Preserving the Scholarly Side of the Web

Adam Mikeal, Cody Green, Alexey Maslov, Scott Phillips, John Leggett
Digital Initiatives Research, Texas A&M University Libraries
{adam, cody, alexey, scott, leggett}@library.tamu.edu

Abstract

This paper presents results of a case study that addresses many issues surrounding the difficult task of preservation in a digital library. We focus on a subset of these issues as they apply to the preservation of scholarly articles encoded in current web standards. We also describe the two common preservation mechanisms, emulation and migration, as well as our selection of the latter for our particular case. Finally, we compare two approaches to migration, automatic and manual, and discuss their strengths and weaknesses in our context. We show that consistent use of open standards leads to more efficient migration processes and issue a "call to arms" to the digital preservation community to ensure that scholarly material currently on the web can be preserved for future generations.

1. Introduction

As the Internet continues to become more entrenched in our society, an increasing amount of scholarly writing is found in various forms on the World Wide Web. This trend is positive in many respects. The distributed nature of the Internet can aid in document preservation when multiple copies are made [19] and the nearly ubiquitous access afforded in many parts of the world serves to increase dissemination and access to important information. But this trend has detrimental effects as well. Inherent features of the web create fundamental problems for preservation. The proliferation of data formats found on the web, nonconformity to existing standards due to the lack of a standards-enforcement body, and the massive scale of the collection all contribute to these problems.

Digital libraries have a vested interest in scholarly writings from the incunabular to the present. In order to prevent this critical body of knowledge from vanishing for future generations, we must find stable methods for preserving the digital scholarly record. Exacerbating this issue is the rapidly evolving nature of the web, where standards change dramatically from one generation to the next. The shortened life cycle found on the web obsolesces both data and the standards that encode that data, creating a moving target for the digital archivist. These issues combine to make automatic preservation strategies difficult at best and often impossible.

Although general in scope and found across the web in collections of all types and sizes, the issues described above are frequently managed on a small scale with a single collection or website. This was the situation we faced with the task of migrating the Journal of Digital Information (JoDI) [27] from its aging site architecture to a modern web application capable of handling the complex workflow associated with a peer-reviewed academic publication.

In this paper, we describe seven issues related to preservation of web-based documents. Then we examine two strategies for addressing these issues: *emulation* and *migration*. Next, we discuss our experiences with these strategies in the context of JoDI. Finally, we present the lessons we learned and issues to be faced by the digital preservation community to ensure preservation of scholarly writings. We believe that research should be directed toward developing tools to aid digital archivists in their preservation efforts. This will ensure that valuable information doesn't become lost to the public, locked away in inaccessible formats.

2. The Journal of Digital Information

JoDI is a peer-reviewed, electronic-only academic journal that was first published in 1997. From its conception, JoDI was intended to challenge traditional print paradigms. It started as a web-based journal long before "web publishing" was a common concept and before the general trend toward open access e-journals [2] [4] [7] [31]. JoDI was never intended to be merely a collection of print-based articles available online—its charter explicitly states that the journal wishes to "encourage the presentation of new data sources, [and] data from experimental work". The submission guide-

lines specifically contain instructions on how to deal with hypertexts and other non-printable media.

In many ways the JoDI conversion process represents a synopsis of preservation problems facing much of the information on the web. Although the articles all share a common context and general document structure, the syntactical differences between the articles was significant, since the encoding used to create the presentation formats changed as web standards evolved.

At the time we began the conversion (October 2005), JoDI was composed of six volumes containing 26 issues and 168 individual articles. Many of the articles were encoded in the HTML format that was current at the time of publication. A significant number of PDF files and supplementary files, such as datasets and JPEGs, were also included. A few of the articles were written as experimental hypertexts. These articles contained many smaller files, one for each node of the hypertext.

Due to the ephemeral nature of the web, standards governing the encoding of information often change before the information has reached the end of its lifecycle. Today, a typical choice for encoding scholarly, text-based material for storage and display on the web might be TEI [32]. Based on the options available in 1997 (e.g. no XSL transformations), it was decided that the canonical version of the articles would be stored directly in HTML. The HTML standard in use at the time was HTML 3.2, with a mixture of earlier standards thrown in. As the journal continued along its publication path, the web—and the encoding standards that dictate the file formats and protocols—continued to evolve. By 2005, the encoding formats used in these articles were already 3 generations old.

