Saving Private Hypertext: Requirements and Pragmatic Dimensions for Preservation

Catherine C. Marshall
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
USA
cathymar@microsoft.com

Gene Golovchinsky
FX Palo Alto Laboratory, Inc.
3400 Hillview Avenue, Bldg. 4
Palo Alto, CA 94304
USA
gene@fxpal.com

ABSTRACT
The preservation of literary hypertexts presents significant challenges if we are to ensure continued access to them as the underlying technology changes. Not only does such an effort involve standard digital preservation problems of representing and refreshing metadata, any constituent media types, and structure; hypertext preservation poses additional dimensions that arise from the work's on-screen appearance, its interactive behavior, and the ways a reader's interaction with the work is recorded. In this paper, we describe aspects of preservation introduced by literary hypertexts such as the need to reproduce their modes of interactivity and their means of capturing and using records of reading. We then suggest strategies for addressing the pragmatic dimensions of hypertext preservation and discuss their status within existing digital preservation schemes. Finally, we examine the possible roles various stakeholders within and outside of the hypertext community might assume, including several social and legal issues that stem from preservation.

Categories and Subject Descriptors
H.5.4 [**Hypertext/Hypermedia**]: *Architectures, Theory, User Issues*

General Terms
Documentation, Human Factors, Standardization, Theory

Keywords
Hypertext, Archiving, Digital preservation.

1 INTRODUCTION
We begin this paper with two cautionary tales to motivate our discussion.

> My [Cathy's] copy of *Uncle Buddy* is still intact in the original box: the booklet, the cassette tapes, and five diskettes. The software is for the Mac, a platform I used when I got this pioneering hypermedia novel. I don't have a Mac anymore; I donated it to the school system about five years ago. I imagine that it's so obsolete by now that it has been recycled to reclaim the precious metals and process the toxic elements. I don't even have a floppy disk drive or cassette player any more. Would the HyperCard-based software run today if I had access to a Mac with a floppy drive? I am told it would, but I have no way to verify this. In fact, it seems that I have no way at hand to read *Uncle Buddy*, although my copy appears to be in good condition. The same is true for my first-edition copy of *Intergrams*. My copy of *Afternoon* no longer has the packaging – I'm no longer sure which platform will read the floppy disk.
>
> In 1995, Judy Malloy and I [Cathy] designed a Web version of *Forward Anywhere*. The user interface and screen design were very simple; the screen design was intended to capture the look of a cathode ray terminal circa 1980 and the user interface was intended to reflect the process used to create the piece. In 2003, Judy received an email from a reader who told her that the "lines" function no longer worked. She forwarded me the message, and I debugged the C code – a server name had changed – and recompiled it after several false starts trying to remember the Unix command line parameters for the compiler and how libraries were linked. I was lucky the bug was simple; I had little recollection of how parts of the code I wrote ten years ago worked.

As we accumulate a significant number of digital artifacts and experience inadvertent losses through technological changes, platform upgrades, and media degradation, our anxiety about preservation grows [22]. Will it be necessary to sneak into a computer museum in the dark of night to read an old email message or a classic work of hypertext fiction? Are we so optimistic that we believe our digital photos will capture a lifetime of memories, accessible on demand in fifty years?

This problem has received greater attention with the advent of significant digitization projects and serious digital library efforts (for example, [2]). In fact, this concern prompted a study by the Commission on Preservation and Access and the Research Libraries Group [6]; various research labs in the commercial sector (for example, IBM, Ricoh, and HP in conjunction with MIT) also have preservation and archiving technology development efforts. Furthermore, monumental projects like Brewster Kahle's Internet Archive or Rick Prelinger's archive of film footage demonstrate that there has been plenty of time and resources invested in on-the-ground problems associated with digital archiving.

The Research Libraries Group's 1996 report described three strategies for digital archiving at the institutional level: (1) refreshing, a process that literally copies the digital objects to be preserved; (2) emulation, which provides a means of reproducing the technological context in which the digital object was

HT'04, August 9–13, 2004, Santa Cruz, California, USA.

originally created; and (3) migration, a means of recreating the digital object within the current technological context [6]. These strategies have been further examined in [22] and [20].

Each of these strategies has its drawbacks. Refreshing addresses the degradation and obsolescence of media and physical storage, but does not take into account the technological context. Thus any substantial change in platform or application format will not be addressed by such a strategy. Emulation takes into account the changes in platform and applications by attempting to preserve them too in some sense. However, it is clear that this strategy is impractical (and insufficiently general) for our purposes, even on an institutional scale: the number of hardware and software combinations that have been used and the range of customized readers that have been developed to distribute literary hypertexts makes the cost prohibitive. The third strategy, migration, while being the most practical and general of the solutions, still doesn't really address the full spectrum of the problem since migration at best may lose vital aspects of the document's form even while preserving its content (which is the flip side of the problem that PDF most specifically addresses).

Other strategies have been proposed as well. The DSpace project, the result of collaboration between HP and the MIT Libraries, is implementing an open source system so that institutions and organizations can maintain their own digital repositories for long-term document storage and preservation [36]. The system's functionality satisfies many research and educational preservation requirements and, as an open source effort, could be extended to meet many of the requirements we identify in this paper for preserving and archiving literary hypertexts.

Alternatively, a recent article in the *Communications of the ACM* proposes a banking-based model of preservation in which an institutional service stores copies of digital materials [17]. There are several problems apparent with this strategy for preserving literary hypertexts. Even given the high level of trust we have in our financial institutions as third-party intermediaries, we are asking them, in essence, to protect and store an abstraction, the amount of money entrusted to them. An equivalent strategy for digital media does not address content destruction; nor does it really tackle the problems identified in the RLG report (except perhaps the problems solved by refreshing).

