating a previously archived object, contextual links need to be re-created and/or updated to account for semantic shift (re-contextualization).

3. CHALLENGES

The introduction of managed forgetting into digital memories is a challenging task and its adequate combination with the goals of preservation has to be carefully investigated implying three key challenges:

- An interdisciplinary concept for flexible and gradual **managed forgetting** that meets **human expectations** and is driven by the goal of the digital memory complementing human memory;
- Development of flexible and multifaceted **information value assessment** methods in support of managed forgetting and in support of resource selection for preservation;
- Development of adequate **forgetting actions** especially for **quality-aware consolidation and concentration** for textual and multimedia content, such as, summarization, aggregation, detection of redundancy, and consideration of diversity.

3.1 Challenge: Meeting Human Expectations

Relevant State of the Art: In the field of psychology, aforementioned works [18, 27, 29] conducted subjective studies in order to shed light on understanding human remembering and forgetting. This can benefit digital preservation methods that aim at complementing the human ability to remember or forget information. From the Human-Computer Interaction (HCI) perspective, works related to digital preservation are, e.g., [4, 7, 8], which focus on system design for supporting the reminiscence of past events.

First Ideas in ForgetIT: Supporting managed forgetting in a digital memory is a novel concept, for which no former experience and best practices exist. It is therefore important to thoroughly analyze the human expectation for this process. An interdisciplinary approach is planned for this purpose. The idea is to investigate, what we can learn from the way a human memory forgets and remembers. Humans are, for example, very effective in (a) rapidly extracting the general gist of an experience, while forgetting many details, in (b) extracting common pattern of similar experiences avoiding the redundant "storage" of such pattern, and in (c) identifying data that are only temporally required and can be forgotten after task completion. Those and further characteristics of human forgetting will be further investigated. Selected characteristics will flow into a model for managed forgetting. The goal is, however, to complement not to copy or replace human memory. This perspective will create the highest benefit in the interaction of humans with digital memory. For analyzing the expectations towards managed forgetting user studies will be performed.

A further important source of inspiration for tailoring the managed forgetting process are the best practices and guidelines, which are already used in libraries and archives for selecting material for retention, transfer and destruction.

3.2 Challenge: Multifaceted Information Value Assessment

Relevant State of the Art: Forgetting basics [1, 9, 10, 22] are based on a decay theory, and an interference theory. There have been some works on modeling a temporal decay function, for example, applied to data streams [19] and exploited in information retrieval [13]. A recent work [20] considers different temporal document priors inspired by retention functions [17] considered in cognitive psychology that are used to model the decay of memory.

First Ideas in ForgetIT: Assessing the information objects in digital memory provides the basis for triggering managed forgetting actions, such as, condensation, contextualization and transition to the archive. We define two complementing information assessment values: *memory buoyancy* and *preservation value*. Memory buoyancy is inspired by the metaphor of information objects sinking down in the digital memory with decreasing importance, usage, etc., which increases their distance to the user. Memory buoyancy is influenced by a variety of factors in the following categories: usage parameters (such as, frequency and recency of use, user ratings, recurrent pattern), type and provenance parameters (information object type, source/creator) and context parameters (such as, relevance of resources as background information, general importance of topic, external constraints), and temporal parameters (age, lifetime specifications). The preservation value reflects the importance that the considered object gets preserved and will be used to decide if and when to archive an information object. Partly, the preservation value is influenced by similar factors as memory buoyancy, but it serves a different purpose: An object with a high value of memory buoyancy might already be moved to the archive (as a copy), because it has a very high preservation value, while staying still in direct uncondensed access to the user; an information object with low memory buoyancy and low preservation value might be preserved only in its condensed version or it might be decided not to preserve it at all. In this activity various factors influencing memory buoyancy and preservation value will be investigated as well as approaches for learning most effective factor combinations. Furthermore, approaches for enabling the user to explicitly and implicitly influence the values for memory buoyancy and preservation value will be developed, e.g., explicit expiry dates and lifetime specifications or tagging objects as non-forgettable.