3. Web Preservation Issues

Many issues exist in digital preservation [10] [13] [15] [21]. In this section we discuss issues that deal specifically with the preservation of hypertexts in today's web ecosystem—the intersection of operating systems, browsers, standards and their implementations.

3.1. Evolving standards

Web standards such as HTML have been constantly changing since the introduction of the web. This rapid evolution in standards can be attributed to the shift from academic interests—which drove the creation of the web—to commercial interests driving the web today [3]. This shift has had a significant impact on both the development of new standards and the evolution of existing standards.

Two significant changes have occurred to the HTML web standard: the transition from SGML to XML and the increasing separation of semantics and presentation. Standards based on the XML paradigm require a stricter encoding than standards based on the SGML paradigm. This transition has raised the effort required to author web documents, while simultaneously creating new opportunities for data interchange.

The early web frequently mixed presentation with semantic markup. Much of the syntax of early HTML was dedicated to formatting the layout of a web page. Recent iterations of these standards have attempted to separate the appearance and content of web pages. These changes in web standards have produced a paradigm shift in the way that web content is conceived and delivered.

Finally, as seen in the HTML standards family, where there have been 5 different standards adopted in the last 10 years [33], the rate of standards evolution is quite rapid. Given the same rate of growth over the next 100 years (a minimal preservation timeframe), there would be 50 separate standards within this single family. As commercial interests continue to influence the growth of the web, there are no indications that this rate of change will decrease.

3.2. Lenient validation

Since the early days of the web, browsers have been notoriously lax in their interpretation—and enforcement—of HTML standards. The result has been a much lower barrier of entry into HTML coding. This encouraged a rapid proliferation in the number of web pages that appeared on the early web and a corresponding diversity in the quality of those documents [18]. Modern browsers have continued to support these inadequately structured documents through the use of separate “standards” and “quirks” modes. In quirks mode, if the document fails to validate, the browser will fall back to a looser interpretation of the standards.

Many browsers currently support this behavior and sites such as the Internet Archive [26] depend on it. However, it is not reasonable to expect that all standards—as well as all misinterpretations of those standards—will continue to be supported in future versions. Ultimately, the sheer number of permutations and contradictory behaviors created by the ever-changing standards will call for the older, least-compatible interpretations to be dropped. These decisions will likely happen without regard for the scholarly value of the endangered documents.

3.3. Inconsistent structure

When working with a set of documents independently authored by multiple individuals, there will often be inconsistencies between the structure and content of each document. These inconsistencies frequently occur even when the authors are working from a common template. This presents a unique set of difficulties for automatic conversion utilities. Although the documents may be syntactically uniform, the underlying semantics can differ enough to make automated approaches infeasible.

Automatically locating a section containing bibliographic references in a document encoded in HTML 3.2 can be a complex task. Any number of methods or labels used for identification—"references", "works cited", and "bibliography"—are all strictly valid according to the standard. Encoding all the various mechanisms necessary for this identification is overly cumbersome. Using techniques such as natural language processing and statistical analysis can help, but these are unlikely to reach the success rate of manual analysis.

3.4. Message preservation

When preserving any document the conservation of the original message is of primary concern. While there is debate regarding the nature of a document [5], we use Levy's definition—a document is the physical manifestation of a message [13]. Extending this definition we recognize three significant aspects of a document: appearance, behavior, and context. All documents will exhibit each of these aspects to some degree. However, each aspect may not be significant to the message of the document.

Figure 1 depicts the relationships between these three aspects. Membership in a set in the diagram indicates whether that aspect has any significant impact on the document's message. If a document's appearance, behavior, and context all have a significant impact on the message, the document will be harder to preserve. A document that does not rely on all three of these aspects will be easier to preserve.

The determination of an aspect's significance hinges on understanding the author's original intent during the document's conception. Judging intent is often a subjective exercise and since the author is frequently uninterested or unavailable for consultation, these decisions must be made by others in the preservation community. Additionally, an author's original intent may have changed since the document's creation or may be unknown even to the author. The following issues will discuss how each aspect—appearance, behavior, and context—affect the preservation of a document's message.