Hypertext fiction occupies a provocative niche in defining requirements and testing solutions for the immense problem of digital archiving (see, for example, the Electronic Literature Organization's PAD initiative[1]). Not only is hypertext fiction a literary effort, it may also represent a software development effort, a sophisticated and often unconventional use of different kinds of digital media, a visual design component, and an exercise in interaction design that may even involve special types of platforms and hardware [14].

Because computer games share important properties with hypertext fiction [1] they may also share many of the same preservation problems, although games tend to be subject to less attention by digital archivists, and more by aficionados of the individual games or gaming platforms. One might envision specialized museum-like efforts to preserve the games with their platforms (akin to the effort of saving unique mechanical arcade games of San Francisco's Playland at the Beach [9]), but it is more likely in everyday gaming situations they will be replaced by games more suitable to current technological and social climate. By contrast, one might envision older works of hypertext fiction to continue to be read alongside newer works; thus hypertext preservation efforts must also be oriented toward access.

Given that there are relevant technological approaches to preserving hypertexts and that there are various related efforts afoot, it is vital to address three questions:

- What aspects of hypertext fiction is it essential to preserve? Is it only what can be captured by a universal format that adequately represents a node-link structure and some aspects of its appearance, a "gold standard" [11] or canonicalization [26] for format? This may be difficult given the variety of standards that have been used to fully represent the content, structure, and presentational aspects of a hypertext, such as Dexter, OHM, or simply the standards used on the Web such as RDF and XML. Furthermore, many important hypertexts cannot be expressed with any such formalism. Or should the effort be more museum-like in its approach to saving aspects of context and use?

- What is an appropriate strategy (or strategies) for carrying out preservation and archiving for hypertext fiction? Can we identify a multi-tiered strategy that is economically viable while still satisfying most of the constraints of good preservation practices?

- Who is responsible for the preservation and archiving of hypertext fiction? If it isn't the author, who decides what is preserved and what is lost? If it isn't the publisher, how will preservation efforts interact with current and future copyright and DRM restrictions? If it isn't the reader, how will records of reader interaction – seen as a fundamental precept of hypertext fiction [30] – be saved for future readings?

In this paper, we will explore all three areas of preservation and archiving given the interesting set of requirements posed by hypertext fiction.

2 WHAT'S IN A HYPERTEXT?

For ordinary digital documents, preservation usually involves capturing the document's content (in terms of the literal text and other embedded media, possibly with semantic tags), its structure (in the sense of functional tags), and some notion of its appearance (how it is rendered on the screen). This representation is then preserved using a relatively stable medium, such as CDROM, as many distributed copies, such as the strategy employed by LOCKSS [7], or using a repository-based strategy such as DSpace [36].

Fully preserving the characteristics of literary hypertexts promises to be more of a representational challenge, not in least dictated by the more prominent role of the reader. Not only must the usual document properties be preserved, but also the characteristics that support the reader's role. At best, this includes:

- The ability to produce all readings that were originally possible;

[1] http://www.eliterature.org/pad/.

- The ability to interact with the work in ways that were supported by the original;
- The machine time implicit in the original work, if appropriate;
- The embodiment of the work's design in the assumed reading environment; and
- Records of reading that capture intangible and tangible interactions, such as history and annotations;

We look briefly at the entailments of each of these preservation requirements, knowing that each contributes to the overall complexity of the problem. Dimension-based solutions to each are also mentioned; later they are covered in greater detail when we discuss the pragmatics of preservation. By looking at these requirements, we take another perspective on the usual emulation or migration strategies that attempt to replicate the software itself; by using these requirements, we may arrive at a different set of partial solutions based on individual requirements or different ways to capture the spirit of the original work.

The ability to produce all readings that were originally possible. From guard fields [3] to Petri net-based hypertexts [35] to hypertexts created on-the-fly [13], many literary hypertexts have relied on a sophisticated notion of which readings are possible within the bounds of unrestricted traversal. A fully preserved hypertext permits all possible readings and precludes readings that were not possible in the original.

The ability to interact with the work in the original way. We may take for granted the mode of interaction used by a literary hypertext. For example, *Intergrams* relies on the reader's ability to hover over a many-layered, partially translucent spatial work to examine a poem's structure; it is a mouse-based interaction that might not be readily replicated given a different input device. While we may have a difficult time imagining a different input device, others have been available in the past, and still others may be developed in the future, as for example the squeezing and tilting interaction described in [16]. The annotation interface to *Forward Anywhere* relies on a reader's creation of freeform digital ink marks using a pen tablet [14]; the mouse does not invite similar interaction.

The machine time implicit in the original work. In her 1998 paper, Luesebrink ventures that, "While readers are not normally conscious of access time in print literature, electronic texts are always pursued in an environment subject to the vagaries of computer speed and software performance." [25] The simple effect of hardware, software, and network performance cannot be neglected; anyone who has run game software on an upgraded platform can vouch for the fact that there are cases in which the unintended speed-up significantly alters the experience of the game. Thus, timing factors in animations and basic interaction should be preserved with the work.

The embodiment of the work's design in the assumed reading environment. While most preservation efforts have taken into account how an electronic document is rendered on the screen, literary hypertexts seem to be even more sensitive to taken-for-granted properties of particular displays, such as the limited screen size of the original Macs, or properties of fonts that may have changed over time. It is surprising, in fact, how text layout may radically change the perception of content as Marshall and Ruotolo saw in their field studies of Microsoft Reader [27]. In this field study, rhyming poetry was reduced to doggerel simply because the rhyming words of the couplets tended to fall on new lines, which caused them to be displayed in isolation. Similarly, early literary hypertexts were often written for the smaller displays available for the Mac; when we see recreations of them now on our typical larger, higher resolution displays, the works are either confined to a screen area that seems too small, or the amount of the text displayed in a browser window is far more than the author intended the reader to see at once, as illustrated in Figure 1 and Figure 2.

