

How People Manage Information over a Lifetime

CATHERINE C. MARSHALL¹

1.1. Introduction

How will we manage our heterogeneous collections of digital information over a lifetime? How will we look through many decades' worth of digital belongings? What will our digital legacy be? We can think of long term PIM issues from three equally important perspectives: (1) storing the digital belongings we have amassed over the years (long term storage); (2) maintaining these belongings in a form that they can be viewed, used, or possibly even changed (preservation); and (3) providing individual and collaborative mechanisms for reclaiming these belongings from long term digital store (access).

1.1.1. Stories characterizing important long-term PIM issues

Many PIM scenarios are focused on immediate information needs and uses: for example, browsing through vacation photos to find one to send to a favorite aunt or gathering the right documents to prepare for a meeting. The following vignettes couple the characters created for this book with interview data collected during recent studies to highlight longer-term issues.

Predicting value and metaphors for long-term access. Brooke tapes two tickets to her dressing-table mirror to remind herself that she's going to a concert Wednesday night with Derek and some other friends. After the concert, she tacks one of the ticket stubs to the wall of her cube at work as a memento of a wild evening. After the start-up goes under and she brings her personal items home, the ticket stub goes into her "treasure box", an old footlocker she keeps in the guest room closet. Originally this personal information is kept in sight to remind Brooke when the concert is and to give her an emotional lift as she anticipates the event. After the concert, the ticket stub changes roles – it becomes a reminder of a fun evening in the recent past. Finally, Brooke puts her ticket stub in a place she expects that she'll encounter it again – possibly many years later – and that this small ragged piece of printed cardboard will evoke a number of associated memories. She'd never think to go looking for a ticket stub, but she knows that when she sees it again, it will not only remind her of the concert, but also of her friends, her musical tastes, and a whole period of her life. The value of the ticket changes as time passes: Brooke could have tossed the stub after the concert or when she took down the things pinned to her cube wall, but instead she has elected to keep it "forever."

Digital context. Alex has performed research on an esoteric topic intermittently since he was in college; it's his secret – not even his family knows he's interested in Rongorongo. He has even

¹ Microsoft Corporation (cathymar@microsoft.com)

published several papers about the Easter Island script and saved numerous resources he has discovered on the Internet. He has cited some of the articles in his publications, but others he has not. In fact, he hasn't even read all of them, but he keeps the whole collection with the idea that he'll refer to the articles when he needs to and read them carefully when he has time. Some are from the grey literature and, as such, have not been published through normal commercial channels (Wood, 1984), but others are retrieved from digital libraries and special collections. Although he acknowledges that it would be easy to recover the resources, he feels that if he lost them he would never be able to remember them all. He has even considered making an index for his collection, so if anything happened to his computer he would know which articles he'd lost.

Distributed storage and format opacity. Jenny is Derek's younger sister. Like many people, Jenny's most important personal information management concern is maintaining her photos so she'll have them when she's older. She's in college now, but she already has collected a substantial number of photos. She cares for some of them herself – particularly those she has taken using her digital camera – but her mother still takes care of those she considers “family photos.” Jenny's own photos are by no means stored in one place: she has printed some to organize in traditional albums and to tape to her bedroom door; others are digital-only, stored on the hard drive of her laptop and on photo CDs in a cabinet. Still other photo CDs are scattered around her room; sometimes they get mixed in with her music CDs and software. She lost several batches of photos in a recent crash that she attributes to a virus she got from sharing music on Limewire. Jenny does not consider herself to be a computer expert and usually enlists Derek's help when she has computer troubles like this virus. She knows her camera can produce several different resolutions and formats – and that some of the applications she uses to manipulate the photos produce files that her friends can't display – but she's never confident of which format to use and just uses the default settings.

Curatorial effort and predicting value. Derek has kept his email for as long as he can remember; he keeps both school-related email, especially for the attachments, and personal correspondence. Both are important to him: his school email serves as a means for maintaining a record of his work – he knows he'll dip back into it when he's studying for the bar – and his personal email acts as a journal and an index into his rich store of memories. It is also storage of the last resort: if he can't find a file in his primary computer's file system, he looks for it in his email. He has even sent attachments and other important bits of information to himself to ensure he has copies available at all times. In spite of the varied and important roles that email plays in his life, Derek maintains that he wouldn't be overly upset if he lost his current email file – all 11,000 messages – in part because it is such a troublesome accumulation of content. He never feels like it has been properly culled and he doesn't know how to back it up.

1.1.2. The importance of long-term personal information management

From these vignettes, and from growing attention in the digital library community (Beagrie, 2005) and the newsmedia (Hafner, 2004), it should be evident that storing and maintaining personal information over the long haul is an important topic that raises particularly challenging issues in a digital environment. These issues cover significant technical, social, and legal territory. Technically, we must address the storage, preservation, and long-term access of digital

materials; socially, we need to consider the roles of various emerging genres (such as blogs and personal web sites) in our culture and over the course of our lives, including how and when we want others to have access to them and how we can make these things we have saved intelligible to others; legally we must consider how personal digital materials interact with the holdings of other online services (such as personal financial records that are held by a bank), as well as how material protected by Digital Rights Management (DRM) (Stefik, 1999) may be preserved, and how a personal archive can contain materials that are drawn from digital libraries and other stores of copyrighted material.

Despite the acknowledged importance of digital personal information, it is difficult to convince many people of the urgency of this problem. On one hand, we have been trained to approach technological progress with an air of optimism: by the time we want to open the fifty-year-old photos of our children, a viewer will have been implemented to decode whatever obsolete format they were stored in and will render the photos with perfect fidelity on the display of the moment. This perspective goes hand-in-hand with the strategy of benign neglect that most people apply to keeping their physical personal materials: photos, letters, legal and financial records, and other important keepsakes are tossed into boxes (or, at best, filed carefully) and put somewhere safe, in an attic, under the bed, in a safety deposit box, or in a closet, and left undisturbed for many years. Thus, many of us believe we are accumulating our valuable digital stuff in a small number of circumscribed places, and when the time comes, we expect to be able to simply pull out the files and look at them in much the same way as we would look at our second grade class picture in the box on the closet shelf.

On the other hand, digital information invites an attitude of radical ephemerality. We have all lost digital materials and by now many people tend to view disk crashes, computer viruses, and media obsolescence with a certain sense of inevitability. They commonly use the metaphor of a house fire, and assert that one must simply move on. From this perspective, we cannot expect to have any of our personal digital information in fifty years; it will be long gone and we might as well get used to it.