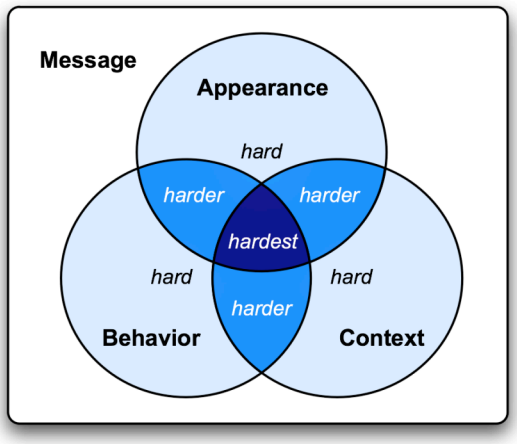

Figure 1. The difficulty of preservation as a document's message is impacted by appearance, behavior, and context.

3.5. Appearance preservation

The appearance of a document may have a significant impact on the message that document conveys to the reader. For scholarly works, it is often the case that the appearance of the document is not of vital significance to the document's message. An example would be a journal article presented in IEEE or ACM format; this change would have little impact on the message transmitted to the reader. However, that same article presented in a child's handwriting in crayon would arguably alter the perceived message.

The paradigm shift described in section 3.1—a move away from mixed semantics and presentation to separated content and appearance—creates difficulties when attempting to preserve the author's original message. During a preservation process the author's original intent must be considered before performing any transformations. Did the author use a table because its contents are actually tabular in nature or because a table was the only means available at the time to format the document in a particular manner?

3.6. Behavior preservation

Behavior is an important concept that must be addressed in any discussion of message preservation. Behavior consists of the mechanisms through which a reader interacts with a document, and presents its own unique difficulties. As the web ecosystem—operating systems, browsers, standards and their implementations—continually evolves, these behaviors constantly change. Some behaviors and instructions present in earlier forms of the web ecosystem—JavaScript, App-

lets, Flash, and ActiveX—may no longer convey the same meaning as they do in the modern environment. For example, a link that includes instructions to open a new window may no longer be appropriate because many users have chosen to disable this feature (most likely due to abuse by commercial advertising). Furthermore, the same message may now be presented by opening a tab instead of a window.

How do these behavioral issues affect the original message? Should they be converted to new behaviors that were not available in the earlier ecosystem? Was the previous behavior simply a choice of a ubiquitous mechanism or did it convey an intrinsic part of the document's message [22]?

3.7. Context preservation

Preserving the context of a document is the most difficult of the three aspects. The context encompasses both the reader's background—economic, socio-political, and academic—and the social and historical milieu that exists at the time of the reading. These issues have been discussed by others [6] [13] [20]. We have chosen to focus specifically on the contextual issues that are unique to the web: the links created by hypertexts, both within and between documents.

To preserve this context, web documents must be resilient to changes in location and reference. Web-based documents do not exist as isolated entities but rather in the greater context of the web. To ensure resilience, a hypertext archivist—an individual responsible for preserving web-based documents—must account for the changes that may occur to documents both in and out of his control. Due to the nature of the web, authors have no control over who links to their work. In addition, throughout a web document's life, its location—represented by its URI—may change as shifts in software mechanisms or organizational control arise. When these shifts occur the relationships between documents must be preserved to maintain link usability.

4. Preservation Strategies

Several viable strategies exist for preserving digital material. This paper assumes that the issues involved in the preservation of the physical bitstream are already resolved and focuses on preserving the usability and readability of web-based documents. Two commonly addressed preservation strategies are *migration* and *emulation*. In this section the strengths and weaknesses of these strategies are evaluated with respect to the previously discussed issues.

4.1. Emulation

Under the emulation paradigm, a document is not altered but is preserved exactly as originally authored. To access the document's content, a user interacts with an interpreter. The interpreter emulates the functionality of the original system that existed at the time of the document's creation [15] [21].

An emulation paradigm is not likely to be feasible in our context for the following reasons. First, multiple interpretations of web standards by modern browsers create incompatible web ecosystems. Each of these ecosystems must be handled by an interpreter. Web authors already have a difficult time creating content that is consistent across the various permutations of web ecosystems. In twenty years this task will be even more difficult. By their very nature, the inconsistencies that create differences within these ecosystems are undocumented and thus become nearly impossible to reproduce. This problem is compounded because we do not necessarily know for which particular ecosystem the author originally intended the document.

Second, the sheer number of permutations for the interpreter to manage is very large. The rapid pace of web standards evolution shows no sign of slowing down at present. It is unreasonable to expect that future browsers—the ostensible interpreter for web documents—will maintain compatibility for all permutations. Instead, it is likely that browsers will drop backward compatibility for particular standards or implementations. This deprecation is likely to occur once it becomes a burden to maintain the standard and/or the number of pages encoded in that standard drops below a certain threshold (likely determined by economic viability).