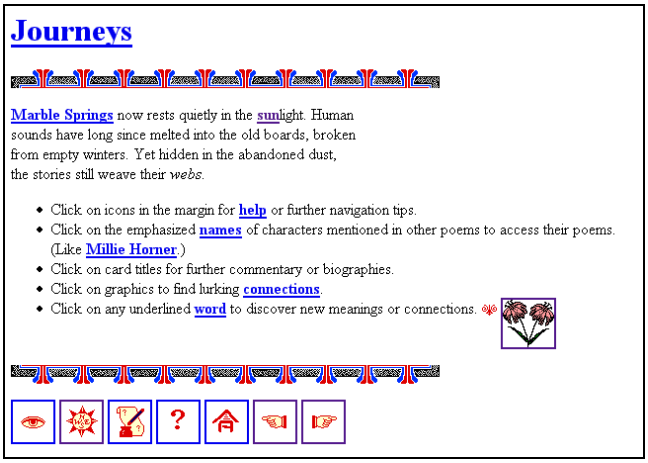

Figure 1. Intended appearance of a *Marble Springs* node, as converted to HTML.

Figure 1 shows a web page of *Marble Springs* [21] as it was intended to be seen in a web browser. This web page was part of an electronic excerpt from the complete work, originally published as a HyperCard stack by Eastgate Systems. Figure 2 shows the same page as it was originally displayed on our laptop computer, allowing the browser window to assume the shape specified by system defaults. While the text remains the same, the unintentionally large window makes the links difficult to understand: "sunlight" near the top of the page links to "*Sun*:" at the bottom. No visual change occurs when the link is traversed, and a reader unfamiliar with the work may be left wondering if the system is broken.

Records of reading that capture intangible and tangible interactions. One of the more interesting developments in literary (and other) hypertexts is that they have been designed to record a reader's detailed interaction with them, thus creating a specific history that can be used to reconstruct the reading or can form the basis of further interaction (that is, the reader can interact with his or her history as well as with the original work, as in [33]). Thus, for an individual reader, a preservation effort might include some means of saving and having continued access to these records. In general, preservation should include the means to add further records of reading, as was possible in the original.

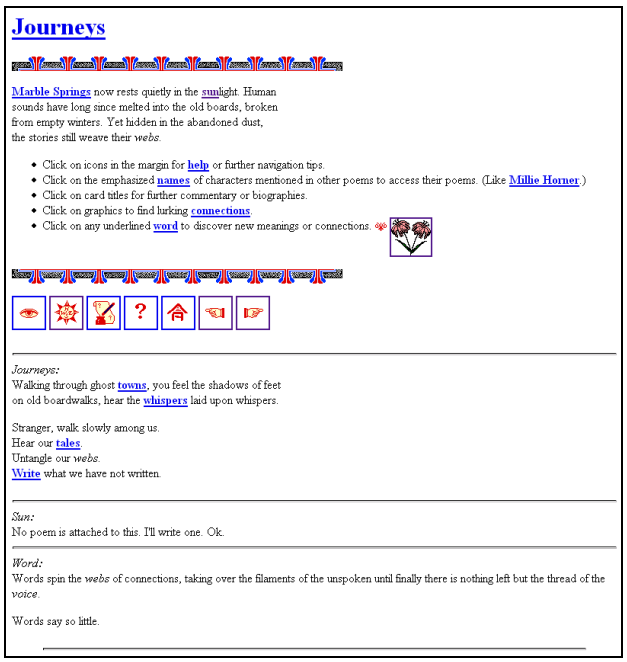

Figure 2. Same web page, in default browser window.

Tangible records of reading include elements like a reader's annotations on the work. If these are scholarly annotations, a preservation strategy might even include them with future editions or versions of the work; if they are personal annotations, they set forth the problem of the sort described above: the reader would like to retain them with the work, regardless of their subsequent value. These records help a reader form a geography of his or her personal digital library and thus should not be ignored by preservation strategies [28].

3 PRAGMATICS OF PRESERVATION

We can preserve every aspect enumerated above only if we preserve the whole work: the media, the interface, the OS, the platform, etc. (We note, however, that few preservation efforts through history have enjoyed the luxury of completely representing all aspects of the work in context.) Once we enter the realm of migration, however, we must choose what to save and how to save it. This is an editing process, not unlike those that preserved the literary works of the ancient Greeks: the written plays differed from the performances, and were further modified when copies were made (manually) throughout their existence.

So what can we save? We can consider preservation along six dimensions: metadata, media, structure, appearance, behavior, and user data. At one extreme of the space, we have just the cataloging record; at the other the work itself. Between, we can imagine a continuum of approximations that capture the media, the structure, the appearance, and the behavior of the original. As the examples that follow illustrate, these dimensions or aspects of a hypertext may be preserved independently, and will be suitable for different kinds of reading and use by future readers.

3.1 Preserving the parts

The dimensions enumerated above vary in the degree to which they are amenable to preservation: some are straightforward, others have not even been considered worthy of preservation. In Table 1, we summarize the set of dimensions in terms of associated preservation efforts, and illustrate them with some familiar examples. The dimensions may also be considered with respect to ease of archiving (see Figure 3). It is relatively easy to preserve the bits that characterize each component; it is harder subsequently to use these bits to reconstruct some dimensions than others. It is interesting to note that although in principle user data is easy to store – these are typically data files – it is often overlooked as an important component of an *interactive* work. Furthermore, it may be difficult to recreate the effect of user data on behavior in some hypertexts.. This is especially (and necessarily) true in the case of institutional repository schemes; they are not geared to personal preservation.