In practice, most people inhabit a space somewhere in between these two caricatures and recognize themselves in both extremes. They know they should make copies of valuable files; they know they should worry about the lifespan of various digital storage media; they believe they should cull the valuable files lest they get lost amid an accumulation of indifferent digital dross; they believe that digital formats can be converted without loss to newer formats; and they suspect that anything they have found rather than created can be found again through clever searching (Marshall et al., 2006). But everyday human actions belie these good intentions: files are not copied; storage media are not refreshed; and digital files accumulate at a frightening rate. Simply put, benign neglect will not be sufficient to keep our digital things safe for a lifetime.

1.2. PIM archiving issues

Libraries and cultural heritage institutions – along with records management organizations – have long grappled with archiving problems and were among the first to address digital archiving issues. These institutions have been joined by other disciplines on the forefront of the production and use of digital information, the digital arts and sciences, as well as folk historians and Internet guardians. But are they all talking about the same thing when they talk about digital archiving? Not really, but there is considerable overlap in their missions and shared concern with literally preserving the bits that represent heterogeneous digital materials, ensuring that we will have access to them and be able to view or use them in the future. At the very least, digital preservation efforts must address issues associated with storage media (e.g. reliability, durability, and media format), hardware and software environments (e.g. operating systems, drivers, and shared libraries), application-specific formats and functionality (e.g. file formats and codecs), and display capabilities (e.g. resolution, fonts, and color). Figure 1 illustrates some of the additional concerns that different disciplinary practices and institutions bring to the table.

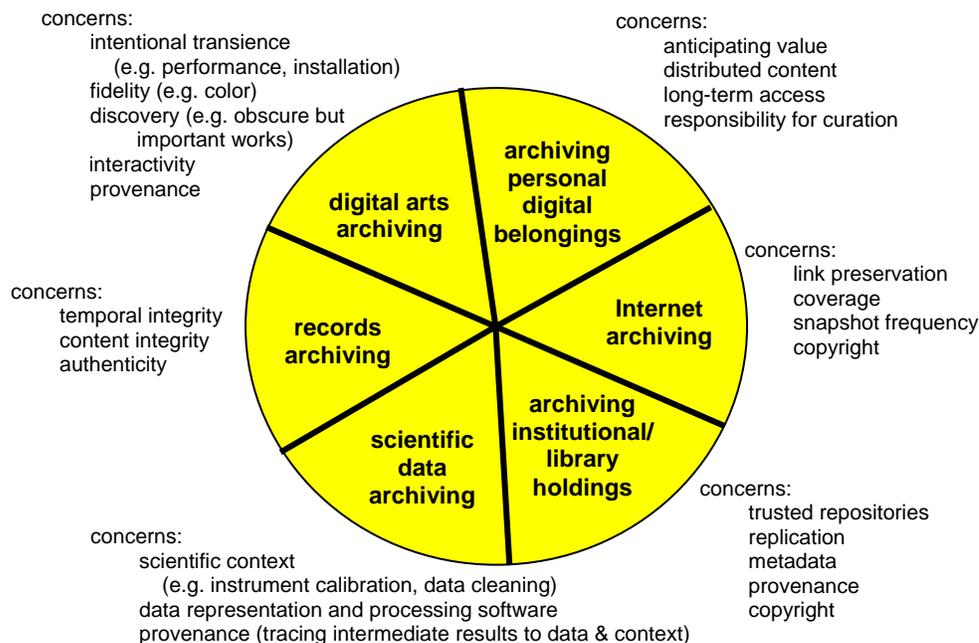


Figure 1. Digital archiving concerns arising from different disciplinary practices and institutions.

This diversity of interests may make the problem seem hopelessly complex; but instead we can view each discipline as providing traction on a slightly different set of issues. For example, e-science archiving efforts are establishing best practices for managing distributed datasets. Digital arts archiving is addressing tricky questions associated with preserving the interactivity or visual fidelity of a hand-crafted work. Records archivists are developing techniques that may be used to maintain the integrity of our personal records. Libraries are driving the development of institutional repositories (Tansley et al., 2003), canonicalization and migration procedures (Lynch, 1999), and replication techniques (Cooper et al., 2002). Finally, the Internet Archive is

implementing policies and methods for preserving large scale hyperlinked structures (Lyman et al., 1998). In short, although the problems are far from solved, substantial efforts and programs are underway (Arms, 2000).

But what of the requirements that are unique to personal digital archiving? Naturally, personal archiving revolves around the same basic technological issues as other types of digital archiving – how materials are stored; how they may be preserved, and how long-term access may be supported – but if we re-examine these issues from a PIM perspective, there is still much work to do. In this chapter, we consider seven key attributes of personal digital belongings that should shape our development of archiving technologies:

- (1) digital material accumulates quickly, obscuring those items that have long term value;
- (2) digital material is distributed over many different off- and online stores, making it difficult to keep track of individual items;
- (3) digital material derives a substantial amount of meaning from context that may not be preserved with an item (for example, references to email attachments, links to Web pages, and application-specific metadata);
- (4) digital material is easily passed around and replicated, introducing a tension between protection (of privacy, copyright, and security) and future access;
- (5) digital formats are not only opaque to many users, but also may be incompatible with available applications or may become obsolete;
- (6) curating personal digital archives is time-consuming and requires specialized skills; and
- (7) our current computing environments have yet to incorporate mechanisms and metaphors that support long term access.

Predicting value. Most personal information in the digital world is not collected intentionally as part of a coherent collection; instead heterogeneous materials accumulate invisibly over time. Because it is inherently difficult to anticipate future value even among professionals trained to evaluate and discard (Levy, 1998; Baker, 2001), personal digital belongings often accumulate very quickly. This trend has been exacerbated by drastic reductions in storage cost; it seems much cheaper to keep everything than it does to assume the overhead of up-front evaluation of individual items (Gray, 2000; Beagrie, 2005). Furthermore, personal digital belongings have varying lifecycles; some of the accumulated items are valuable for a decade, some for a lifetime, and some beyond, and few of us can tell the difference. The complexity of the digital lifecycle lends further credence to the argument that we should keep everything and worry about decoding or evaluating it later (Gemmell et al., 2002; Cutrell et al., 2006).