3.3 Challenge: Flexible Forgetting Actions

Relevant State of the Art: Relevant research areas to forgetting actions for quality-aware consolidation include document summarization, duplicate detection, and diversity analysis. Automatic document summarization [26] is aimed at extracting the semantic content from a document in order to produce a well-formed and grammatical summary of what the document or document set is about and what its broad content is. Aforementioned works on detecting duplicate or near-duplicate documents has been mainly focused on different similarity metrics [5, 6, 28]. In the area of information retrieval, there is an interplay between redundancy, diversity and interdependent document relevance [3, 21].

First Ideas in ForgetIT: There are several forms of forgetting that will be supported including: changing the ranking of the "forgotten" object in a result list or not showing it as a result at all, replacing the object by a summary

object, marking the object as a deletion candidate etc. As an extreme the process will also support deletion as a forgetting option. Furthermore, managed forgetting will be used in several places of the information and preservation lifecycle: for focusing the content in active use, for helping in preservation decisions and for revisiting preservation decisions within the archive (gradual forgetting). Clearly there will be no one size fits all for managed forgetting, either. It is planned to define an adaptable framework for the managed forgetting process, which fixes the principle mechanisms of the process and can be customized along different dimensions: the parameters that are used for information assessment, the threshold used for memory buoyancy and preservation value for triggering forgetting actions and the options of forgetting considered. We will also investigate the use of a policy framework that supports the definition of different forgetting policies. Policies have been shown to be an intuitive and powerful tool in the area of security management, e.g., for specification of access rights. In the preservation context, besides customizing the forgetting process, policies also can capture external constraints, such as legal preservation requirements or business requirements (e.g., to make sure that information pertinent to obsolete product versions is preserved). Furthermore, we will also investigate into methods for detecting redundancies and for condensating textual as well as multimedia information objects.

4. DELETION BEHAVIOR IN ONLINE SOCIAL BOOKMARKING: A CASE STUDY

To support our motivation of systematic forgetting, we conducted a case study of analyzing deletion behavior in Online Social Bookmark and Publication Management System - BibSonomy [2]. The web-based system supports team-oriented publication management and social bookmark sharing. BibSonomy offers users an ability to categorize and archive two types of resources, i.e., *bookmarks* and *literature references*. In particular, a user can upload and share a resource, or label them with arbitrary words, so-called *tags*. In addition, an uploaded resource can also be deleted from the system by its owner when needed.

A formal model for BibSonomy is given as follows: U, T, and R are finite sets, whose elements are called users, tags and resources, respectively. Y is a ternary relation between them, i.e., $Y \subseteq U \times T \times R$, whose elements are called tag assignments, and the set P of all posts is defined as $P = \{(u, S, r) | u \in U, r \in R, S = T(u, r), S \neq \emptyset\}$ where, for all $u \in U$ and $r \in R$, $T(u, r) = \{t \in T | (u, t, r) \in Y\}$ denotes all tags the user u assigned to the resource r. The principal unit of our analysis is a post p, which is a transaction made when inserting a resource to the system. Based on the BibSonomy data model described in [2], there can be more than one transaction records associated to a resource uploaded. This is because a transaction record will be created for *each tag* assigned to the inserted resource. In this study, we do not leverage user tag information, and all transaction records belonging to the same resource ID will be regarded as one unit of study, or a *post* in our case. Thus, a post p is defined as a tuple $(u, r, time(r))$, where a user u is the owner of a resource r uploaded at *time(r)*.

In order to motivate the concept of managed forgetting, we investigate deletion processes manually performed by users over time, so-called *deletion behavior*. We obtained the publicly-available data dumps of BibSonomy consisting

Table 1: Statistics of distinct posts per user.