Third, intra-document links may present a problem for an emulation strategy. Link semantics may evolve in future standards that create incompatibilities with previous versions. Link resolution systems in use today could become incompatible for technical or political reasons. The global nature of the Internet ensures that interrelated content crosses geopolitical boundaries. These boundaries create points of potential instability, as illustrated by recent news reports of DNS name resolution conflicts [8].

Finally, the emulation paradigm presents a further disadvantage by shifting much of the preservation burden from the archivist to the user. The untrained user is ill equipped to deal with these issues of preservation: selecting which interpreter to use, evaluating between competing interpreters, and determining the validity of an interpreter's output. Instead, these issues should be addressed and managed by trained hypertext archivists.

4.2. Migration

Under the migration strategy, documents are periodically translated from older formats to current formats, and users are able to access preserved documents using current tools [30]. This strategy requires continual maintenance to succeed, as documents cannot be left in older formats too long, lest the tools used to interpret those formats become unavailable. In general there are two approaches to migration, automatic and manual. In addition, link migration is a critical issue separate from these two approaches.

4.2.1. Automatic Migration. When migration is performed automatically, algorithms are used to convert from one format to another without human intervention. The main strengths of automatic migration are cost, consistency, and portability.

When considering a large volume of documents, this approach is cheaper than comparable manual methods. Since an algorithm is predictable, the results of the transformations will be uniformly consistent. Also, once an automatic conversion has been developed, it has the potential for reuse in other migration efforts.

The weaknesses of this approach all stem from an algorithm's inability to understand the message on which it operates. An algorithm is unable to consider the author's original intent when deciding how to transform documents. Similarly, an algorithm is unable to create missing information that is implied or encoded in the natural language of the document. This shortcoming can be demonstrated in the use of the `blockquote` element in the HTML specification family. Often this element is used solely for way it alters the document's appearance, rather than the intended purpose of conveying semantic information. Automatic conversions would have a difficult time differentiating between these two uses of this element.

Another problem arises when attempting to impose an external structure. For example, given the task of dividing a document into header, body, and footer sections, an automated algorithm would have difficulty selecting the appropriate elements for each section. The intended purpose would be readily available to a human reader but transparent to an algorithm.

Various conditions must be present before automatic conversion becomes feasible. Primarily, the original data should be consistent. Data consistency enables the automatic conversion to be precisely tuned to the particularities of the documents being migrated. This is significant, because the cost advantage of an automatic migration process is only realized when it is amortized over a large volume of documents.

4.2.2. Manual Migration. Manual migration efforts may include automatic processes, but generally involve significant human intervention at some point during the process. This approach has both strengths and weaknesses in our context. The strengths lie in dealing with problems of lenient validation, author intent, and migration of structures. Although manual processing does not completely resolve the problem of determining author intent, humans are better suited to resolve the issue. Whether migrating the structure of a document into a new template or imposing external structure onto an otherwise flat document, humans are simply better at categorizing sections of documents based upon their content and context.

The manual conversion paradigm does have certain disadvantages: the high cost relative to an automatic process and the lack of scalability when dealing with large collections. In addition, there are a variety of human factors to consider, such as errors and inconsistency prompted by the tedious nature of the task. Manual conversions are costly endeavors requiring a time investment proportional to the size of the document collection. Often, significant training may be required before an individual can participate in the conversion process. When time is an important factor in the project, an additional complication is introduced because the greater the size of the collection, the greater the number of people that must be trained.

Finally, because of the subjective nature of the decisions made during a manual migration process, the results of different individuals may be divergent. This side-effect is to be expected, since the very reason human interjection is needed—the ability to make a subjective judgment call—means that different individuals will reach different conclusions regarding the same questions. A quality control is needed in the process to mitigate this side-effect.

4.2.3. Link Migration. Whether manual or automatic migration strategies are used, link migration must be considered. The specific method used to preserve links depends on whether the hypertext archivist has control over the pointer and/or target component of the link. From this classification of control we derive three types of links based upon their preservation properties: internal links, in-coming links, and out-going links.

Internal links in this context refer to links where both the pointer and target component are under the control of the same archivist. This classification includes two subtypes: the pointer and target refer to the same document, and the pointer refers to a separate document under the archivist's control. Internal links can be preserved by updating the link pointer and target simultaneously. A change in a target's location must trigger a modification to all associated pointers.