We assume the presence of a bibliographic record of the work. Not only would we not be able to refer to a work without it, but also this is the most straightforward dimension to preserve. Standards (e.g., MARC, Dublin Core, etc.) are available to record the metadata, and preservation efforts in this domain are well underway. Despite the seemingly straightforward nature of bibliographic metadata, it is not always simple choose the appropriate standard; nor is it a given that their constituent fields can be readily mapped to one another [27].

Since hypertexts often include text, images, and other media as lexia, these can be preserved independently of the work, and a variety of ongoing preservation efforts based on refreshing (when

Table 1. Dimensions of hypertext preservation

Dimension	Example	Status
Metadata	MARC, Dublin Core, RDF, ISBN	Preservation efforts underway; multiple formats
Media	Images, text, etc.	Preservation through refreshing and standard format choices
Structure	May be implemented as standard node-link graphs; may include constraints such as guard fields; may be realized as translucence and simultaneities (see [32])	Basic node-link structure can be represented in OHM, for example. Transformations to other representations are possible. Problems arise when structure is entangled with behavior; otherwise tractable. No canonical representation for spatial hypertext
Appearance	Font metrics, display color characteristics, screen resolution, default window sizes	There is no canonical way to specify appearance. Color and fonts are platform-specific.
Behavior	Fluid Hypertext [38]	Preservation is not feasible without preserving important aspects of the executable and the platform, either through emulation or migration (re-implementation).
Reader data	Annotations, bookmarks, state, history, user settings	These kinds of personal data are frequently entirely overlooked when considering software and documents.

used in conjunction with well-established data formats) should ensure the longevity of these components of hypertext works. Preservation of the media alone may have some value as well: Reviewing these components could give the future reader some sense for what was there, without revealing how the materials were related, displayed, or used. Some readers have taken this approach to "reading" hypertext even in the presence of the whole hypertext to gain insight into the work's scope [10].

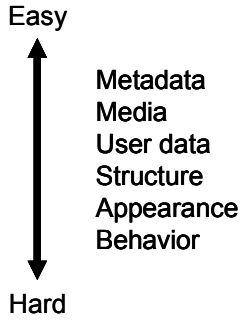

Figure 3. Dimensions of preservation organized by increasing difficulty. Note that user data is often overlooked.

By structure, we mean the traditional graph that describes the connectivity among nodes of a hypertext. Although simple embedded hypertext representations such as HTML do not make structure explicit, it can nonetheless be extracted and preserved [15]. Preserving structure alone would give a future reader a sense for the complexity of the work without revealing the details. A table of contents is one such structural representation.

Appearance is often important to a hypertext. Spatial hypertexts are an obvious example, but even primarily textual works may use juxtaposition and layout to achieve specific effects. We sometimes become aware of inadequate representations of appearance when a piece of software run in an emulator does not look quite right or translation (migration of content and structure alone) has resulted in an altered visual representation due to functional discrepancies between host systems. An example of this phenomenon was aptly illustrated by translations of NoteCards data files (called NoteFiles) to HyperCard using Bornstein and Riley's hypertext translation facility [4]. Having a few screenshots of a system in use can often give a more accurate depiction of the work than prose describing its various features. It is important to note that such seemingly trivial aspects as font and color are often not preserved across different platforms. In the second author's experience, a difference between Windows 98 and Windows NT fonts caused one layout-sensitive program to crash due to different pagination breaks in the two systems.

Some hypertexts rely heavily on interaction, expressed as reader-initiated animation, stateful navigation, or other forms of manipulation. In such cases, system behavior often becomes the most important characteristic of the work, the quality that sets it apart. Unfortunately, interactivity is often the most difficult aspect to capture as it relies on computation that may be difficult to replicate without access to the original software.

Finally, some dynamic hypertexts use interaction histories, bookmarks, annotations, and various user profile data to influence reading sequences. This sort of user profile data typically goes unnoticed – often it is conspicuous only after its loss, such as when software is reinstalled on a new computer, or recovered from backup. When considering preservation schemes in which readers can retain personalized information, particular attention must be given to how and where hypertexts store such data.

Of course, some literary works deliberately defy preservation efforts. Individual copies are meant to have a finite lifespan (as in Gibson's *Agrippa* [12], which self-destructs) or a limited readership (as in Jackson's *Ineradicable Stain* [19], which is composed entirely of single word tattoos on several thousand people; only the tattooed participants are permitted to read the entire work).

3.2 Preserving the whole

The previous section offered a catalog of parts-specific preservation, in the hope that the collection of parts can be used to reconstitute the whole. In some cases, unfortunately, it may be too expensive or otherwise infeasible to preserve the original hypertext. Yet since hypertexts are inherently about interaction, we should consider ways of capturing that even if we lose some of the details of the work. The following is one speculative idea for capturing the essence of a hypertext instead of its components.

A video of someone using a system can reveal much of its spirit, even if it doesn't allow a future reader to experience the whole work. It does, however, lock us into the narrow view of the system as demonstrated for the camera. We are all familiar with the film showing Douglas Englebart using Augment, but we know little of the system beyond what was shown. While more rewarding than a static picture or a textual description, the video gives an impoverished sense of what it was like to use the system.