Type	Max	Avg	Std
All	119,678	370.87	2872.39
Bookmark	58,144	171.91	1292.09
Bibtex	119,678	198.96	2556.16

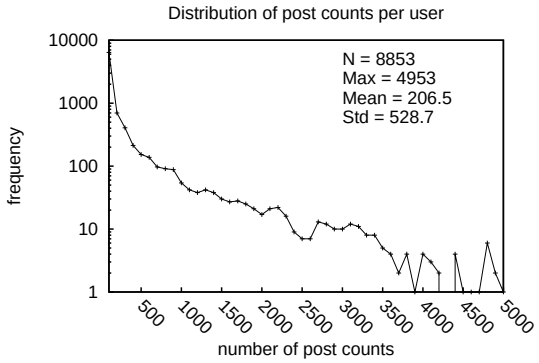

Figure 2: Distribution of post counts per user.

of 15 data snapshots, i.e., 2006-06-30, 2006-12-31,..., 2012-01-01, 2012-07-01, 2013-01-01, where the average time distance between any two snapshots is approximately 6 months. The dataset does not contain information about user names and demographics, thus our analysis was studied unobtrusively and compiled anonymised client based Web transaction logs. As of 1 January 2013, the number of active users in this study is 8,928 users, and basic statistics about distinct posts per user are shown in Table 1. The maximum numbers of bookmarks and bibtex posted per user are 58,144 and 119,678, respectively. On average, there is about 370 resources posted per user and the average of bookmarks and bibtex posted per user are 171 and 198, respectively.

As mentioned in [2], there are non-human users that automatically insert posts, e.g., the DBLP computer science bibliography. Therefore, we ignored such users with more than 5,000 posts from the analysis. To this end, we conducted the study in total of 8,853 users. Figure 2 shows the distribution of the number of distinct resources (post counts) per user. We conducted a detailed analysis by dividing users with respect to the number of their resources posted in total, into three groups: Group1 (10-100 posts), Group2 (101-1,000 posts), and Group3 ($>$ 1,000 posts). Our hypothesis is that different groups of users can shed light on the different characteristics of deletion behavior among users who share posts from very few to very many.

Deletion behavior was studied by computing the number of posts added or removed made by each user at different time snapshots. For a given user u, the number of posts *added* at a particular time snapshot t_i can be computed as the *difference of two sets*, namely, the set of posts at current time t_i and the set of posts at the previous time snapshot t_{i-1}: $add(u, t_i) = P(u, t_i) - P(u, t_{i-1})$, where the type of post $p \in P$ can be either a bookmark or a publication reference (denoted *bibtex*). On the contrary, the number of posts *removed* at a particular time snapshot t_i can be computed as the difference of the set of posts at the previous time snapshot t_{i-1} and the set of posts at current time t_i: $remove(u, t_i) = P(u, t_{i-1}) - P(u, t_i)$.

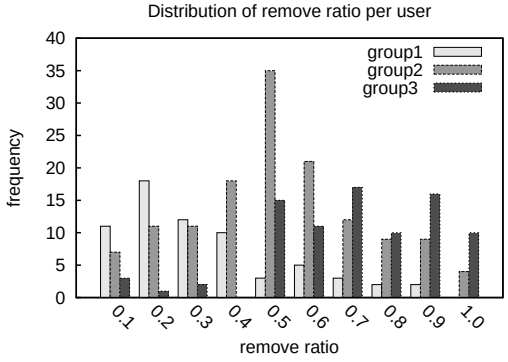

Figure 4: Remove ratios among different groups.

The trend over time of posts added or removed on average among three different groups is illustrated in Figure 3. In general, the results exhibit highly similar trends among different groups. Our observation is that, at each time snapshot, the number of added posts is greater than the number of posts removed in most cases, for all groups. This results in the increasing number of posts accumulated over time. For Group1 and Group2, the number of posts of the type *publication* is slightly higher than *bookmark*. It can suggest that users in the first two groups mostly share publication references than bookmarks, whereas the number of bookmarks posted by Group3 users are significantly higher than publication references.