In-coming links are pointers found on web pages not under the control of the archivist that target material under the archivist's control. These in-coming links may present a challenge because the original location of the target must be maintained to preserve the validity of the link. To do so the archivist has three options: preserve the original location, create a pointer to the new location, or use a transient location mechanism. Preserving the original location is the most attractive option but may be difficult—if not impossible—due to architectural or political changes. When a target must change locations and there is no pre-existing transient location mechanism, the alternative is to create a pointer at the original location that resolves to the new target. A transient location mechanism is a system that assumes the document may change locations. The CNRI handle system, DOIs, and PURLs are all examples of such a system [24] [25] [28]. However, the use of these systems requires coordination at the time of original authoring.

Out-going links point outward to target resources not under the control of the archivist. These outgoing links present challenges because the documents to which they point may not remain valid or may contain modified content. In these cases there are two options: point to the original location whether it exists or not, or point to an archived copy of the target. While the first option may produce a broken link, it continues to communicate to the user the author's original intended target location. The second option is problematic because it requires archiving the target, although someone else (e.g., Internet Archive) may already be doing this.

Finally, for each of these link types the author's objective must be considered. For example, the author may have intended for the targeted document to be continually fresh when viewed by the reader. The particular strategy used to preserve a link may require subjective analysis in order to determine the author's original intent.

5. Case Study

This case study examines our experience in preserving the Journal of Digital Information, a web-based peer-reviewed journal. This process encompassed three goals: an immediate goal, a short-term goal, and a long-term goal.

1. The immediate goal was to enable an upgrade to a new publishing system. The new web-based architecture improved the peer review workflow, enabled harvesting via OAI-PMH [12] [23], and allowed multiple publication formats. It was determined that the Open Journal System (OJS) [29] met these requirements.

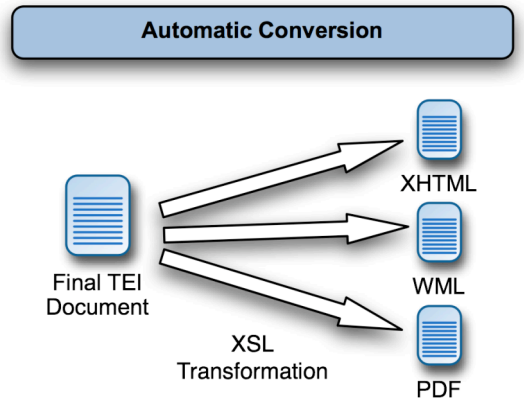

Figure 2. Representation of automatic TEI conversion.

2. The short-term goal afforded by the change in architecture was to enable better data analysis over the documents. The ability to extract references, figures, tables, titles, authors, and other basic document structure was important. Mining this data would enable the journal to offer improved features to readers.

3. The long-term goal was migration to a stable preservation format. After evaluating other standards, TEI was chosen as the long-term preservation format because of its unique text encoding properties. This encoding choice would allow for multiple presentation formats based on a single archived source (Figure 2).

To accomplish these goals two strategies were attempted. A fully automatic migration strategy was tried first. When that approach failed, efforts were redirected towards a manual strategy. The following sections describe our experiences with these strategies in greater detail.

5.1. Automatic Migration

The first strategy attempted was an automatic migration. To accomplish this strategy we used a two-stage process (Figure 3). The first stage involved the application of HTML Tidy to fix many of the issues associated with the older HTML code—misuse of elements, improper nesting of tags, mixed semantic and presentation data, etc. The next stage was to convert the articles—now validating XHTML documents with consistent syntax—to a permanent storage format using the XSL styling language. The format chosen for preservation was TEI, under the assumption that it would have greater longevity than HTML, and as a

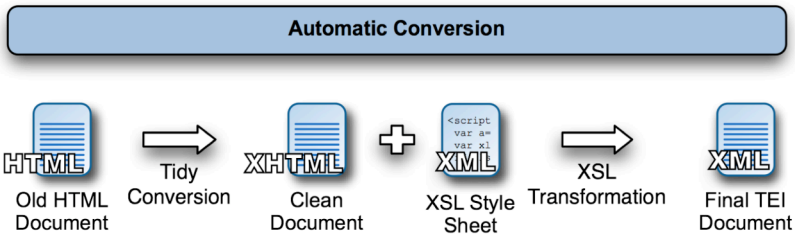

Figure 3. Representation of the automatic conversion process.

published, open XML schema it could be converted to a wide variety of display formats.

