Preservation Options for HTML 4

Date: August 2, 2005
Author: Grace Carpenter

Introduction

Because HTML is being phased out by the W3C, it is virtually guaranteed that future generations of browsers will not support it, and that non-XHTML documents will become unrenderable.

Forward Migration (to XHTML)

Pros:

• Possibility of preserving formatting and most functionality
• Object remains in an open, human-readable format

Cons:

• Although process can be automated, some human oversight/intervention will often (usually?) be necessary

Due to the long history of browser non-conformance to standards, and the vast number of HTML documents created with that non-conformance in mind, it seems unlikely that migration from one HTML version to the next can ever be done on a fully automated basis. In many HTML documents the relationship between style and meaning is too tightly coupled to undo without human involvement.

• May introduce some browser incompatibilities

Migration to a non-HTML format

There are really no other markup languages that are renderable by web browsers, and that could act as alternatives to HTML/XHTML.

HTML 4 to PDF

Pros:

• Can preserve most formatting and minimal functionality (links)

Cons:

• Migration from open to proprietary format
• Migration from human-readable to binary format
• Many aspects of an HTML could potentially be lost, such as scripting.

Appendix: Possible Migration Tools

Tidy.exe http://tidy.sourceforge.net/

This is the best-known tool for "cleaning up" HTML (and potentially migrating it to XHTML). Integrating it into the DSpace ingest process seems like a possibility for enforcing "clean" HTML or well-formed XHTML code. Possible drawbacks:

• It looks like there's no up-to-date java library/API. DSpace would probably either have to call it from the command line, or perhaps forward the user to one of the web pages that supplies a front end to Tidy (or perhaps a web service?).

• Tidy has numerous options that would probably need to be configured by each individual system administrator; it would be hard to streamline its integration.
• HTML produced by Tidy is cleaner, but not necessarily valid, and therefore needs to also go through a validation step.

W3C Validation Service http://validator.w3.org/
Since Tidy doesn't necessarily produce *valid* XHTML, a second tool is necessary to validate the cleaned-up output. This service is provided by the W3C validation service; however, the user would have to have enough technical knowledge to interpret the validator's output (not always simple), and then fix the underlying causes of errors.

HTML Kit http://www.chami.com/html-kit/
HTML Kit is an integrated environment that allows the user to run Tidy.exe on a file and then call the W3C Validator through the same GUI. A study done for the Smithsonian Institute Archives on preservation of web resources [Dollar] concluded that using HTML Kit to unify the migration/validation process was more time efficient than running separate programs.

References

[Dollar] Dollar Consulting, "Archival Preservation of Web Resources: HTML to XHTML Migration Test Technical Considerations, Evaluation, and Recommendations" Smithsonian Institution Archives; 1 July 2002.
http://www.si.edu/archives/archives/dollarrpt2.html