We can, of course, record on video a hypertext as it is read. Unfortunately, that will yield a linearization of the work that lacks alternatives that are essential to hypertext, or the alternatives will be traversed in the same order each time we play the recording. One possible way of capturing multiple paths may be drawn by analogy with the physical world. Several projects have using a variety of techniques to capture three-dimensional representations of city streets (e.g., [23], [39]) by recording scenes with moving cameras. The computer synthesizes the multiple feeds and yields a navigable structure, permitting the user to seemingly fly (drive) through the streets in an arbitrary manner.

We can imagine capturing some kinds of hypertext interaction in short movies (optimistically produced in a ubiquitously used digital video format) or Flash animations, and then linking them according to the original hypertext structure, possibly in a hypervideo system such as HyperHitchcock [34]. Although these hypertext fly-throughs may lack some of the interactivity, they may give a more dynamic recording of the experience of using the system. A consistent recording method decreases the costs of maintaining the playback system, and the animations preserve the visual appearance (but probably not the feel) of the original.

There are losses, however: if a node is not accessed during the recording phase, it is not available for playback, particularly if point-to-point links are the only interaction style available in the recording medium. This situation is not uncommon: Mark Bernstein notes that 16 of 28 hypertext titles that he published had Jane's spaces – nodes inaccessible through links.[2] A media-centered approach to archiving would capture these hidden nodes, whereas a recording would not.

[2] http://markBernstein.org/December02.html#note_3200

Recording also fixes the navigation sequence, and may thereby constrain the scope of the work available for playback, particularly if an initial selection forecloses on certain possibilities. Authors of such works might consider creating different recordings that explore mutually-exclusive paths, not unlike alternate movie endings or the male and female versions of the *Dictionary of the Khazars* [31].

Certainly we should note that in these cases, we have simply deferred the problem, moving from one digital platform to another. But in the event that emulation or migration is difficult or costly, a careful choice of a recording format makes it more likely that some of the more ephemeral aspects of the work's appearance and interactive behavior will survive.

4 RESPONSIBILITY FOR PRESERVATION AND ACCESS

When we talk about preservation and archiving of literary hypertexts, the question that cannot be ignored is who will bear the responsibility for carrying out any of the strategies we set forth in this paper. Will it be the reader, who now owns the work in question, and may want to read it again in the future? Will it be the author, who understands the intent of the work and who has much vested interest in the work's stability and continued reader access to the work? Will it be a traditional institution like a university library or museum whose mission includes preservation and access, and whose expertise is well aligned with the problems presented by literary hypertexts? Will it be an independent philanthropic organization charged specifically with this job? Will it be the works' publishers, who have a financial interest in their continued availability? Or is it destined to be a for-pay service, used by whomever has sufficient interest and resources.

4.1 Stakeholder roles

We explore each possible stakeholder in turn, since each suggests different issues and uncovers different assumptions. Special attention is given to author-driven preservation, since choices that the author makes during the work's creation will bear on its eventual sustainability.

Reader-driven preservation. Many of us don't realize it yet, but if we are to access, read, or even view our digital materials in the future, we currently shoulder an implicit burden of keeping them up-to-date with platforms, software, and standards. No-one is keeping an eye on us to ensure we're handling our digital photos appropriately to be able to look at them when we're ninety; if the jpeg format were to be superceded by another data representation or if a media type – for example, the CDROM – were to fall into disuse, it would be up to us to anticipate transitions and migrate our digital photos to the new storage medium. Furthermore storage media like CDROMs and DVDs do not have the same lifespan in the hands of consumers as they do in a setting geared for preservation. Whereas an organization concerned with preservation may implement processes that safeguard the media, most consumers may not have the capability to address (or even awareness of) this problem until it is too late. Similarly, it can be argued that the hypertexts we have bought and may want to re-read are our own responsibility, just as it is our responsibility to keep our important paper books dry and safe from silverfish, book lice and book worms.

More difficult still is the question of preserving the outcomes of an individual's readings. If a migration strategy has been implemented by another stakeholder in the hypertext ecology (for example, the work's publisher), it will still be up to readers to preserve the data they themselves have generated – annotations, bookmarks, paths, and so on. As we have noted, without attention to this detail, it is difficult for a reader to know what to do to migrate personalization and recorded history to an updated work.

Author-driven preservation. Because many literary hypertexts are experimental and "difficult", their readership is limited at best. Authors are anxious to promote the widest possible dissemination of these works by keeping them evergreen, accessible within today's reading environments. This amounts to a preservation effort, usually involving migration. We have already observed authors undertaking such efforts, sometimes moving hypertexts from outdated authoring/reading environments like HyperCard to the Web so they remain accessible. Certainly in the case of literary hypertexts that require a substantial amount of custom software (e.g. Rosenberg's *Intergrams* [32] or Waldrip-Fruin's *Impermanence Agent* [37]) or in which the code is an explicit part of the work (as Cayley discusses in [5]), it almost seems appropriate that the author bears some responsibility for preservation and archiving. In these instances, the interactivity that is implemented through the code, or the code itself, is vital to the reading experience.

Thus authors' decisions about writing and programming environments used to create their literary works have implications for preservation. Our dissection of the preservation effort suggests that several strategies – practiced early, practiced often – may produce hypertexts that are more readily archived: more easily converted into new representations, more robust in the presence of change in the platforms on which they run, or more suitable to run in emulators. This list is by no means exhaustive, and authors should not infer from it that we advocate uniformity of expression. Rather, there are tradeoffs to be made of which authors should be aware.