The problems we encountered during this process mirrored the issues described in section 3. Lenient validation and evolving standards were partially addressed by the use of HTML Tidy, but the problems of inconsistent structure and message preservation proved more difficult. Inconsistent structure—that is, inconsistency between article structure and content within the limits of the overall format—proved to be an intractable issue for an automatic conversion. Neither Tidy nor XSL could accurately force the source HTML into a consistent TEI-like structure without significant loss of information. While all of the documents in our collection were academic articles with similar structures, many details of the individual articles were unable to be mapped to appropriate structures.

The constantly evolving nature of the web standards used in the documents was also problematic. Deprecated tags could be detected and replaced by Tidy, but certain tags had been obsolete long enough that it was difficult to automatically map the content to new constructs. Choosing the appropriate mechanism for representing the older tags often meant choosing between several methods of encoding. Each method might maintain or lose a different part of the author's original message.

While the structural problems could potentially have been mitigated by encoding special cases for every new case that was encountered, determining author intent for these cases was a more serious problem. Since our target format TEI does not support presentation aspects to the same extent as HTML, we were forced to discard most appearance detail. This could potentially cause a loss of some semantic meaning when the author employed a lenient use of HTML. A manual strategy could potentially determine when presentational aspects contained semantic meaning and re-encode them in the new format correctly.

5.2. Manual Migration

After our experience with the automatic strategy, we redirected our efforts to a manual strategy. We employed a multi-step process combining some automatic methods with human review (Figure 4). The first step—as during the automatic approach—involved the use of HTML Tidy to correct misuse of elements, improperly nested tags, and other minor issues. Next, a common structural template was created to impose structure on the documents. The structure categorized individual components of the document as belonging to front matter, back matter, or the body. It further canonically identified such common components as the title, author, abstract, references, footnotes, figures, and acknowledgements. The documents were then manually translated into the new common structural template.

During this process we were forced to make decisions regarding author intent. This entailed deciding which presentational elements contained semantic meaning and needed to be preserved in the new format. Similarly, we were required to identify which elements held no semantic meaning but were used purely for appearance and could be discarded, if necessary. For example, there were cases of the `blockquote` tag being used merely to indent the text, and not to indicate quoted material. During this process we also encountered issues of document behavior, such as links that opened in other windows. These behavioral aspects

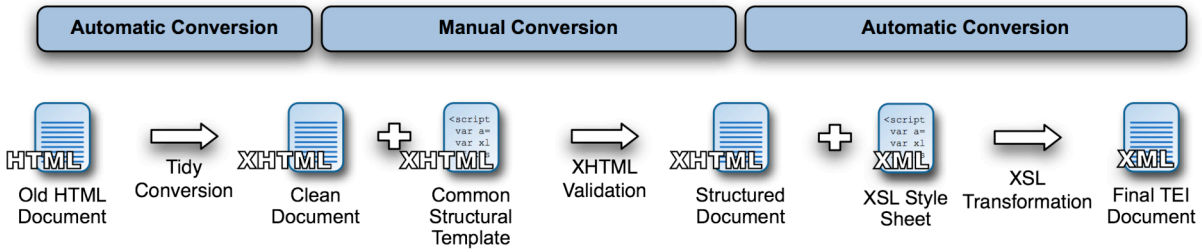

Figure 4. Representation of the manual conversion process.

had to be similarly analyzed for any semantic intent, and the instructions translated into appropriate encodings. Finally, after the documents were translated they were validated for consistency.

The manual conversion process was able to successfully address the issues raised in section 3. As in the automatic migration process, the use of Tidy was able to mitigate the issues of lenient validation and evolving standards. The human decision making abilities were able to address the issues of inconsistent structure and message preservation. By making judgment calls regarding specific elements' semantic meaning, we were able to translate the elements into purely semantic markup for future preservation.

5.3. Link Migration

As the documents were migrated into a new software architecture several link preservation issues were encountered. As discussed previously there are three types of links: internal, in-coming, and out-going. For JoDI each of these link classifications were addressed separately.

Since both the pointer and the target were under our control for internal links, the preservation method was obvious. During the manual conversion process we updated any pointers encountered to match the new location of the target documents in the revised architecture.