However, there have been serious critiques of the “keep everything” approach because of its eventual consequences, the effect that having too much has on human attention (Jones, 2004). There is also a profound question about the emotional or economic viability of keeping everything, even if it isn’t visible, as suggested in (Czerwinski, et al. 2006); discarding unwanted or unpleasant material is often a valuable exercise and it may even be legally mandated in some situations (for example, personal information that belongs to an employer).

Distributed storage. Personal information is often stored in a distributed way – on the hard drives of different home and work computers, on removable storage media, on remote servers, on personal devices, and printed as hardcopy. Not only does this distribution cause people to lose track of where specific digital belongings are and which version of a file is the most recent; digital storage media are also vulnerable to failure, obsolescence, and loss and may be unreliable from the outset.

One way digital belongings become distributed is through replication, often for safety's sake. For example, a person might save and share favorite pictures in multiple places: on a local hard drive, on external media such as CDs, on photo-sharing services such as Flickr, on personal web sites (often provided through hosting services), on camera memory cards, and on friends' and family members' computers (possibly in their email). While this replication provides a simple and effective safety net, it also introduces curatorial complications such as identifying the preferred version for a particular use (e.g. choosing among photos stored at different resolutions) or retaining appropriate metadata (e.g. the real creation date).

Digital context. Archival best practices require careful attention to preserving the material's original *context*. Thus when we wrestle with the problem of the long-term maintenance of personal digital belongings, we can't simply attend to each digital object as a stand-alone item, but rather we need to capture the characteristics that make it part of a whole. For example, a digital photo is more than just the image content: it is also the metadata the camera records, the metadata the photographer adds (for example, tags to identify the subject of the photo); and its membership in a set of photos the photographer has taken at the same event. It may be attached to an email message, included in a blog posting, or shared on a web site. One of the most powerful aspects of maintaining digital materials in their original form (rather than, for example, saving them in print) is how this context may be kept alive over time. However, preserving a large, distributed, linked structure and its metadata is a daunting problem (Lyman and Kahle, 1998). Furthermore, context may be lost when digital belongings are moved to new computers or accessed by different applications. For example, file creation date is often lost when the file is copied or moved. From our vignette describing Derek's personal digital library, it is easy to see how the metadata may be as important (or more important) than the item itself.

Conflicting interests of protection and long-term maintenance. Individuals, institutions, and organizations all have an interest in protecting their digital belongings from unwanted access, unauthorized use, or digital piracy. They gravitate toward strong solutions such as encryption or DRM in addition to minimal protection such as passwords. While these solutions provide attractive short term protection, over the long term they may inhibit the ability to preserve and access the data they are protecting (Lavoie and Dempsey, 2004). For example, passwords and encryption help us keep personal information private; however, they may also render it inaccessible as passwords are forgotten or encryption keys are lost. Copyright protection such as DRM may make publishers more comfortable with the safety of their intellectual property; however, DRM can prevent individuals from taking appropriate preservation measures (such as creating extra copies or migrating content to newer formats).

Format opacity, incompatibility, and obsolescence. Most straightforward strategies for the long-term management of personal information are complicated by format opacity and a growing number of standard formats, many of them incompatible, and some already obsolete. Format opacity mainly arises from the desire to keep complexity hidden from the user. For example, instead of asking a user whether she wants to encode her video in MPEG-1 or MPEG-2, an interface might ask whether the video will be saved on a CD or a DVD. The user may not be aware of the consequences of using one format over another; for example, codecs may be unavailable for certain platforms.

Curatorial effort. Institutional archiving is time-consuming and costly. Often solutions that require user action (for example, converting files to canonical formats like PDF or assigning metadata values) also mandate that institutions introduce incentives (or punishments) to ensure participation (Beagrie, 2003). This problem is exacerbated for PIM. Many individuals keep their computing environments running by enlisting ad hoc IT support – relatives, friends, or colleagues who perform common system administration tasks on their behalf, such as installing new applications, setting up new peripheral devices, or coping with malware infections (Marshall et al., 2006). These ad hoc IT people have varying degrees of knowledge and sometimes come into conflict with one another. Any sort of long term personal archiving technology will need to consider the demands it places on this variable support environment; the success of a solution may depend on its ability to run without significant intervention.

Long term access challenges. One of the most provocative problems that arises from personal archiving is one of access. Clearly you can't deliberately look for something that you don't remember you have, so even the most functional desktop search is no panacea (Marshall and Jones, 2006). In fact, field studies often reveal that not only do people forget particular items that they've saved; they also forget entire categories of saved material or places that they've stored treasured items (Whittaker & Sidner, 1996; Bruce, et al., 2004; Marshall and Bly, 2005). They believe they have kept things that they haven't and they are surprised by what they do find in their long-term stores. This experience of keeping valuable material (or what is conceived of as valuable material at storage time) in a specific well-known place – such as the box under the bed – leads us away from the desktop metaphor and into a realm of place and value. While these storage places for valued material may share some of the properties of the desktop – they may be well organized and highly structured, or they may be informal catch-alls, or some of both – the desktop is aimed at short term information management strategies, not for keeping a lifetime's worth of belongings.

1.3. Case study: a long term collection of email correspondence

In this section we describe an effort to assemble a single archive from six years' worth of email between two correspondents, M and Q; this description captures some archiving issues and allows us to introduce technologies to address them. Over the six year period, close to 1400 unique messages – 800 pages of text – were exchanged. Both correspondents value and save

their email, much as other people value and exchange photos.² Unifying separate email streams into a single archive for storage is important for both preservation (canonicalizing the formats) and access (mostly done through selective browsing, not through targeted search).

In the past, benign neglect was sufficient to ensure that old letters would survive into the future; people would bundle them together with a ribbon, put them in the box under the bed, then be able to read them sixty years hence. Putting only six years' worth of email into a coherent archive entailed a significant amount of effort and was at times a test of IT skills. In principal, it would have required far less effort with automated support for "real" archiving.³ We will explore each of the issues we discussed in the last section as they apply to this email archive.

Predicting the value of email. Even though both correspondents thought of the entire archive as valuable chronicle of their lives during the 6 year period, the individual exchanges varied greatly in their archival worth. This value discrepancy mainly stemmed from the fact that the email spanned the normal range of message types (Boardman et al., 2004). Messages used to coordinate face-to-face meetings ("See you at Bridgepointe at 1:30") or just to keep in touch may be less valuable later on than messages that were more substantive⁴. Even the substantive messages varied greatly in tone and worth. Some contained everyday news; others were ruminative; and still others were accounts of crises and remarkable circumstances.

