Would you like to remember (almost) everything? If yes, you might reconsider after reading the next few stories.

Have you ever heard of Ireneo Funes? He is the main character in “Funes the Memorious,” a short story by the Argentinian writer Jorge Luis Borges, first published in 1942. Ireneo can remember everything, but it turns out that this bogs him down, and he becomes lost in details. He spends his time investigating ideas like inventing a new system of enumeration, which gives each number a unique name (such as “Maximo Perez” instead of “seven-thousand-and-fourteen”), or enumerating and cataloging all his memories in a day. Coping with new details every second makes him unable to focus and to generalize, and he spends his short life lost in thought.

This phenomenon also happens in real life. Jill Price is a California woman who through a rare combination of genes remembers virtually everything that has happened to her since the age of 15. "Starting on Feb. 5, 1980, I remember everything. That was a Tuesday." She also remembers the sensations that came with each event, which makes looking back painful at times. Thus, remembering dominates her life.

“The Web remembers everything” is another phrase that is used a lot these days. While this is not true—Web archiving in reality is spotty at best—things do turn up on the Web that were not meant to be remembered. In addition to the two stories already cited, Viktor Mayer-Schönberger has discussed the “Drunken Pirate” case of Stacy Snyder, a teacher in training and student at the Millersville University School of Education. In 2006, Snyder posted a photo on MySpace with the caption “Drunken Pirate” that showed her drinking at a private party wearing a pirate head. This photo subsequently lead to her being denied a teaching degree, with the dean saying that she was promoting drinking in “virtual view” of her underage students. You can still find the photo online.

Prompted by such cases, the “right to be forgotten” has been a theme of the European Commission in recent years. We invite the interested reader to read the Snyder case as well as an interesting discussion printed as a special symposium issue of the Stanford Law Review online on “The Privacy Paradox.”

A new presentation by the European Commission on the “right to be forgotten” ruling of the EU Court on 13 May 2014 is informative as well; the presentation discusses the scope of this ruling and its relationship to the proposed EU Data Protection Regulation.

In this article, we mainly focus on this question: What should we remember and thus archive, and what can we forget? Each year makes it easier to accumulate large numbers of photos and videos in the social and personal digital space. Their long-term existence is mostly driven by chance rather than by clear guidelines or rules for archiving them. Thus, unfortunately, cases of nonintended disappearance of personal photos happen much too often, caused by a similar lack of survival control as in the cases of extensive survival of “youthful” photos in the public space. How should we proceed? Is it really as easy as providing a “date to forget” to any digital information for an improved survival control, as Mayer-Schönberger speculatively has suggested?

Managed Forgetting and Contextualized Remembering

The European project ForgetIT (www.forgetIT-project.eu) is investigating the introduction of a form of digital or managed forgetting into information management environments. The project, managed by the L3S Research Center at the Leibniz Universität Hannover, focuses on the idea of making more conscious decisions about which content is really important, and
thus should be preserved safely, and which content we can (and should) forget. Although we often follow a keep-it-all strategy for our content, in reality the content is subject to a form of random forgetting, for example, as a result of hardware crashes or when storage formats and technologies become obsolete. What we need instead is a principled strategy of forgetting, preservation, and remembering.

Inspired by the central role human forgetting plays in helping us stay focused, in ForgetIT, managed forgetting supports resource selection for preservation. Preservation refers to technological and organizational methods, which ensure a long-term survival and accessibility of content over several decades and longer. At first glance, forgetting and preservation might seem to contradict one another. However, because preservation comes at a cost, it is important to make conscious decisions about which resources to invest in. In ForgetIT, managed forgetting is complemented by synergetic preservation, which makes preservation technology easier to use in personal information management. Finally, the third cornerstone of the project is contextualized remembering, which equips preserved content with sufficient and evolving context information. Such a managed approach will keep our archives understandable and useful, thus ensuring a type of semantic survival of the selected resources. This is inspired by the constructive nature of the human remembering process. Personal preservation—how we deal with our personal information such as photo collections over time—is one of the application domains of the ForgetIT project.