In addition to raw counts, we also computed *remove ratio* as a fraction of the number of time snapshots a user deleted at least one post. For example, a user u has been a member since 2006, and the user deletion activity is observed 10 times during 15 snapshots in time. Thus, *remove ratio*(u) equals to $0.67 = (10/15)$. Figure 4 illustrates the distribution of users' remove ratios among different groups. The results show that the group of users with fewer posts (Group1) has fewer deletion activities, while the group with more posts (Group3) tends to delete more often.

What trigger a deletion process? Does the number of current posts or that of newly added posts influence the deletion? We sought to answer such questions by performing a correlation analysis by correlating: 1) deletion activities with the total number of posts (or bookmarks or bibtex) and 2) deletion activities with the number of *added* posts (or bookmarks or bibtex). Note that, we only considered any user u with *remove ratio*(u) ≥ 0.5. Table 2 shows the correlation results of deletion activities over time with the total number of posts (**Post**), the number of *added* posts (**Post+**), the total number of bookmarks (**Bookmark**), the number of *added* bookmarks (**Bookmark+**), the total number of bibtex (**Bibtex**), and the number of *added* bibtex (**Bibtex+**), respectively. In general, it can be observed that deletion is highly correlated with the number of resources added, but not the number of total resources users currently possess. Finally, Group1 shows highest correlation results between deletion and added resources in most cases.

Our final analysis is to determine whether given resources are still accessible online. This is motivated by the recent study *Losing my revolution: how many resources shared on social media have been lost?* by SalahEldeen and Nelson [24]. The work has estimated that 27% of resources shared in social media are lost and not archived after 2.5 years. Table 3 shows the total numbers and percentage of resources that were accessible online using their URLs (retrieved on 28 April 2013). On average, there are less than 83% of bookmarks and less than 74% of publication references that were still accessible. This observation suggests that it is important to automatically identify unavailable resources and trigger a forgetting action, e.g., tagging objects as forgettable or deletion, in order to help user handle obsolete information.

Table 3: Resources accessible on 28 April 2013.

	#Bookmark (%)	#Bibtex (%)
Group1	715 (87.73%)	546 (78.56%)
Group2	5,074 (81.34%)	4,396 (73.39%)
Group3	24,909 (78.21%)	3,984 (69.48%)

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented our vision for integrating the concept of managed forgetting into a joint information and preservation management process. This work is still in a very early phase. Nevertheless, we wanted to take the opportunity to discuss the idea of managed forgetting in the preservation community. As a consequence there is still a rich set of future work ahead of us, including: Foundations for the managed forgetting process building upon interdisciplinary work with cognitive psychology; a substantiated information value assessment model in support of the information value dimensions memory buoyancy and preservation value. This also includes the identification of the set of measurable parameters best to be used for estimating those values; and experiments for better understanding the constituents and mechanisms of managed forgetting, e.g., interactions with photo collections, and revisiting behaviors for Web users as well as organizational information seekers.

Acknowledgments This work was partially funded by the European Commission Seventh Framework Program under grant agreement No.600826 for the ForgetIT project (FP7 / 2013-2016).

6. REFERENCES

[1] P. Barrouillet, G. Plancher, A. Guida, and V. Camos. Forgetting at short term: When do event-based interference and temporal factors have an effect? *Acta psychologica*, 142(2):155–67, Feb. 2013.

[2] D. Benz, A. Hotho, R. Jäschke, B. Krause, F. Mitzlaff, C. Schmitz, and G. Stumme. The social bookmark and publication management system BibSonomy. *The VLDB Journal*, 19(6):849–875, Dec. 2010.

[3] Y. Bernstein and J. Zobel. Redundant documents and search effectiveness. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 736–743, 2005.

[4] S. Bowen and D. Petrelli. Remembering today tomorrow: Exploring the human-centred design of digital mementos. *International Journal of Human-Computer Studies*, 69(5):324–337, May 2011.

[5] A. Z. Broder. Identifying and filtering near-duplicate documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, COM '00, pages 1–10, 2000.