This unpredictable variety of content ensures that there is no easy way to filter it for the most interesting messages; nor would a search capability stand alone as a mode of access to the collection. Much like other material saved over a long period of time, some crucial elements have been forgotten, and others have changed in function. In the next section, we will discuss the implications of value on heuristics for visualizing and accessing a long-term accumulation of digital belongings.

Distributed storage of email. Many people manage multiple email accounts. They may have a variety of motivations for doing so: to separate personal and work correspondence; to address various limitations of their various computing environments (for example, some email accounts are better suited to receive or send attachments and some have size limits); or to maintain a multiplicity of identities coupled with divergent interests (for example, people who participate in online dating or auctions often have separate identities for these specialized activities). Yet the exigencies of real-world situations dictate that this separation may not be strictly maintained: people receive personal email at work; their separate identities blend; and they violate their own policies for using one account over another. To make matters more complicated, some email storage is server-based; some is client- or device-based; and some allows people to maintain distributed collections. Thus, the trend is toward distributed materials.

² It should be noted that the literary letter is considered an imperiled form due to the shift to electronic communication (Donadio, 2005); we expect email archiving to become part of managing special collections.

³ Many email systems allow one to archive messages, but it is not archiving in the sense that the word is used here. It is archiving *relative to that particular email application* and it is primarily designed to sort inactive mail from active mail.

⁴ Although, much like the ticket stub in our introductory vignettes, a brief coordination message may change its role over time, after the event has taken place.

Our illustrative case is no exception. Over the six years covered by the archive, Q and M each used six different email accounts, some personal and some professional, to support their correspondence. While the email in question formed a single conceptual collection, the *de facto* archive was fragmented and difficult to re-integrate. Figure 2 illustrates this fragmentation; each email account is shown according to the duration it was active. The yellow bars represent local storage; magenta corresponds to device-based storage; and turquoise indicates client-side storage. For example, from this visualization, we can see that during the second quarter of 2000, Q’s mail might have been sent or received using any of four different accounts and M’s might have been sent or received from any of three accounts.

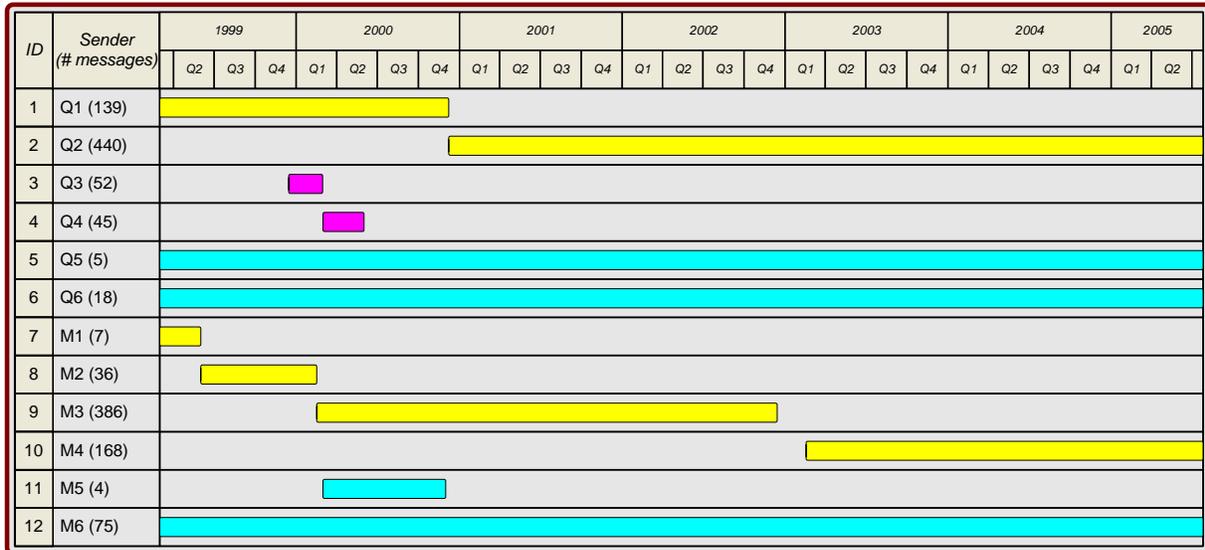


Figure 1. Long-term use creates a distributed collection, with multiple email accounts active at any given time.

The flip side of distribution and replication is redundancy. A coherent browseable archive must anticipate redundancy to support future access. Some redundancy is planned as risk-reducing replication (for example, an entire personal email file may be copied to a server) and other redundancy stems from common practices such as including the whole of the thread in the current message or cc’ing oneself to make the email accessible from another account. Over time, redundancy is both a safety net and an annoyance. For example, an included thread may contain the only copy of a message; on the other hand, messages sent to two or three accounts means that there will be multiple copies of these messages in a merged archive.

Digital context as attachments and URLs. Increasingly, personal email is not a stand-alone collection of messages that refer only to one another (Whittaker et al., 2006). Two distinct phenomena add complexity to the long-term sustainability of email: (1) references to external material using URLs or links to web pages; and (2) attached material such as formatted documents, photos, and other email (which paradoxically may be in a format unintelligible to the individual’s own email application). Some of these references and attachments are transient and do not contribute substantially to the value of the exchange such as jokes and newspaper clippings sent as “thinking of you” gestures (Marshall and Bly, 2005). But others, such as personal photos, are valuable additions to the archive; in fact, email is an important mechanism

for keeping this kind of material (Bruce, et al., 2004). Furthermore, some of these references, while not significant in and of themselves, may be vital to understanding an elliptical discussion. Yet it is relatively common to lose both Web references and attachments over the life of an email archive, in spite of existing functionality that makes it easy to save them at the time of receipt.

Let's look at a few of the attachments and references to identify some common problems. Most simply, attachments may be missing: they may have been downloaded and deleted at the time of receipt. Similarly, references may point to missing Web pages or sites, which is unsurprising since the average Web page is only active 44 days (Lyman, 2002). Sometimes the Web pages themselves have been attached to messages; unless the sender has been careful, this may result in incomplete content (e.g. images or component frames may be missing). Furthermore Web pages are at best fluid; the content may be changed so as to render the reference nonsensical.

