# Content Permanence Via Versioning and Fingerprinting

Jonathan Simonson [†]
j.simonson@computer.org

Daniel Berleant [‡]
berleant@iastate.edu

Ahmed Bayyari [†]
akb@engr.uark.edu

[†] Department of Computer Science & Computer Engineering
University of Arkansas
Fayetteville, Arkansas 72701-1201, USA

[‡] Department of Electrical & Computer Engineering
Iowa State University
Ames, Iowa, 5011-0002, USA

## ABSTRACT

Referencing documents on the Web is becoming increasingly popular due to the convenience provided to both readers and publishers. Unfortunately this convenience can become just the opposite when referenced documents are altered or removed. This lack of content permanence is of particular concern for works of lasting appeal. To address this problem, a scheme is proposed that both encourages content permanence and detects document version tampering.

**KEYWORDS:** Content permanence, security, hypertext referencing, versioning, and electronic publishing.

## INTRODUCTION

Hypertext and the growth of the Internet have made the Web an immensely valuable information resource. Authors are referencing documents on the Web with increasing frequency. Unfortunately the content permanence of these referenced documents does not parallel that of documents distributed on write-once media. Documents on write-once media, once distributed, are difficult to recall and replace. The result is that previous document editions are allowed to co-exist with newly created editions. On the other hand, documents distributed on the Web are typically replaced automatically as master copies are updated. Thus as master copies are altered, references tend to lose their integrity. Though updates to documents are necessary, prior editions should be preserved and made accessible to help ensure content permanence and reference integrity.

One solution to this problem is the centralized archiving of document editions by a trusted third party (TTP). Through this approach document editions can be secured against tampering. The Xanadu project [4], among others, provides this capability. A further partial example is the Internet Archiving project [3], which captures snapshots of the Web.

A decentralized approach is another alternative. It requires authors and publishers to maintain their own document edition archives. [2, 5, 6, 7] provide a few examples. The systems described, however, were not designed to detect or prevent version tampering.

The work presented here takes a hybrid approach. It is decentralized in its archiving of documents but centralized in its means of detecting document tampering. Authors and publishers are thus able to maintain their documents locally while a remote site provides the means of verifying document edition integrity.

## SYSTEM OVERVIEW

To facilitate the modification of documents on the Web while encouraging content permanence, a form of versioning is needed. Via versioning, referencing authors are able to provide hypertext links to specific document editions. This helps ensure reference integrity since old editions continue to exist as new editions are created. If, however, these document editions are controlled locally, a means of testing edition integrity must exist.

By involving a remote site that maintains digital digests or fingerprints of each document edition, tampering becomes detectable. It is such a system that is proposed here and is currently being developed.

The means by which tampering is detected is not unlike that used by such companies as Surety Technologies and Digistamp. The proposed system works as follows. At the local site, when a document edition is submitted to the version control system for Web access, a fingerprint is generated and then archived at a remote TTP site. The fingerprint is signed by the TTP site and returned to be associated with the document edition. When the document is requested via a hyperlink in a referencing document, the hypertext browser, if designed and configured correctly, can detect and notify the reader of any tampering by recomputing the fingerprint and comparing it with the attached fingerprint. The document could also be transfered to the TTP for verification.

By creating such a system in which tampering is detectable, content permanence is more likely to be maintained.

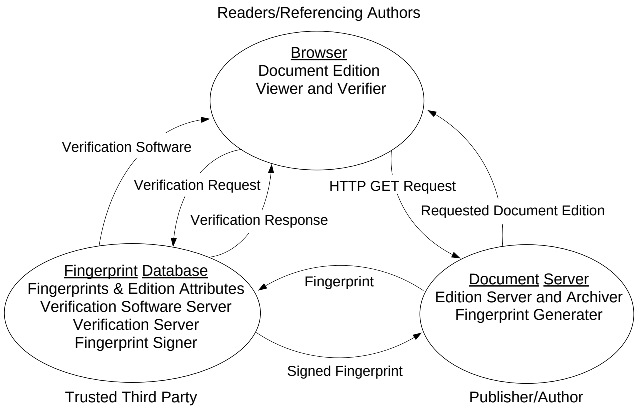

**Figure 1: System Components**

## SYSTEM IMPLEMENTATION

The system is composed of three main components: a document archiver and server, a fingerprint database, and a document viewer and verification client or browser. These components and their interactions are illustrated in Figure 1.