Virtual machines

Hypertexts that rely on embedded code to implement behavior (rather than on an external program such as a web browser) are more likely to be available in the future if the code runs in a virtual machine (e.g., Flash animations, Smalltalk, Java, etc.) than if they are written as platform-specific executables. One direct advantage is that a hypertext, once developed, is more likely to run on a variety of platforms. Another is that the larger the installed base of the virtual machines, the less likely that programs written for those virtual machines will become obsolete. Of course there are no guarantees: ten years ago, it seemed perfectly natural to create HyperCard hypertexts. Only five years later the practice seemed almost completely abandoned, the software only runs in the "classic" emulation mode on the new OS, and while the viewer is still available for download on the Apple web site, the authoring software is no longer supported. The process of organizational loss has begun.

External link representations

Although there are dissenting voices, external link representation is a well-accepted aspect of many hypertext models. External link representations can be checked for errors, which may save many applications from breaking in ugly ways. They also make it more likely that this important structural aspect of a work will be preserved correctly and that media that make up the hypertext nodes can be archived through existing means.

Recordings of interaction
Video should not be overlooked as a preservation format. Make a recording of a session with your hypertext, and you are likely to see it played back in ten years on a cell phone. The odds of the system running on a readily-accessible machine are much lower.

Server-side computation
The immediacy and ubiquity of distribution is one of the lures of Web-based hypertext development. If the hypertext is essentially static, it is amenable to archiving using the component-based strategies discussed above. If, however, the hypertext includes a significant server-side computational component (whether CGI, ASP, JSP, PHP – the list is ever expanding), the author must take specific precautions to ensure the integrity of the piece. In addition to the standard problems of code portability, authors should strive to remove dependencies on the particular server of the original deployment. One possible strategy to test the archivability of a work is to try to republish it on a different machine while it is still in active use.

Institutional preservation. During the dark ages, preservation of manuscripts was the domain of monks, and later, of rich patrons: a wealthy nobleman who had fifty copies of his favorite play made for his friends around the Mediterranean greatly increased the chances of that work's survival into the present. More recently, institutions like libraries and museums have been the locus for preservation of physical artifacts; they have assumed this role in the digital world as well, although not without difficulties since digital preservation is a complex problem and unfortunately resources for such efforts have been thin at best.

Some libraries such as Alderman Library at the University of Virginia, the New York Public Library, and Harvard University Library have collections of literary hypertexts that appear in their online catalogs as electronic resources. Thus libraries may have full digital metadata describing the works, but the works themselves are available as physical media to be checked out in much the same way as one would check out a book. To date, we know of no specific preservation efforts at these libraries; the works are still in circulation, and ultimately a preservation effort would require more than refreshing or replacing the storage medium. Furthermore we should note that most of the stakeholders we discuss in this section have no particular provision for preserving readers' data unless this data is considered part of a special collection (for example, the personal library of a well-known person).

Museums do preserve computers and the software that runs on them (see, for example, the Computer History Museum)[3]; as such, they could realistically be expected to preserve literary hypertexts along with the platform on which they run. If museums were apt to take on this responsibility, we could expect to at least have literary hypertexts remain part of the cultural record, although access would be limited to a particular place. It is ironic that our early electronic hypertexts would have to be handled analogously to the early hand-coded(!) books.

Publisher-driven preservation. As long as a work continues to be in demand, we might expect publishers to perform migration-style preservation. That is, the work may be re-implemented on viable hardware and software platforms. Like authors, publishers are the most likely to have the motivation and to hold the appropriate copyrights to perform such a translation. However, migration is expensive and publishers must be expected to perform a cost/benefit analysis to determine whether they can and will be responsible for supporting continued access to a literary hypertext.

Philanthropic organizations as the agents for preservation. Recently, the Electronic Literature Organization has undertaken a specific effort (called PAD), that states as its mission:

> The Preservation, Archiving, and Dissemination (PAD) project seeks to identify threatened and endangered electronic literature and to maintain accessibility, encourage stability, and ensure availability of electronic works for readers, institutions, and scholars.[4]

This mission seems to be in accord with standard (and laudable) archiving goals, but we still must ask who does the work of preservation, for it would seem to be a formidable task in the case of valuable literary hypertexts that don't run on standard hardware and software platforms, or that use existing platforms in unusual ways, at the very fringe of their capabilities. If we look at concrete plans for representing the platform, software, and content of digital data like Lorie's Universal Virtual Computer [24], it is easy to see how preservation and archiving can turn into an expensive process.

Furthermore, even if the code and media of a particular hypertext are encoded in a universal encoding, that is no guarantee that the hypertext could be reconstituted later. Code relies on libraries, libraries on other libraries and on operating systems. It is impossible (and arguably meaningless) to encode an entire OS to preserve a work of literature, and there is no guarantee that the virtual machine would execute the code correctly.

What of the Internet Archive? Won't that strategy at least capture the most straightforward of electronic writing (in particular, node-link hypertexts that only rely on embedded code to implement any specific behaviors for the piece)? It seems that it will, at least for now, but it will not preserve Web-based literary hypertexts that use scripts to generate web pages.

Preservation as a service. It is not inconceivable that preservation and archiving will join other document services offered either on a for-pay basis, or for as an adjunct to library services (see [18]). However, in this case, the devil is in the details: will these services be able to handle many of the literary hypertexts? Who will make the inevitable authoring decisions involved in converting between formats? Who will own the copyright on these migrations?

4.2 The perfect archivist

From examining these possibilities, we can identify three types of problems that lead to specific characteristics of a literary hypertext archivist. The first type of problem stems from the efforts of readers and authors: they are not concerned with preservation beyond making the work accessible in the "next" environment; it is doubtful that any individual effort will make literary hypertexts available into the future. This is due in part to the "closed" nature of typical reading environments: it is hard for the reader to disassemble the work into its components and to reassemble it for some other reading environment. The difficulties

[3] http://www.computerhistory.org/

[4] This mission description is taken from http://www.eliterature.org/pad/

are as much technological as they are legal. The second type of problem is associated with the sustainability of the institution or organization that shoulders the responsibility. The third has to do with expertise: does the organization or institution have the competence to understand the important characteristics of the work to preserve and the skills to undertake the preservation effort? Note that these two abilities may be quite different.