[6] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe. Collection statistics for fast duplicate document detection. *ACM Trans. Inf. Syst.*, 20(2):171–191, Apr. 2002.

[7] D. Cosley, V. S. Sosik, J. Schultz, S. T. Peesapati, and S. Lee. Experiences With Designing Tools for Everyday Reminiscing. *Human–Computer Interaction*, 27(1-2):175–198, 2012.

[8] M. Crete-Nishihata, R. M. Baecker, M. Massimi, D. Ptak, R. Campigotto, L. D. Kaufman, A. M. Brickman, G. R. Turner, J. R. Steinerman, and S. E. Black. Reconstructing the Past:

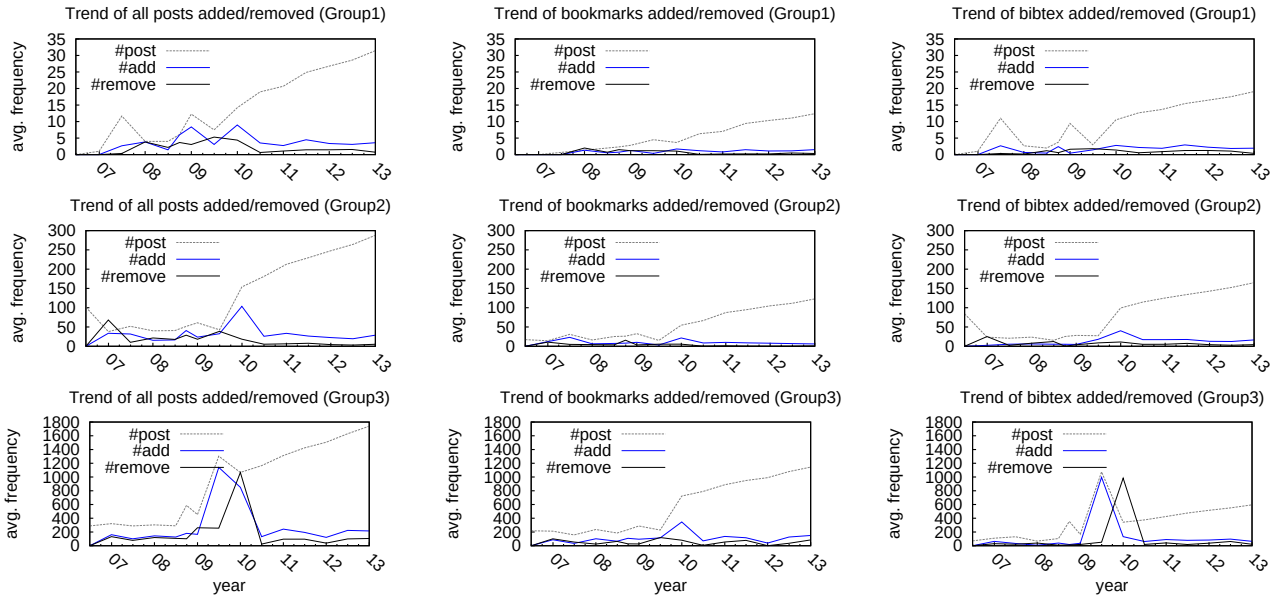

Figure 3: Trend over time of posts added/removed by user.

Table 2: Correlation of deletion behavior with the number of posts, bookmarks and bibtex.

	#Users	Post	Post+	Bookmark	Bookmark+	Bibtex	Bibtex+
Group1	13(66)	0.2632	0.5924	0.0875	0.3214	0.3413	0.6994
Group2	65(137)	0.2140	0.4247	0.1274	0.4195	0.3503	0.6249
Group3	73(85)	0.0918	0.3183	0.0845	0.3615	0.1591	0.4793

Personal Memory Technologies Are Not Just Personal and Not Just for Memory. *Human–Computer Interaction*, 27(1-2):92–123, 2012.

