FILE FORMAT

File Format Name: Hypertext Markup Language

File Extension(s): HTML

Date Introduced: Hypertext Markup Language (First Version), published June 1993 as an Internet Engineering Task Force (IETF) working draft (not standard). HTML 2.0 published November 1995.

Dates in Use: 1990-present. In common use by the World-Wide Web (WWW) global information initiative since 1990, until the date when HTML 2.0 was introduced.

Variations: Some file formats, such as HTML, or the source code of some particular programming language, are in fact also text files, but adhere to more specific rules, which allow them to be used for specific purposes. There is no official standard HTML 1.0 specification because there were multiple informal HTML standards at the time. Subsequent versions of HTML (3.2, 4.0, 4.01, ISO HTML and XHTML) were published as recommendations by the World Wide Web Consortium (W3C), and provided many new capabilities such as support for tables, text flow around figures and the display of complex math elements. They were designed to be compatible with HTML 2.0. The most common extension for files containing HTML is .html, however, older operating systems, such as DOS, limit file extensions to three letters, so an .htm extension is also used. Although perhaps less common now, the shorter form is still widely supported by current software.

Developers: The initial edition was provided by Tim Berners-Lee, and is considered by most to be the definitive HTML 1.0. It was further developed by the IETF with a simplified SGML syntax. SGML stands for Standard Generalized Markup Language and is a metalanguage in which one can define markup languages for documents. HTML is now an international standard (ISO/IEC 15445:2000). The World Wide Web Consortium (W3C) maintains later HTML specifications.

Open Source/Proprietary: Since it is managed by the W3C, it is in a sense open source, although it any modifications have to follow regulated procedures. In accord with the W3C Process Document, a Recommendation progresses through the maturity levels of Working Draft (WD), Candidate Recommendation (CR), and Proposed Recommendation (PR), culminating ultimately as a W3C Recommendation (REC). A Recommendation may be updated by separately-published Errata until enough substantial edits accumulate, at which time a new edition of the Recommendation may be produced (e.g., XML is now in its third edition).

Associated Operating System: The Hypertext Markup Language (HTML) is a simple

markup language used to create hypertext documents that are platform independent, that is, portable from one platform to another. It is an SGML (Standard Generalized Markup Language), and is widely regarded as the standard publishing language of the World Wide Web. It is compatible with both Mac and PC platforms.

Associated Application(s): Requires a web browser, such as Navigator or Internet Explorer, for viewing. Some others: Firefox, Mosaic, Safari. The World Wide Web primarly uses HTTP to serve HTML documents to users. In order to do this correctly, it is necessary for the document to be described correctly: the necessary metadata includes the MIME (Multipurpose Internet Mail Extensions) Type (typically "text/html").

Associated Media: (storage): These files do not tend to be large, and can be stored in most types of media. There is no particular media associated with it.

Compression: HTML has no compression.

Primary Usage: SGML is a language for describing markup languages, particularly those used in electronic document exchange, document management, and document publishing. HTML is an example of a language defined in SGML.

SGML has been around since the middle 1980's and has remained quite stable. Much of this stability stems from the fact that the language is both feature-rich and flexible. This flexibility, however, comes at a price, and that price is a level of complexity that has inhibited its adoption in a diversity of environments, including the World Wide Web.

HTML, as originally conceived, was to be a language for the exchange of scientific and other technical documents, suitable for use by non-document specialists. HTML addressed the problem of SGML complexity by specifying a small set of structural and semantic tags suitable for authoring relatively simple documents. In addition to simplifying the document structure, HTML added support for hypertext. Multimedia capabilities were added later.

Risks: In a remarkably short space of time, HTML became wildly popular and rapidly outgrew its original purpose. Since HTML's inception, there has been rapid invention of new elements for use within HTML (as a standard) and for adapting HTML to vertical, highly specialized, markets. This plethora of new elements has led to compatibility problems for documents across different platforms.

As the heterogeneity of both software and platforms rapidly proliferate, it is clear that the suitability of 'classic' HTML 4 for use on these platforms is somewhat limited. Future (generally upwardly compatible) versions of HTML with new features will be released with higher version numbers, and may not fit all platforms.

These risks hold true for all file formats:

1. Reliance on software and hardware to be able to read them
2. Unstable storage media – need correct storage otherwise will deteriorate
3. Version control – records are easy to amend which makes it difficult to ensure

authenticity, integrity and validity of the record

4. Ease of destruction – once record is lost, it is very difficult or impossible to retrieve

5. Decision to preserve the record needs to be taken when the record is created and needs to take on board not only the life of the record but also the life of the system in which the records are stored

6. Requires new workflows and new responsibilities; and ownership needs to be defined within the institution

Conservation Actions: All of the file formats found on the Internet can be broken into one of two types: ASCII format and binary format. ASCII files are text files you can view with any word processor. Binary files contain non-ASCII characters. If you display a binary file on your screen, you will see a lot of strange symbols and characters.