Thus retrieving and preserving the content at the time of original access seems like a sensible strategy for ensuring that the intelligibility of the exchange is maintained. But this not only creates technical problems – it is yet another window onto the general problem of preservation – it also requires attention to the fair use provisions of copyright law. How much external content may be kept without further permission? What about attachments like newspaper clippings that are free when they are current, but must be retrieved for a fee later on? It is easy to see that any solution is not without attendant complications.

Conflicting interests of protection and maintenance: the case of the missing password.

Personal email is usually private; thus in the case study we might expect problems arising from the correspondents' efforts to ensure privacy. Fortunately this did not involve tackling the problem of missing encryption keys; instead passwords were used to protect email stored locally.

Unsurprisingly, all of Q's email is password-protected. Server-side passwords are usually easy to recover by contacting a system administrator or issuing an automated password request. In fact, these email stores will not become problematic until the account owner is no longer capable of accessing them; at that point, the tide will turn, since no provision has been made for changes in who may see the email (Czerwinski et al., 2006). This seemingly remote possibility has proven to be a very real obstacle in military situations in which soldiers' email cannot be accessed by relatives after the soldiers' deaths.

The inactive Microsoft Outlook file (Q1 in Figure 2) is the stream we will focus on. It was stored locally on a household computer, one shared by other family members and houseguests; thus password protection seemed prudent. Over time the password was forgotten and could not be reconstructed by normal means such as trial and error or consulting likely-looking scraps of paper. The email was adjudged to be sufficiently valuable to justify the cost of attempting a break-in. Efforts to locate instructions and a utility to remove a local password were successful, but opened the door to malware (the software could not be obtained from a trusted source) and damaged message headers, including dates. From this simple example, it is apparent how security interests may render older personal information inaccessible. Security may be too effective to be so easily thwarted; issues of trust may make it difficult to choose an acceptable solution if it involves running a strange executable; or the problem may be too longstanding to be addressed by solutions currently available.

Format incompatibility of merged email. Streams Q1 through Q6 are stored in four different email formats, reflecting the different applications used to manage the email. If these streams are stored without the application used to manage them (a common archiving scenario), they may be indecipherable later; they will require canonicalization either at merge time or when they are displayed. Not only must the formats be renderable, but also they must be made *readable*: an important consequence of format differences, genre shifts, and mail application conventions, defaults, and options is that a merged stream may be difficult to read. For example, forwarding, replying, and device characteristics all result in odd line breaks. Although on the face of it, this seems like a relatively minor problem, it detracts from the experience of reading emotionally valuable old email.

Although it can be argued that email presents special formatting issues, there are more profound challenges in other media, especially as it accumulates over time. In some cases, such as digital video, shifting standards and ever-improving resolution and capabilities represent a rendering quagmire – just preserving the ability to view the digital object is the central aim of many standardization efforts and much research. To achieve some kinds of stability, such as stabilizing a document's appearance, special archival formats such as PDF/A have been proposed (LeFurgy, 2003); to maintain the editability of a document, others have speculated that fully emulative solutions are necessary (Rothenberg, 1998), although such emulation solutions must be justified by use, since full emulative fidelity may be more expensive than it looks at the outset (Reichherzer & Brown, 2006); still other digital belongings may require attention to interactivity (Waldrip-Fruin, 1999; Marshall and Golovchinsky, 2004). This aspect of preservation is perhaps the most thoroughly investigated in a variety of domains; it is wise not to become too embroiled in this single issue at the expense of the others.

Curatorial effort in assembling and maintaining an email archive. Our observations about the time-consuming and often difficult nature of digital curation are born out by the case study. Not only was the construction of the unified archive labor intensive; even maintaining the separate streams of email over long periods presented challenges. As we have already noted, it is not uncommon for an individual to manage multiple email streams involving different applications, devices, and stores.

First, let's consider the personal/professional separation (the Outlook email files in our case study); these are tied to work email accounts on Exchange servers. Curation usually starts to be an issue when the individual leaves a job; personal email must be culled from the work email that must be left behind. Moving Outlook .pst files is not straightforward; they are not located in a place in the file system that consumers are apt to be familiar with, nor can they be moved using a straightforward copy. For those people who consider email as a fundamental PIM record, the files may also be very large. Furthermore, attachments require separate attention if they are to be moved and tracked with the email stream; they are often stored in temporary directories that are ignored by utilities. Maintaining the Outlook .pst files thus requires specialized knowledge and cannot be left to benign neglect.

Next, let's consider the ubiquitous free server-side email accounts. Often they seem to require little or no maintenance; in fact, consumers are increasingly relying on the archival properties of

these email accounts (Marshall et al., 2006). However, there is no assurance in the End User License Agreements that this email will be stored permanently. Without curatorial attention, these accounts represent a paradoxically chimeral form of archiving: most consumers have no idea how they would reclaim the content of this type of email stream if they decided to abandon the account or the provider stopped offering this service. The demonstrated consumer impulse is to go through the messages, copying them one by one if the email is perceived to be at risk. Surely this is not a viable curatorial strategy.

Long term access of email in the case study. It is easy to defer the problem of long term access to the powers of search: desktop search and techniques like Implicit Query (Cutrell et al., 2006) have improved greatly in recent years. But is search going to be the only mode of access to many years' worth of personal digital belongings? Certainly it will address information needs scenarios: a lost piece of financial information, a dimly remembered photo from one's childhood, or the name of a friend of a friend. However, in our case study, much of the reason for keeping the correspondence is far less tangible; it will be used to remember times, people, and places not documented anywhere else.

That said, it is a considerable volume of text. How might interesting or valuable content be highlighted? It is tempting to think that we can organize an archive according to the information in a personal gazetteer. Analogous to a geographical gazetteer, a personal gazetteer might draw on records from an individual's calendar, contacts, and maps to create a database of important people, places, and events. This technique would combine the work on gazetteer-based digital reference services (Buckland, 2004) with automated personal media organizers such as (Graham et al., 2002).

However, such heuristics should be approached with caution. If we look at the material gathered in the case study, important events (birthdays, holidays, and the like) are acknowledged, but they are more often the occasion for platitudes and brief greetings ("As the Subject line says, Happy Birthday! ... We'll have to do something to celebrate.") On the other hand, remarkable events – uncomfortable Match.com coffee dates, alarming nightmares, and unexpected conversations – are more likely to spur the kind of documentation that is evocative and worth re-reading. Break-ups are not on the calendar, nor in a hypothetical personal gazetteer:

“Unfortunately, things didn't turn out better than expected this time. D. broke up with me yesterday evening. (Actually, I had to practically play 20 questions to get him to actually do it.) He wants to be friends. Gross. I had to tell him twice I thought he should leave before he actually got up and left.”