In-coming links proved to be problematic. We created a two-phased solution to solve this problem. During the first phase the existing target locations are preserved exactly as they exist in the previous architecture. The second phase involves implementing a pointer system to create a mapping between the old locations and the new locations.

Finally, for out-going links we decided to simply leave them as originally authored. It would be possible archive all documents linked to by JoDI articles. However, more problematic is the issue of archiving the context of those targeted documents. The context of the targeted document includes any documents to which that document also points. This quickly leads to an overwhelming number of documents to be archived. Although the Internet Archive appeared to be an attractive option, we chose against it because its cache is incomplete, and may not contain an appropriate version for the date in question. Furthermore, the Internet Archive presently relies on an emulation strategy to preserve documents, which may not be viable in the long-term for web documents.

5.4. Migration Analysis

Two migration strategies were used to accomplish our preservation goals for JoDI: automatic and manual. After attempting an automatic process and failing, a manual strategy proved successful. Unfortunately, the manual process required a significant investment of time and resources, requiring approximately 120 man-hours for only 168 articles. This time investment makes this strategy unattractive for future migration projects.

Once we completed the manual migration to strict XHTML, the previously attempted automatic migration was successful. Success was achieved because the documents used strict validation and a common document structure. This imposed structure removed much of the ambiguity involved in determining author intent. It is clear that had stricter standards and document structure been enforced from the beginning—whether at the point of authorship or publication—our attempts at automatic migration would have found more success.

6. Lessons Learned

One of the primary lessons we learned from this experience is that automation can and should be applied to the preservation process. In the specific case described here—the JoDI conversion—HTML Tidy was essential in addressing the issues of lenient validation and evolving standards. Other tools, like XSL transformations, also proved invaluable in imparting consistent structure to otherwise diverse documents.

However, issues exist that prevent automation or make it difficult to apply. The most fundamental of these issues is the inconsistency of document structure and the lack of instruction regarding the author's original intent. For example, it is difficult to use an automated process to separate presentation and semantic data. This is particularly true where similar markup is used to convey both types of data. In such cases, manual conversion—in combination with automated processes—is the only efficient migration method to preserve the documents without losing the author's message.

Addressing these issues at the time of the document's creation (or subsequent migration) can reduce the need for manual work later in the preservation process. If the documents adhere to a consistent standard and share a common structure that provides semantic information about its content, many of the aforementioned issues—authorial intent, message preservation, etc.—can be addressed in an automated manner.

7. Future work

Multiple prospects exist for further research in this area. One potential approach is the introduction of common document structures that indicate author intent. These common structures could be domain-specific templates for scholarly writing, email, blogs, commercial venues, governmental sites, etc.

Another area of work would involve evaluating the feasibility of removing support for loose validation from browsers. Commonly known as "quirks" mode, the current practice—dating back to the earliest browsers—has been to silently accept and render malformed HTML. Many questions should be considered. What would be the economic impact of such a change? Should this change occur suddenly or incrementally? How would this affect continuing migration efforts? How would such a change affect projects that currently rely on quirks mode (such as the Internet Archive)?

Finally, the community could also benefit from an investigation into the creation of better automation tools. The current trends in this research area include the application of natural language processing techniques and statistical analysis approaches [11] [1].

8. Conclusion

In a sense, this paper is a "call to arms" to the digital preservation community. At the moment, the web is full of documents that are encoded in a variety of nearly obsolete formats. The documents vary greatly in content, structure and quirks introduced by their respective authors. Steps must now be taken to preserve these documents for future generations, despite the difficulty of the task. Also, steps should now be taken towards greater consistency and adherence to standards—both for newly created works and migrated documents—otherwise the same problems encountered in this case study will continue to recur.

9. Acknowledgements

The authors would like to thank Jay Koenig for his helpful criticism and Valerie Reese for the many hours of effort that enabled us to complete the JoDI conversion project.

10. References

[1] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley Longman Publishing, Boston MA, 1999.

[2] C. W. Bailey, *Open Access Bibliography: Liberating Scholarly Literature with E-Prints and Open Access Journals*, Association of Research Libraries, 2005.

[3] T. Berners-Lee, "WWW: past, present, and future", *Computer*, IEEE, October 1996, pp. 66–77.

[4] A. Buckholtz, "Returning Scientific Publishing to Scientists", *The Journal of Electronic Publishing*, August 2001.

[5] M. K. Buckland, "What is a 'Document'?", *Journal of the American Society for Information Science*, John Wiley & Sons, Inc., September 1997, pp. 804–809.