[9] H. Ebbinghaus. *Memory: A contribution to experimental psychology*. Teachers college, Columbia university, 1913.

[10] A. Heathcote, S. Brown, and D. J. K. Mewhort. The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, 7:185–207, 2000.

[11] X. Jin, J. Jiang, and J. L. de la Rosa. Protage: Long-term digital preservation based on intelligent agents and web services. *ERCIM News*, 80:15–16, January 2010.

[12] L. Johnston. We are all digital archivists: Encouraging personal digital archiving and citizen archiving. In *Proceedings of iPres 2011 - 8th International Conference on Preservation of Digital Objects, Signapore, November 2011*, 2011.

[13] X. Li and W. B. Croft. Time-based language models. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 469–475, 2003.

[14] C. C. Marshall. Challenges and opportunities for personal digital archiving. In C. A. Lee, editor, *I, Digital: Personal Collections in the Digital Era*, pages 90–114. Society of American Archivists, 2011.

[15] R. Mayer, S. Pröll, and A. Rauber. On the applicability of workflow management systems for the preservation of business processes. In *Proceedings of iPres 2012 - 9th International Conference on Preservation of Digital Objects, Toronto, Canada, October 2012*, 2012.

[16] V. Mayer-Schönberger. *Delete - The Virtue of Forgetting in the Digital Age*. Morgan Kaufmann Publishers, 2009.

[17] M. Meeter, J. M. J. Murre, and S. M. J. Janssen. Remembering the news: Modeling retention data from a study with 14,000 participants. 33(5):793–810, 2005.

[18] L. Mickes, T. M. Seale-Carlisle, and J. T. Wixted. Rethinking familiarity: Remember/Know judgments in free recall. *Journal of Memory and Language*, 68(4):333–349, May 2013.

[19] T. Palpanas, M. Vlachos, E. Keogh, D. Gunopulos, and W. Truppel. Online amnesic approximation of streaming time series. In *Proceedings of the 20th International Conference on Data Engineering*, ICDE '04, pages 338–349, 2004.

[20] M.-H. Peetz and M. de Rijke. Cognitive temporal document priors. In *Proceedings of the 35th European conference on Advances in Information Retrieval*, ECIR'13, pages 318–330, 2013.

[21] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. *SIGIR Forum*, 43(2):46–52, Dec. 2009.

[22] D. C. Rubin, S. Hinton, and A. Wenzel. The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25(5):1161–1176, 1999.

[23] M. Runardotter, H. Quisbert, J. Nilsson, A. Hägerfors, and A. Mirijamdotter. The information life cycle : issues in long-term digital preservation. *Arkiv, samhälle och forskning*, 1(1):17–29, 2006.

[24] H. M. SalahEldeen and M. L. Nelson. Losing my revolution: how many resources shared on social media have been lost? In *Proceedings of the Second international conference on Theory and Practice of Digital Libraries*, TPDL'12, pages 125–137, 2012.

[25] R. Schmidt. An architectural overview of the scape preservation platform. In *Proceedings of iPres 2012 - 9th International Conference on Preservation of Digital Objects, Toronto, Canada, October 2012*, 2012.

[26] K. Spärck Jones. Automatic summarising: The state of the art. *Inf. Process. Manage.*, 43(6):1449–1481, Nov. 2007.

[27] J. A. Sumner, S. Mineka, R. E. Zinbarg, M. G. Craske, S. Vrshek-Schallhorn, and A. Epstein. Examining the long-term stability of overgeneral autobiographical memory. *Memory*, Feb. 2013.

[28] M. Theobald, J. Siddharth, and A. Paepcke. Spotsigs: robust and efficient near duplicate detection in large web collections. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 563–570, 2008.

[29] N. Unsworth, B. D. McMillan, G. A. Brewer, and G. J. Spillers. Individual differences in everyday retrospective memory failures. *Journal of Applied Research in Memory and Cognition*, 2(1):7–13, Mar. 2013.