It is the unexpected nature of these events, and the details told in accounts of them, that make the email interesting to browse.

Message characteristics such as length, structure, threadedness, and temporal proximity are apt to be better predictors of value than content semantics; these intrinsic properties are readily available from the header (if it is intact) or from straightforward analysis. Unsurprisingly, in our example corpus, longer messages were more interesting than shorter ones. Messages with many short paragraphs signified newswiness. Longer messages sent after a long period of silence often

flagged an important event. On the other hand, the tail of a longish thread – often consisting of short messages sent close together – usually proved to be of less long term interest. As past studies have shown, the subject field is often uninformative in predicting the value of a message (Whittaker and Sidner, 1996). In this case, many subjects were simply openings (e.g. “wassup?” or “you’re probably in Portland”) and had little to do with the extended content of the message. As Whittaker et al. noted, message replies are a convenient way of re-initiating contact, so a message with a subject line, “Re: Amused about the slammer,” probably has strayed from the original subject. In practice, even though both correspondents were familiar with the other’s email address and used autocomplete, about 2/3 of the messages were written using the “Reply to” feature, often with little concern for whether they were actually replies.

How can we visualize a collection like this as a whole? Work on visualizing email archives has focused on the ebb and flow of relationships among correspondents or social ties (Perer et al., 2005; Donath, 2005) and is more oriented toward characterizing academic collaboration, rather than on this more ruminative or emotive kind of correspondence between two friends. In the concluding section, long-term access techniques for other types of material will be discussed in greater depth.

Despite the fact that email might be considered a very different kind of PIM content than, say, a digital photo collection or an accumulation of personal documents in a file system, the case study demonstrates that email raises many of the same long term issues related to storage, preservation, and access as these other types of personal digital content.

1.4. Looking Forward

Given the seriousness and ubiquity of the issues, several questions come to mind. Is personal digital archiving an insoluble problem? Should we give up and simply regard our digital belongings as transient? Or should we be digital Pollyannas and keep buying more and more storage with the hopes that both storage media and the bitstreams written on them will still be viable and decodable when we want to see them again?

It is clearly more productive to focus on the most important and tractable of the issues now while we can still implement up-front strategies and before too much has been lost. Problems directly related to storage, preservation, and access require technology development; other issues – for example, the balance between content protection using various security mechanisms and the ease of reclaiming the content in individual and collaborative settings – must be worked out socially and legally as well as technologically. Still other issues, such as curation, rely on pragmatic attention to the dovetail between technology and practice. Even if in principle people are happy to take curatorial responsibility for their digital belongings, we have observed that in practice, people don’t apply even the simplest of digital safekeeping strategies and their beliefs about digital media, technologies, and the networked information environment are often contradictory and riddled with misconceptions (Marshall et al. 2006).

In the storage arena, tools for managing distributed content and techniques for stabilizing digital references are two areas for future work that emerge readily from this discussion. The management of distributed content can take advantage of a federated index (Jantz, 2005) rather than fully merging streams into a single database. It is unrealistic to conceive of one's digital belongings as being held in a single store, especially in the intermediate term; we can expect personal information to continue to be distributed among multiple stores including server-based email, digital libraries and archives, trusted personal records repositories held by institutions like banks and the government, and storage-based media sharing services like Flickr.

Similarly, sophisticated preservation techniques – e.g. format registries (Arms and Fleischhauer, 2005), universal virtual machines (Lorie, 2002), format migration services (Hunter et al., 2004), or emulation and decoding services (Heminger and Robertson, 1998) – and digital archiving practices (Hodge, 2000; Smith, 2005) are under development in academic and institutional settings. There is ample reason to believe that these techniques and practices will be applicable to personal digital materials, especially if digital belongings are preserved according to the best practices of specific genres. Records may best be preserved with the specialized techniques developed by records archivists; the consumer imaging industry has developed its own techniques and standards for archiving image collections (for example, see <http://www.everplay-spec.org/>); and web-based material may benefit from lazy preservation (Smith, et al., 2006).

However, it is important to acknowledge that the cost structure of these preservation activities will be different for individuals than they are for institutions; cost/benefit trade-offs must be evaluated. Before we jump into a costly emulation strategy, for example, it would be wise to consider use. Are we emulating a complicated digital application that runs in a particular computing environment so we can render a modified photograph? The emulation will be far more important if we want to further modify the photograph than it will be if we simply want to view it. Many promising preservation strategies involve encapsulating multiple representations of a single digital object (Jantz and Giarlo, 2005). These preservation strategies allow the original version to be retained for reasons of provenance. Furthermore, if any canonicalization or migration steps are loss-y, they can be redone later. Finally, this approach supports emulation if it required, but does not mandate it if it is not.

Access technologies provide us with the most wiggle-room in the immediate future. They can safely be developed over time on top of the other technological solutions, as accumulations of personal digital belongings expand and our understanding of long term use continues to develop. Promising personal information access directions include:

- (1) Automatically generated visualizations that provide us with an overall gestalt of what we have (see Chapter 7);
- (2) Manually defined and circumscribed digital places and geographies that give us the digital equivalent of “the box under the bed” (for the most valued stuff) and “remote storage lockers” (for the things we aren't sure we'll continue to want);
- (3) Heuristics for detecting the relative value of individual items, because people demonstrate the worth of their belongings much more reliably than they declare it; and

- (4) Methods and tools for examining individual items that reveal an item’s provenance and help increase its intelligibility (to oneself and possibly to others), for example allowing a user to distinguish among related copies of an item (Marshall, 2006).