**Which Photos Should We Archive?**

In a first systematic research work on automated photo selection for preservation, we asked 35 users to select 20 percent from one of their own collection of photos using the metaphor of a digital vault (see Figure 1). This analysis of untagged photo collections of up to several hundreds of photos per collection and more than 8,000 images overall was used to analyze the factors (and features) that dominate photo selection decisions in real-world settings—for example, how might you select photos from a collection brought back from a holiday trip? The photo selection task was accompanied by a user survey, which elicited attitudes towards preservation and perceived photo selection criteria.

The captured selection decisions of the users were employed for learning the most dominant features for automated photo selection methods. To develop these methods, two observed behaviors of users in photo selection were considered (see Figure 2): First, users tended to follow a reduction-oriented strategy of getting rid of duplicates and near duplicates to gradually reduce the decision space. This behavior inspires the development of reduction-oriented methods for automated photo selection. Second, users tried to cover different subevents with the set of photos that they selected. This motivates the development of coverage-oriented methods, aiming to cover all the subevents of a holiday trip, for example. Event coverage was also often named as an important selection criterion in the accompanying survey. It has to be noted, however, that subjective measures such as that a photo “evokes a memory” or “is important to me” were rated even higher as reasons for photo selection by the users in the survey. This suggests that automated photo selection is a difficult task driven by a variety of partially subjective factors.

Starting from the observed reduction-oriented and coverage-oriented user behavior, we developed and compared coverage-based and reduction-oriented methods for photo selection. The enabling mechanism underlying both approaches is photo importance prediction, which is learned based on a rich set of features including quality-based, face-based, and collection-level features. A special focus was
placed on concept-based features, which rely on a method for detecting high-level concepts in a photo in a two-step process, thus at least partially capturing the semantics of a photo. For the coverage-oriented approach, this is combined with a semantic clustering method, which relies on concept detection and subsequently clusters photos based on those concepts and time, aiming for an event-based grouping of the photos. Importance prediction for the individual photos is used to select photos from the individual clusters. For the reduction-based approach, duplicates are aggregated based on a relaxed near-duplicate detection method, and subsequent photo selection (based on importance prediction) considers them as a single photo.

The experimental evaluation of the two methods inspired by human behavior shows that both the reduction- and coverage-based approaches are promising solutions to address the difficult task of selecting photos for preservation, with the former outperforming the latter. In analyzing the relevant features, we clearly observed that the use of semantic information (concepts and faces in image) and collection-based features are core ingredients for tackling this problem. Our experiments also confirmed the user survey results, suggesting that image quality is only a secondary criterion for this photo selection task.

Other Work on Photo Selection
Automated photo selection has been studied in a variety of settings including photo summarization, selection of representative photos, and automated creation of photobooks and collages. A variety of factors influences the selection of photos from a photo collection. Clearly, the mix of factors depends on the photo selection purpose. For example, we would use different criteria to pick photos for sharing on Facebook than to decide which photos to preserve for the long term.

The development of photo selection methods also has to take into account photo taking and management practices. Early work showed a change in such practices implied by the growing proliferation of digital photography (such as the large increase in the number of photos). Relevant social practices surrounding photos are photo sharing practices driven by the digital medium and social media, the increase in more mundane motives for taking photos (such as food), and the mix of purposes for photo taking (and preservation) including self-presentation, social aspects, remembering, and reflection. (Previous research provides a good overview of such social practices.)