[6] M. Chalmers, "A Historical View of Context", *Computer Supported Cooperative Work (CSCW)*, Springer Netherlands, August 2004, pp. 223–247.

[7] R. Crow, *SPARC Institutional Repository Checklist & Resource Guide*, The Scholarly Publishing & Academic Resources Coalition, November 2002.

[8] M. Geist, "China and the break-up of the net", *BBC News Online*, BBC, accessed 23 May 2006, URL: <http://news.bbc.co.uk/1/hi/technology/4779660.stm>.

[9] S. Granger, "Emulation as a Digital Strategy", *D-Lib Magazine*, October 2000.

[10] M. Hedstrom, "Digital Preservation: A Time Bomb for Digital Libraries", *Computers and Humanities*, Kluwer Academic Publishers, 1998, pp. 189–202.

[11] P. Jackson and I. Moulinier, *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*, John Benjamins Publishing Company, Philadelphia, 2002.

[12] C. Lagoze, H. Van de Sompel, "The Open Archives Initiative: Building a Low-Barrier Interoperability Framework", *1st ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM, Roanoke VA, June 24–28 2001, pp. 54-62.

[13] D. M. Levy, "Heroic Measures: Reflections on the Possibility and Purpose of Digital Preservation", *Proceedings of the 3rd ACM Conference on Digital Libraries*, ACM, Pittsburg PA, June 24–27 1998, pp. 152–160.

[14] D. M. Levy, C. C. Marshall, "Going Digital: A Look at Assumptions Underlying Digital Libraries", *Communications of the ACM*, ACM, April 1995, pp. 77–84.

[15] R. Lorie, "A Methodology and System for Preserving Digital Data", *Proceedings of the Joint Conference on Digital Libraries 2002*, ACM, Portland OR, July 14–18 2002, pp. 312–319.

[16] C. C. Marshall, G. Golovchinsky, "Saving Private Hypertext: Requirements and Pragmatic Dimensions for Preservation", *Proceedings of the Fifteenth ACM Conference on Hypertext and Hypermedia*, ACM, Santa Cruz CA, August 9–13 2004, pp 130–138.

[17] A. T. McCray, M. E. Gallagher, “Principles for Digital Library Development”, *Communications of the ACM*, ACM, May 2001, pp. 49–54.

[18] H. W. Poole, C. J. P. Moschovitis, *The Internet: A Historical Encyclopedia, Volume III: Chronology*, ABC-CLIO, 2005, p. 123.

[19] V. Reich and D. S. Rosenthal, “LOCKSS: A permanent web publishing and access system”, *D-Lib Magazine*, Corporation for National Research Initiatives, June 2001.

[20] J. Rothenberg, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*, Council on Library & Information Resources, Washington DC, January 1999.

[21] J. Rothenberg, “Ensuring the Longevity of Digital Documents”, *Scientific American*, January 1995, pp. 24–29.

[22] B. Seaman, “Interactive Text and Recombinant Poetics”, *First Person: New Media as Story, Performance and Game*”, MIT Press, Cambridge, 2004.

[23] H. Van de Sompel, C. Lagoze, “The Santa Fe Convention of the Open Archives Initiative”, *D-Lib Magazine*, Corporation for National Research Initiatives, February 2000.

[24] The Corporation for National Research Initiatives Handle System, URL: <http://www.handle.net/>.

[25] The International DOI Foundation Digital Object Identifier System, URL: <http://www.doi.org/>.

[26] The Internet Archive, URL: <http://archive.org/>.

[27] The Journal of Digital Information,
pre-migration URL: <http://jodi.tamu.edu/>,
post-migration URL: <http://journals.tdl.org/jodi>.

[28] Online Computer Library Center Persistent Uniform Resource Locator, URL: <http://www.purl.org/>.

[29] Open Journal Systems published by The Public Knowledge Project at Simon Frasier University Library, URL: <http://pkp.sfu.ca/ojs/>.

[30] *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*, Commission on Preservation and Access and The Research Libraries Group, Washington DC and Mountain View CA, 1996.

[31] Scholarly Publishing and Academic Resources Coalition, URL: <http://www.arl.org/sparc/>.

[32] The Text Encoding Initiative: Yesterday's information tomorrow, URL: <http://www.tei-c.org/>.

[33] The World Wide Web Consortium (W3C), URL: <http://www.w3.org/>.