### Document Archiver and Server

The document archiver and server resides at each publisher's site. In the process of archiving, it generates each edition's fingerprint and transfers it to a TTP fingerprint database. Then it associates the returned signed fingerprint from the TTP site with the archived document to enable integrity checking. The archiver is comprised of 1) a versioning system, 2) an Apache Web server and an associated module for extracting requested editions from the versioning system [6], and 3) software integrated with the versioning system to handle the generation of fingerprints and their transfer to the TTP site. The first two items are available and ready for use. The Concurrent Versions Systems (CVS) [1] is used as the underlying versioning system, though other versioning systems such as WebDAV [8] could also be used.

### Document Fingerprint Database

The TTP fingerprint database server is used to maintain information regarding each document edition for which a fingerprint has been submitted. This information, along with the fingerprint, is used to verify document edition integrity. Fingerprints are signed by the server and returned to author sites to be associated with document editions. The server will be designed with a Web interface for edition verification via document transfer. In addition, it will provide software necessary for edition integrity checks directly by a browser in the form of a plug-in or a separate application.

### Document Viewer and Verification Client

Requests for document editions are made via a hypertext browser which serves as the document viewer and verification client. The use of a plug-in or other associated application may be necessary to carry out the actual verification though some browsers such as Netscape Communicator may already have this ability builtin.

Placing editions within signed Java archive (JAR) files is a possible alternative to the use of a plug-in assuming the browser automatically checks the JAR file signature. This is not known to be the case, for instance, with Netscape Communicator where a software install appears to be required before the signer can be identified.

Ultimately, for readers to know if a document edition has been altered, browsers or other applications must be available to verify the integrity of the document, or the reader must be able to verify directly with the fingerprint database server.

## CONCLUSION

By having a system that provides content permanence via versioning and fingerprinting, readers and referencing authors can be better assured of reference integrity then they would otherwise through versioning or fingerprinting alone. Such a system is valuable to both those wishing to reference documents that may undergo changes and those wishing to have their own documents referenced. Ultimately such a system would increase the reliability of the Web as a reference source.

## REFERENCES

1. P. Cederqvist. *Version Management with CVS for CVS 1.9.27.* Sigmum Support AB, Sweden, 1993.

2. H. C. Davis. Referential integrity of links in open hypermedia systems. In *Hypertext 98*, pages 207–216. ACM, June 1998.

3. B. Kahle. Preserving the internet. *Scientific American*, 276:82–83, March 1997.

4. T. H. Nelson. The heart of connection: Hypermedia unified by transclusion. *Communications of the ACM*, 38(8):31–33, August 1995.

5. R. Pettengill and G. Arango. Four lessions learned from managing World Wide Web dignial libraries. In *Digital Libraries '95: The Second Annual Conference on the Theory and Practice of Digital Libraries*, pages 177–180. Springer-Verlag, Heidelberg, June 1995.

6. J. Simonson, D. Berleant, X. Zhang, M. Xie, and H. Vo. Version augmented URI's for reference permanence via an Apache module design. In *Proceedings of the 7th International World Wide Web Conference*, pages 337–345, April 1998.

7. F. Vitali and D. G. Durand. Using versioning to provide collaboration on the WWW. In *World Wide Web Journal: Fourth International World Wide Web Conference Proceedings*, pages 37–50. World Wide Web Consortium, December 1995.

8. E. J. Whitehead and M. Wiggins. WEBDAV: IETF standard for collaborative authoring on the Web. *IEEE Internet Comuting*, 2(5):34–40, 1998.