In particular with HTML, untidy HTML is a reason for some unease about the management of a Web resource. While early versions of HTML had poorly defined structure, the recent redefinition of HTML in the context of XML (XHTML) has now formally defined HTML structure. The TIDY tool makes it possible to determine how well an HTML document conforms to this structure, revealing the sophistication and care of the page's manager.

Efforts of the web development community have led to a new thinking in the way a web document should be written; XHTML epitomizes this effort. Standards stress using markup which suggests the structure of the document, like headings, paragraphs, block quoted text, and tables, instead of using markup which is written for visual purposes only, like <font>, <b> (bold), and <i> (italics). Some of these elements are not permitted in certain varieties of HTML, like HTML 4.01 Strict. CSS (Cascading Style Sheets) provides a way to separate the HTML structure from the content's presentation, by keeping all code dealing with presentation defined in a CSS file. This could aid in the future retrieval of the content if need be. There are discussions of XML as a long-term preservation format, which would involve conversion of older files into it.
Note: This file type can become infected by viruses and should be carefully scanned prior to storage.

Resources:
http://www.faqs.org/rfcs/rfc1866.html
http://www.w3.org/TR/2001/WD-xhtml1-20011004/
http://en.wikipedia.org/wiki/Html#Version_history_of_the_standard

FILE FORMAT

File Format Name: Digital Negative Specification

File Extension(s): DNG

Date Introduced: Sept. 27, 2004

Dates in Use: Sept. 27, 2004 - present.

Variations: None so far. DNG was intended to be a unified raw format.

Developers: Adobe Systems Incorporated

Open Source/Proprietary: The Digital Negative Specification (DNG) is an ostensibly royalty-free raw image file format from Adobe Systems. The same day it was announced, Adobe introduced Digital Negative to the market with its free Adobe DNG Converter program. According to Adobe, Digital Negative was a response to demand for a unifying camera raw file format. Like PDF, Adobe has promised not to enforce the "proprietary" aspect of it.

Associated Operating System: Any platform that supports Adobe programs.

Associated Application(s): Adobe Photoshop Elements 3.0 supports Digital Negative.

Associated Media: (storage): As of 2005, a few camera manufacturers have announced support for DNG, including Leica (native camera support) and Hasselblad (export). Other manufacturers, however, appear to have little interest making their raw files easier to read: cameras from Canon, Nikon, Sony and others include elements of encryption designed to make it harder for others to decode the format. The Leica Digital Module R (DMR) was the first camera back to use DNG as its native format.

Compression: It is uncompressed. Shooting Raw images means photographers can avoid dealing with the compression and loss of image quality involved with shooting JPEGs.

Primary Usage: a new unified public format for raw digital camera files. The company also launched a free software tool, Adobe DNG Converter, which translates many of today's popular raw photo formats into the new .DNG file format, compliant with the Digital Negative Specification.

Risks: Many photographers work in Raw-format files from their digital cameras and are frustrated by the many versions that exist, which vary not just from manufacturer to manufacturer but also from camera to camera.

Adobe is letting any manufacturer that wants to use the format in its cameras, printers and software applications do that for free without any limitations in the hopes of encouraging them to accept it as the standard, but this has yet to happen. Since each manufacturer uses a proprietary format (There are already 75 different formats after just two years of Adobe supporting Raw) that is specific to its cameras and most seem reluctant to accept Adobe's offer, it seems that this file format might become just another format that may need conversion in the future. Only if vendors start supporting this format will it be successful. The market will decide the success.

Conservation Actions: The use of proprietary raw files as a long-term archival solution carries risk, and sharing these files across complex workflows is even more challenging. The use of file formats which have been well documented, have undergone thorough testing and are non-proprietary and usable on different hardware and software platforms minimizes the frequency of migration and reduces the risk and costs in their preservation. Similarly utilizing formats that have been widely adopted minimizes risk, as it is more likely that migration paths will be provided by the manufacturers and a degree of "backward compatibility" will be available between versions of the file format as it evolves. It is important to note that backward compatibility is rarely maintained for more than one or two previous versions and that the "window of opportunity" to migrate is therefore relatively brief. Until DNG format is widely accepted, it is unadvisable to store files in this format.

Resources:

http://www.adobe.com/aboutadobe/pressroom/pressreleases/200409/092704DNGQUOTES.html
http://en.wikipedia.org/wiki/Digital_Negative_Specification
http://en.wikipedia.org/wiki/Camera_raw
http://www.dpconline.org/graphics/medfor/formats.html