Features that have been investigated in the context of automated photo selection are extracted from the content itself (visual features), its context (temporal information), or photo annotations. They include more objective factors such as photo quality, aesthetics, and depicted entities, especially people. Congcong Li and his colleagues presented a method for translating important criteria for photo aesthetics into methods that can be applied for automated photo selection. However, photo selection is also driven by more subjective factors, for example, when a picture captures an important moment or when it displays what the user wanted to capture. A qualitative study on “keep or delete” decisions for photos showed the importance of subjective criteria in photo selection. In this study with 74 participants, three categories were identified as the most frequent reasons for deleting photos: reaction (including photos that have a negative

![Figure 2. Automated photo selection for preservation. Following user observed behaviors, the developed methods use reduction-oriented strategy to prune collections and coverage-oriented methods to ensure all subevents are represented.](image)
connotation for the user or are not sufficiently interesting to the user), aesthetics (including quality), and no show (the photo does not show what it is supposed to show). The importance of subjective factors was also confirmed by another study, where the inclusion of features learned from the eye-tracking behavior (as a signal of personal interest) of a user into the photo selection task considerably improved results.

In addition, photo selection does not consider individual photos in isolation. Therefore, collection-based features such as coverage are also considered in photo selection approaches, thus reflecting human behavior in photo selection. For example, one photo summarization approach used a multigoal optimization approach for combining coverage of the concept space defined by the original collection and diversity of the selected photo collection. An analysis on the collection level can also uncover signals for increased user engagement while photo taking. This might for example result in a higher number of near duplicates or in shorter time interval between subsequent shots. Such signals can be important evidence for deciding which photos or photo groups might be more important to the user.

Remembering the Right Things

To close the loop of forgetting and remembering in a digital memory, it is also necessary to consider what is remembered and how “memories” are recalled. One cultural probe study showed that memories are not restricted to special events; people also like to remember more mundane and everyday events of the past because they are regularly repeated, of social value, or exemplary for a person’s character. One of the further identified reasons for remembering past things is temporal contrast, such as things that were done differently in the past. Such differences in habits and situations also impose challenges in bringing back digital “memories.”

For this case, work on contextualization suggests an approach for adding context information in a time-aware way for easing interpretation of older content (such as after decades). Furthermore, it is also important for digital remembering to find relevant memories in the large amount of stored digital memories.

We investigate a related topic as part of our ERC grant project Alexandria (http://alexandria-project.eu), where we are interested in finding the most relevant results in large-scale, multiversion, noisy document repositories, as we have with Web archives. One approach we evaluated in the Alexandria project was to consider historical query intent, investigating for each aspect of a given query, what time point was this aspect most important and preferred over other time points? This helps, for instance, to “remember” the history of a politician or a controversial topic as reflected in an archive, in an adequate and diverse way.

When Not to Share Photos

As mentioned in the introduction, photo sharing and privacy problems often arise when photos are visible outside the originally intended scope. These might be social scopes, such as private versus work environments or friends versus family environments, or a temporal scope such as teenage photos in an employment environment.

To address this problem, we developed a machine-learning-based method for reducing the risk of unintentional exposure of private photos (see Figure 3). Based on an average community notion of privacy elicited from a social game, we designed an automatic privacy classification method. Our experiments show that by exploiting a set of high-level and object-centered features such as the occurrence of faces, edge-direction coherence, and SIFT (scale-invariant feature transform) features, considerable classification performance can be achieved, even without taking into account textual features.
As a final remark, let us refer back to Lawrence Lessig’s “Code and Other Laws of Cyberspace,” first published in 1999, with an updated version in 2006. In this book, Lessig brilliantly reminded us that protection and regulation of privacy (and other values) is not done only by law, but also by our norms, markets, and obviously the architecture of the systems we build as computer scientists. Let us build the right code to manage forgetting and remembering on the Internet and the Web. Let us build code to help protect privacy in the right way!

References


Wolfgang Nejdl is a professor of computer science and the executive director of the L3S Research Center at the Leibniz University of Hannover, Germany. Contact him at nejdl@l3s.de.

Claudia Niedere is a research group leader in the L3S Research Center at the Leibniz University of Hannover, Germany, and is the coordinator of the European project ForgetIT. Contact her at niedere@l3s.de.