Thus, to implement preservation and archiving, a responsible entity:

- Is willing to assume a general effort;
- Is a sustainable organization or institution;
- Is skilled in preservation; and
- Has a good understanding of literary hypertexts.

It is easy to see that this is a tall order. It may require coordinated efforts on the part of authors, publishers, services, and institutions. Readers may have a role to play as well if they are to maintain viable copies of their favorite works, complete with the records of interaction that make the works their own.

4.3 Associated Issues

Identifying who will preserve, archive, and make accessible hypertext literary works is only part of the problem. Because digital preservation is difficult, potentially costly, and must be addressed in the near term if we are to meet the challenge of a rapidly changing technological environment, several issues are likely to arise that go beyond the scope of preserving physical literary works (although sometimes these issues do arise with works in other media like film). These include legal issues such as copyright and permissions, and social or cultural issues such as selecting which works – among many – are actually preserved.

Many of the works that would be preserved in this scheme are still under copyright. Yet, in our discussion of potentially responsible parties, it is easy to see that preservation efforts may be undertaken by someone who is not the copyright holder. Similarly, permissions in a multimedia work may be complex. Works such as Coverley's *Califia* [8] use popular music, for example. How will preservation efforts be affected by the permissions of portions of the work? Will this fuel additional controversy for the responsible party? Since law is not our area of expertise, and may involve the legal systems of different nations, we will not address this issue further, but rather just suggest that it is one to watch.

The issue of which works to preserve is a difficult one indeed. Many traditional literary preservation efforts are based on the apparent universality of the work, its long-term cultural impact, and whether – given the forces of time – the work still is valued. Technology changes so quickly that we do not have the luxury of neglect to determine what we save and what we shrug off as lost. The authors and original community (with its unavoidable politics) are still active; the distance of time and evolving institutions cannot constitute a second opinion.

5 CONCLUSION

We do not intend to prescribe a solution to the preservation problems we enumerate in this paper. Instead, we want to provide requirements, to describe the lay of the land, and to suggest a framework for undertaking what may be an immense and anxiety-provoking effort. Preservation is necessarily partial (for a variety of cultural and contextual elements are bound to be lost from the original reading experience), but instead of taking that limitation as a given and throwing our hands up in the air, we wanted to take a thoughtful look at the problem and examine the kinds of scaffolding we would need to build to support preservation within the tradeoffs of representational complexity, artistic integrity, institutional cost, and other pragmatic concerns, not the least of which is our inability to predict the future.

Many digital preservation efforts are already underway. It is now up to us, as a community, to build on them and add our unique requirements to theirs. In short, we want readers to continue to search for closure in *afternoon* fifty years hence, to appreciate the Pynchonesque humor in *Victory Garden*, to wander through *Califia* with bemused appreciation of the musical genres of the late 20th century, and – most importantly – to understand the roots of literary hypertexts by experiencing them as fully as possible.

6 REFERENCES

[1] Aarseth, E.J. *Cybertext: perspectives on ergodic literature*. Johns Hopkins University Press, Baltimore, MD, 1997.

[2] Arms, C. Keeping Memory Alive: Practices for Preserving Digital Content at the National Digital Library Program of the Library of Congress. *RLG DigiNews*, *4*, 3 (June 15, 2000).

[3] Bernstein, M. Storyspace 1. In *Proceedings of HT'01* (Århus, Denmark, August 14-18, 2001). ACM Press, New York, NY, 2001, 172-181.

[4] Bornstein, J. and Riley, V. Hypertext Interchange Format -- Discussion and Format Specification. In *Proceedings of the Hypertext Standardization Workshop* (Gaithersburg, MD, January 16-18, 1990). NIST Special Publication 500-178, NIST, Gaithersburg, MD, 1990, 39-47.

[5] Cayley, J. The Code is not the Text (unless it is the Text). *Electronic Book Review* 09-10-02. Available at http://www.electronicbookreview.com/v3/).

[6] Commission on Preservation and Access and the Research Libraries Group. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. Research Libraries Group, Mountain View, CA, 1996. Available at http://www.rlg.org/ArchTF/tfadi.index.htm.

[7] Cooper, B.F. and Garcia-Molina, H. Peer-to-peer data trading to preserve information. *ACM Transactions on Information Systems*, *20*, 2 (April, 2002), 133-170.

[8] Coverley, M.D. *Califia*. Eastgate, Watertown, MA, 1998.

[9] Davenport, J. Defending a Museum. *Online Preservation*, August 23, 2002.

[10] Douglas, J. Y. *The End of Books? Or Books Without End? Reading Hypertext Narratives*. University of Michigan Press, Ann Arbor, MI, 2000.

[11] Gemmell, J., Bell, G., Lueder, R., Drucker, S. and Wong, C. MyLifeBits: Fulfilling the Memex Vision. In *Proceedings of ACM Multimedia'02* (Juan-les-Pins, France, December 1-6, 2002). ACM Press, New York, NY, 2002, 235-238.

[12] Gibson, W. *Agrippa*. Available at http://www.williamgibsonbooks.com/source/agrippa.asp

[13] Golovchinsky, G. What the query told the link: the integration of hypertext and information retrieval. In *Proceedings of Hypertext '97* (Southampton, UK, April 6-11, 1997). ACM Press, New York, NY, 1997, 67-74.

