Society of Archivists Data Standards Group

A Guide to Archival and Related Standards
Standards applicable to archives; for the digital delivery of repository guides, finding-aids, and images of material from collections.

Title: Hypertext Markup Language

Name of Standards Developing Organisation:
W3C (World Wide Web Consortium)

Current version
HTML5 Draft Standard (W3C Working Draft 16 November 2010)

Replaces

Abstract
HTML is the predominant markup language for Web pages. It was created by Tim Berners-Lee, the inventor of the Web, as a means to provide basic hypertext navigation, menus and simple documentation. It has become incredibly popular and increasingly sophisticated, reflecting the amazing growth and development of the World Wide Web.

Description
HTML stands for HyperText Markup Language. It is a set of markup tags used to describe Web pages. Markup tags are simply pairs of tags, using angled brackets, with a forward slash to indicate the closing tag. A Web browser, such as Internet Explorer or Firefox, can read the HTML markup and display the page. The browser uses the tags to interpret the page.

A set of tags might indicate that content should be a heading, or that it is a list, or that it forms a paragraph. For example, <h1>An Introduction to HTML</h1> shows content within a heading tag. The browser will interpret this as a heading. Paragraphs are indicated by the use of the opening <p> tag and the closing </p> tag.

HTML provides <body> start and end tags to indicate that all the content within these ‘wrapper’ tags is to be displayed. It also provides <header> tags for the content that comprises the header of the page, that is, the metadata about the page, such as its title, the date it was created, who created it and any modifications made.

One of the most powerful features of HTML is the ability to create hyperlinks between pages. This is done by using the <a> tag to define a link. The link is put within what is called an attribute, in this case one called ‘href’. So, a link to the Archives and Records Association would be:

<a href=http://www.archives.org.uk/>Link to the ARA</a>.
You can see that this markup provides the URL, or location, of the Web page, and it provides some text for the link that is displayed in the page.

An HTML element is defined as everything within a tag, from the opening tag to the closing tag. An element may or may not include an attribute, such as the ‘href’ attribute shown above. Most HTML elements can be nested, for example, the content is nested within the <body> tag.

Here is an example of very simple HTML:

```html
<html>
<head>
<title>An Introduction to HTML</title>
</head>
<body>
<h1>An Introduction to HTML</h1>
<p>HTML stands for HyperText Markup Language. You can find out more about HTML at <a href=http://www.w3.org/html/>http://www.w3.org/html/</a>.</p>
</body>
</html>
```

You can see that there are <html> opening and closing tags wrapping all of the content, and then tags to indicate the header information and the main body of text.

HTML has gone through a number of revisions since it was first created by Tim Berners-Lee. HTML5 is now available as a draft standard. It is designed to help create powerful interactive Web based applications. The ability to support Application Programming Interfaces (APIs) is becoming increasingly important, as APIs enable software applications to interact with each other. This helps promote the sharing of content, moving us further towards the principle of open data. It is also designed to be effective on different devices – not just desktop PCs but mobiles and other hand-held devices. HTML5 is much more prescriptive than previous version in terms of telling browsers what to do with the markup. One of the problems with HTML has been that browsers will interpret the tags differently, so that designers can never be sure how their pages will render. The idea is to make browsers more interoperable.

Web pages are often created using editing software such as Adobe Dreamweaver. This allows authors to use a WYSIWYG editor (What You See Is What You Get), so that they do not have to understand the tags and attributes. However, it is important to have an understanding of how to create accessible Web pages that conform to the Disability Discrimination Act, section III, 1999 (DDA). The Act states that websites must be accessible, but gives little information on how to achieve this. The W3C provides useful Web Content Accessibility Guidelines (WCAG) as part of their Web Accessibility Initiative.
HTML provides tags to identify content and it can provide tags for display. However, it is good practice to use CSS (Cascading Stylesheets) for specifying how to display HTML elements. Stylesheets enable the content (the HTML) to be separated from the presentation to a large extent. In fact, HTML was never really intended to contain tags for formatting; it was intended to define content – ‘this is a heading’, ‘this is a paragraph’, etc. Stylesheets can specify the size, weight and colour of text, whether text is indented, how lists are rendered, and all sorts of other layout issues.

In his book, Weaving the Web, Tim Berners-Lee describes the basis of the World Wide Web:

“The art was to define the few basic, common rules of 'protocol' that would allow one computer to talk to another, in such a way that when all computers everywhere did it, the system would thrive, not break down. For the Web, those elements were...universal resource identifiers (URIs), the Hypertext Transfer Protocol (HTTP) and the Hypertext Markup Language (HTML).” (Weaving the Web, Tim Berners-Lee, Texere 1999)

It is URIs that are used to identify resources, including web pages, and HTTP is the protocol enabling the transfer of information across the Web. HTML is the ‘lingua franca’, the basic warp and weft of the Web.

Next month
Next month we will look at a number of international reference code systems defined by ISO (The International Organization for Standardization).

Jane Stevenson
The Archives Hub, Mimas, The University of Manchester