Table 1 summarizes long-term PIM technologies; it is not intended to be a comprehensive set of technologies, but rather it is representative of promising directions.

issue	Item-level technologies	Repository-level technologies	Accumulation-level technologies
Predicting value	Heuristics for assessing item value as a function of demonstrated worth, emotional impact, creative effort, and reconstituteability.	Services that maintain records of value and context; access methods that rank items by value.	Inter-repository services, e.g. tools that help users distinguish among multiple copies of an item.
Distributed storage	Automated replication methods; digital object surrogates that represent content held elsewhere.	Repositories that combine surrogates for distributed objects with local content.	Federated indices; inter-repository communication.
Digital context Protection v. maintenance	Up-front techniques for gathering and storing context for objects and collections.		
Format	Mechanisms to track protection schemes and help consumers recover protected items and collections.		
Curatorial effort	Digital object models that encapsulate multiple representations.	Repository-based automated canonicalization and migration.	Format migration services; Emulation and decoding services; Format registries.
Long-term access	Services similar to Symantec’s Genesis and Microsoft’s OneCare, where much of the curatorial effort is performed remotely on the consumer’s behalf.		
	Retrieval time tools for inspecting an item’s provenance and context	Circumscribed digital places coupled with collection visualization techniques	Exploratory visualization and knowledge discovery to discover patterns

Table 1. Examples of promising technological approaches to address each of the seven issues.

Understanding individuals’ needs and the characteristics of the global information environment will help us develop a viable approach to personal archiving. A strong use perspective can prevent us from taking on the most general – and often the most costly and most difficult – problems. For example, suppose we agree that to preserve email context, we need to save external Web references. Understanding anticipated use will tell us whether we need to preserve a single destination page or crawl the destination site and preserve multiple pages and their interconnections (Lyman et al., 1998). It will also tell us whether the page can be preserved as a static view, where the emphasis is on content and visual appearance, or whether the destination’s full interactivity needs to be maintained (Waldrip-Fruin, 1999; Marshall and Golovchinsky, 2004). Understanding the global information environment will tell us whether we need to cache the page or pages up-front or whether we can rely on lazy preservation methods to reconstruct it later (Smith, et al., 2006).

In 1997, Terry Kuny suggested that we may be entering a digital dark ages due to our lack of a coherent socially-based strategy for addressing the long term issues raised by the transition from physical documents to digital (Kuny, 1998). Not only are our cultural assets at risk; so too are our personal digital belongings. Personal information management must be approached not only with an eye toward the information overload and task management problems of here and now, but also with attention to how our digital belongings will survive into the future.

1.5. References

- Arms, C. (2000). Keeping Memory Alive: Practices for Preserving Digital Content at the National Digital Library Program of the Library of Congress. *RLG DigiNews*, 4(3).
<<http://www.rlg.org/preserv/diginews/diginews4-3.html>>
- Arms, C. and Fleischhauer, C. (2005). Digital Formats: Factors for Sustainability, Functionality, and Quality. *Proceedings of IS&T's Archiving 2005 Conference*. Springfield, VA: Society for Imaging Science and Technology, 222-227.
- Baker, N. (2001). *Double Fold: Libraries and the Assault on Paper*. New York: Random House.
- Beagrie, N. (2003). *National Digital Preservation Initiatives: An Overview of Developments in Australia, France, the Netherlands, and the United Kingdom and of Related International Activity*. Washington, DC: Council on Library and Information Resources.
<<http://www.clir.org/pubs/abstract/pub116abst.html>>
- Beagrie, N. (2005). Plenty of Room at the Bottom? Personal Digital Libraries and Collections. *D-Lib Magazine*, 11(6). <doi:10.1045/june2005-beagrie>
- Bell, G. (2001). A Personal Digital Store. *Communications of the ACM*, 44 (1), 86-91.
- Boardman, R. and Sasse, M.A. (2004). "Stuff Goes into the Computer and Doesn't Come Out": A Cross-tool Study of Personal Information Management. *Proc. CHI'04*. New York: ACM Press, 583-590.
- Bruce, H., Jones, W., Dumais, S. (2004). Information behaviour that keeps found things found. *Information Research*, 10 (1).
- Buckland, M. (2004, March 3-5). *Going Places in the Catalog: Enhancing Scholarly and Educational Resources with Geospatial Information*. Paper presented at WebWise 2004: Sharing Digital Resources, Chicago, Illinois.
- Commission on Preservation and Access and the Research Libraries Group. (1996). *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. Mountain View, CA: Research Libraries Group. <<http://www.rlg.org/ArchTF/tfadi.index.htm>>
- Conway, P. (1996, March). *Preservation in the Digital World*. Washington, DC: Council on Library and Information Resources.
- Cooper, B.F. and Garcia-Molina, H. (2002). Peer-to-peer data trading to preserve information. *ACM Transactions on Information Systems*, 20 (2), 133-170.
- Cutrell, E., Dumais, S.T., and Teevan, J. (2006). Searching to Eliminate Personal Information Management. *Communications of the ACM*, 49 (1), 58-64.
- Czerwinski, M., Gage, D., Gemmell, J., Marshall, C.C., Perez-Quinones, M., Skeels, M., and Catarci, T. (2006). Digital Memories in an Era of Ubiquitous Computing and Abundant Storage. *Communications of the ACM*, 49 (1), 45-50.
- Donadio, R. (2005, September 4) Literary Letters, Lost in Cyberspace. New York: *New York Times*.
- Donath, J. (2004). *Visualizing Email Archives (Draft)*.
<<http://smg.media.mit.edu/papers/Donath/EmailArchives.draft.pdf>>
- Gemmell, J., Bell, G., Lueder, R., Drucker, S. and Wong, C. (2002). MyLifeBits: Fulfilling the Memex Vision. *Proceedings of ACM Multimedia'02*. New York, NY: ACM Press, 235-238.
- Graham A., Garcia-Molina, H., Paepcke, A., and Winograd, T. (2002). Time as Essence for Photo Browsing Through Personal Digital Libraries. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*. New York, NY: ACM Press, 326-335.
- Gray, J. and Shenoy, P. (2000). Rules in Data Engineering. *Proceedings of IEEE International Conference on Data Engineering*. Los Alamitos, CA: IEEE Press, 3-12.