[14] Golovchinsky, G., and Marshall, C. Hypertext Interaction Revisited. In *Proceedings of Hypertext '00 (*San Antonio, TX, May 30-June 3, 2000). ACM Press, New York, NY, 2000, 171-179.

[15] Grønbæk, K. and Trigg, R.H. Toward a Dexter-based model for open hypermedia: unifying embedded references and link objects. In *Proceedings of Hypertext '96* (*Bethesda, MD, March 16 - 20, 1996*). ACM Press, New York, NY, 1996, 149-160.

[16] Harrison, B. L., Fishkin, K.P., Gujar, A., Mochon, C., Want, R. Squeeze me, hold me, tilt me! An exploration of manipulative user interfaces. In *Proceedings of CHI'98* (Los Angeles, CA, April 18-23, 1998). ACM Press, New York, NY, 1998, 17-24.

[17] Hart, P. and Liu, Z. Trust in the Preservation of Digital Information. *Communications of the ACM*, *46*, 6 (June 2003), 93-97.

[18] Hunter, J. and Choudhury, S. A Semi-Automated Digital Preservation System based on Semantic Web Services. To appear in the *Proceedings of JCDL 2004* (Tucson, AZ, June 7-11, 2004). ACM Press, New York, NY, 2004.

[19] Jackson, S. *Ineradicable Stain*. Available at http://ineradicablestain.com/skin.html

[20] Jantz, R. Public Opinion Polls and Digital Preservation: An Application of the Fedora Digital Object Repository System. *D-Lib Magazine*, *9*, 11 (November 2003).

[21] Larsen, D. *Marble Springs*. Eastgate, Watertown, MA, 1993.

[22] Levy, D.M. Heroic measures: reflections on the possibility and purpose of digital preservation. In *Proceedings of Digital Libraries '98* (Pittsburgh, PA, June 23-26, 1998). ACM Press, New York, NY, 1998, 152 – 161.

[23] Lippman, A. Movie Maps: An Application of the Optical Videodisc to Computer Graphics. *Computer Graphics (Proc. SIGGRAPH'80)*. ACM Press, New York, NY, 1980, 32-43.

[24] Lorie, R. A methodology and system for preserving digital data. In *Proceedings of JCDL'02* (Portland, Oregon, July 14-18, 2002). ACM Press, New York, NY, 2002, 312-319.

[25] Luesebrink, M.C. The moment in hypertext: a brief lexicon of time. In *Proceedings of HT'98* (Pittsburgh, PA, June 20-24, 1998). ACM Press, New York, NY, 1998, 106 – 112.

[26] Lynch, C. Canonicalization: A fundamental tool to facilitate preservation and management of digital information. *D-Lib Magazine*, *5*, 9 (September 1999). Available at doi:10.1045/september99-lynch

[27] Marshall, C.C. Making Metadata: a study of metadata creation for a mixed physical-digital collection. In *Proceedings of Digital Libraries '98* (Pittsburgh, PA, June 23-26, 1998). ACM Press, New York, NY, 1998, 162-171.

[28] Marshall, C.C. Reading and Interactivity in the Digital Library: Creating an experience that transcends paper. In *Proceedings of the CLIR/Kanazawa Institute of Technology Roundtable* (Kanazawa, Japan, July 3-4, 2003). 5.4.1-20.

[29] Marshall, C.C. and Ruotolo, C. Reading-in-the-Small: a study of reading on small form factor devices. In *Proceedings of JCDL'02* (Portland, OR, July 14-18, 2002). ACM Press, New York, NY, 2002, 56-64.

[30] Moulthrop, S. Pushing Back: Living and Writing in Broken Space. *Modern Fiction Studies, 43* 3 (Fall, 1997), 651-675.

[31] Pavic, M. *Dictionary of the Khazars*. Vintage Press, New York, 1989.

[32] Rosenberg, J. *Intergrams*. Eastgate, Watertown, MA, 1995.

[33] Shipman, F. and Hsieh, H. Navigable History: A Reader's View of Writer's Time. *New Review of Hypermedia and Multimedia*, *6* (2000), 147-167.

[34] Shipman, F., Girgensohn, A., and Wilcox, L. Hyper-Hitchcock: Towards the Easy Authoring of Interactive Video. In *Proceedings of INTERACT 2003* (Zurich, Switzerland, Sept. 1-5, 2003). IOS Press, Amsterdam, The Netherlands, 2003, 33-40.

[35] Stotts, P.D., and Furuta, R. Petri-net-based hypertext: document structure with browsing semantics. *ACM TOIS*, *7*, 1 (January, 1989), 3-29.

[36] Tansley, R., Bass, M., Stuve, D., Branchofsky, M., Chudnov, D., McClellan, G., and Smith, M. The DSpace Institutional Digital Repository System: Current Functionality. In *Proceedings of JCDL'03* (Houston, TX, May 27-31, 2003), ACM Press, New York, NY, 2003, 87-97.

[37] Waldrip-Fruin, N. Hypermedia, eternal life, and the impermanence agent. In *SIGGRAPH 99 Electronic Art and Animation Catalog* (Los Angeles, CA, August 8-13, 1999). ACM Press, New York, NY, 1999, p. 90.

[38] Zellweger, P.T., Mangen, A., and Newman, P. Reading and writing fluid hypertext narratives. In *Proceedings of Hypertext 2002* (College Park, Maryland, June 11-15, 2002). ACM Press, New York, NY, 2002, 45-54.

[39] Zheng, J.Y. Digital Route Panoramas. *IEEE Multimedia* (July 2003), 57-67.