- Hafner, K. (2004, November 10). Even Digital Memories Can Fade. New York: *New York Times*.
- Hart, P. and Liu, Z. (2003). Trust in the Preservation of Digital Information. *Communications of the ACM*, 46 (6), 93-97.
- Hedstrom, M. and Montgomery, S. (1998). *Digital Preservation Needs and Requirements in RLG Member Institutions*. Mountain View, CA: Research Libraries Group.
- Heminger, A.R. and Robertson, S.B. (1998, November 21). *Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents*.
<http://tuvok.au.af.mil/au/database/research/ay1996/afit_la/rober_sb.htm>
- Hodge, G.M. (2000). Best Practices for Digital Archiving: An Information Life Cycle Approach. *D-Lib Magazine*, 6 (1).
- Hunter, J. and Choudhury, S. (2004). A Semi-Automated Digital Preservation System based on Semantic Web Services. *Proceedings of JCDL'04*. New York, NY: ACM Press, 269-278.
- Jantz, R. (2005). Digital Preservation: Enabling Technologies for Trusted Digital Repositories. *D-Lib Magazine*, 11 (6). <doi:10.1045/june2005-jantz>
- Jantz, R. and Giarlo, M.J. (2005). Digital preservation: Architecture and technology for trusted digital repositories. *D-Lib Magazine* 11 (6). <doi:10.1045/june2005-jantz>.
- Jones, W. (2004, March 3). Finders, keepers? The present and future perfect in support of personal information management. *First Monday* 9 (3).
<http://www.firstmonday.org/issues/issue9_3/jones/>
- Kuny, T. (1998). The Digital Dark Ages? Challenges In The Preservation Of Electronic Information. *International Preservation News*, 17 (5). <<http://www.ifla.org/VI/4/news/17-98.html>>
- Lavoie, B. and Dempsey, L. (2004). Thirteen Ways of Looking at...Digital Preservation. *D-Lib Magazine* 10 (7/8). <doi:10.1045/july2004-lavoie>.
- LeFurgy, W.G. (2003, December 15). PDF/A: Developing a File Format for Long-Term Preservation. *RLG DigiNews* 7 (6).
- Levy, D.M. (1998). Heroic measures: reflections on the possibility and purpose of digital preservation. *Proceedings of Digital Libraries '98*. New York, NY: ACM Press, 152 – 161.
- Lorie, R. (2002). A Methodology and System for Preserving Digital Data. *Proceedings of JCDL'02*. New York, NY: ACM Press, 312-319.
- Lyman, P. (2002, April). Archiving the World Wide Web. In *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*. Washington, DC: Council on Library and Information Resources, 38-51.
- Lyman, P. and Kahle, B. (1998). Archiving Digital Cultural Artifacts: Organizing an Agenda for Action. *D-Lib Magazine* 4 (7). <<http://www.dlib.org/dlib/july98/07lyman.html>>
- Lynch, C. (1999). Canonicalization: A fundamental tool to facilitate preservation and management of digital information. *D-Lib Magazine*, 5 (9). <doi:10.1045/september99-lynch>
- Marshall, C.C. (2006, August 10). *Why a corpus-topics-relevance judgments framework isn't enough: two simple retrieval challenges from the field*. Paper presented at the SIGIR 2006 Workshop on Evaluating Exploratory Search Systems, Seattle, Washington.
- Marshall, C.C. and Bly, S. (2005). Saving and Using Encountered Information: Implications for Electronic Periodicals. *Proceedings of CHI'05*. New York: ACM Press, 111-120.
<<http://www.csd.tamu.edu/~marshall/p440-marshall.pdf>>
- Marshall, C.C., Bly, S., and Brun-Cottan, F. (2006). The Long Term Fate of Our Personal Digital Belongings: Toward a Service Model for Personal Archives. *Proceedings of IS&T's Archiving*

- 2006 Conference. Springfield, VA: Society for Imaging Science and Technology, 25-30.
<<http://www.csdl.tamu.edu/~marshall/archiving2006-marshall.pdf>>
- Marshall, C.C. and Golovchinsky, G. (2004). Saving Private Hypertext: Requirements and pragmatic dimensions for preservation. *Proceedings of ACM Hypertext 2004*. New York, NY: ACM Press, 130-138.
- Marshall, C.C. and Jones, W. Keeping Encountered Information. (2006). *Communications of the ACM*, 49 (1), 66-67.
- Oltmans, E., van Diessen, R., and van Wijngaarden, H. Preservation Functionality in a Digital Archive. (2004). *Proceedings of JCDL'04*. New York, NY: ACM Press, 279-286.
- Perer, A., Shneiderman, B., and Oard, D.W. (2005, June 3). *Using Rhythms of Relationships to Understand Email Archives*. Paper presented at the 22nd Annual Symposium of the Human-Computer Interaction Laboratory, University of Maryland, College Park, MD.
- Reichherzer, T. and Brown, G. (2006). Quantifying Software Requirements for Supporting Archived Office Documents using Emulation. *Proceedings of JCDL'06*. New York, NY: ACM Press, 86-94.
- Rothenberg, J. (1995). Ensuring the Longevity of Digital Documents. *Scientific American*, 272 (1), 42-47.
- Rothenberg, J. (1998). *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Washington, DC: Council on Library and Information Resources.
- Segal, R. and Kephart, J. (1999, May 1-5). *MailCat: An Intelligent Assistant for Organizing E-Mail*. Paper presented at the Third International Conference on Autonomous Agents, Seattle, WA.
- Smith, J.A., McCown, F., and Nelson, M.L. (2006) Observed Web Robot Behavior on Decaying Web Subsites. *D-Lib Magazine*, 12 (2). <doi:10.1045/february2006-smith>
- Smith, M. (2005, July). Eternal Bits: How Can We Preserve Digital Files and Save Our Collective Memory? *IEEE Spectrum*, 22-27.
- Stefik, Mark. (1999). *The Internet Edge: Social, Technical, and Legal Challenges for a Networked World*. Cambridge, MA: MIT Press.
- Tansley, R., Bass, M., Stuve, D., Branchofsky, M., Chudnov, D., McClellan, G., and Smith, M. (2003). The DSpace Institutional Digital Repository System: Current Functionality. *Proceedings of JCDL'03*. New York, NY: ACM Press, 87-97.
- Waldrip-Fruin, N. (1999). Hypermedia, eternal life, and the impermanence agent. *SIGGRAPH 99 Electronic Art and Animation Catalog*. New York, NY: ACM Press, p. 90.
- Whittaker, S., Bellotti, V., and Gwizdka, J. (2006). Email in Personal Information Management. *Communications of the ACM*, 49 (1), 68-73.
- Whittaker, S. and Sidner, C. (1996). Email Overload: Exploring Personal Information Management of Email. *Proceedings of the CHI'96*, New York, NY: ACM Press, 276-283.
- Wood, D. N. (1984). The collection, bibliographic control and accessibility of grey literature. *IFLA Journal* 10 (3), 278-282.