



Tools for Supporting Educational Technology

# Producing text content for the web

UNSW

EDUCATIONAL DEVELOPMENT & TECHNOLOGY CENTRE



**Author: Belinda Allen**

**Date: September 2002**

**Enquires should be directed to:**

Educational Development and Technology Centre (EDTeC)  
The University of New South Wales, Sydney NSW 2052.

Tel: (02) 9385 2458

Fax: (02) 9385 2463

Email: [edtec@unsw.edu.au](mailto:edtec@unsw.edu.au)

Website: [www.edtec.unsw.edu.au](http://www.edtec.unsw.edu.au)

**© The University of New South Wales**

All material in these notes is copyright. Apart from fair dealing for the purposes of private study, research, criticism or review, as permitted under the Copyright Act, no part may be reproduced by any process without written permission of the copyright holder.

# Contents

---

## Producing text for your web site

Overview .....	2
Content structure .....	2
Creating documents.....	3
Formatting text .....	3
Font sizes.....	3
Style sheets .....	4
Using tables .....	4
Converting Word documents.....	4
Removing Word-specific code .....	4
Retaining formatting when saving Word as HTML .....	5
Using Word/Dreamweaver templates.....	5
PDF .....	6
OCR.....	6
Additional resources .....	8



# Producing text content for the web

---

## Overview

Producing text for the web requires particular considerations if the content is to be effectively communicated, and accessible to as many users as possible. Simply saving a printed document as HTML, or scanning it in as an image, is not an effective use of the medium. It is advisable to be aware of issues regarding online literacy, and technical aspects of content production, before you create your web site.

The 'Reading Online' site has some interesting papers on the subject of reading and literacy in the educational online environment:

[http://www.readingonline.org/articles/art\\_index.asp](http://www.readingonline.org/articles/art_index.asp).

Issues that are relevant include:

- Content structure
- Creating your document
- Formatting your document
- Converting Word documents
- Using PDF (portable document format)
- Using OCR (optical character recognition)
- Copyright issues

These issues are discussed here, and links for further exploration and explanation are provided.

## Content structure

Text content for web pages generally requires a different structure than that used in lecture notes or academic papers. It is more difficult to read large amounts of text on screen, and the usual tendency for the screen viewer is to scan the screen for items of interest. For this reason, chunking of content, with appropriate headings and sub-headings, and the use of formatting such as bulleted or numbered lists, or selective indenting, is necessary.

It is useful to follow a 'journalistic' approach to formatting content – that is, provide a brief overview of the content (maybe in point format), then successively elaborate the content, providing links to in-depth explorations of topics, and to documents designed for print. Write as succinctly as possible for content that you intend to be read on screen.

The advice given in 'Accessibility in Web Design' for structuring of content is good advice for general practice:

<http://www.edtec.unsw.edu.au/inter/dload/webmedia/accessibility/content.html>.

There is also some information on accessibility of HTML generally:

<http://www.edtec.unsw.edu.au/inter/dload/webmedia/accessibility/html.html>

Web usability guru, Jakob Nielsen has produced these resources relating to writing for the web:

<http://www.useit.com/papers/webwriting/>.

He recommends using the 'inverted pyramid' approach to content structure: <http://www.useit.com/alertbox/9606.html>, and has some other tips for writing on the web: <http://www.useit.com/alertbox/9703b.html> and how people read on the web: <http://www.useit.com/alertbox/9710a.html>.

## Creating documents

### Word processors vs. web editors

While your document will eventually need to be in HTML format for web delivery, there is no doubt that web editors are generally very clumsy word processors. Conversely, word processing applications, such as MS Word, while their web compatibility is fast improving, are not great web editors. The ideal work process will involve creating and doing basic formatting in a word processor, then importing into a web editor for final formatting.

Unfortunately, while most text editors allow you to save as HTML these days, the HTML generated is often far from ideal, and a web editor will usually do a much better job of generating HTML code.

If you have access to a good web editor (Macromedia Dreamweaver is recommended), then the most sustainable and accessible approach to document production is to produce text in a word processor, paste as unformatted text into the web editor, and apply formatting via HTML style sheets (see Style Sheets).

If, like many, you are much more familiar with your word processing software, and would prefer to format the document that way, before creating HTML, that is possible, although there are some pitfalls to be aware of (see 'Converting Word Documents').

It is also worth considering if you wish to make print versions of the text content, in which case formatting in your word processor for printing before converting to HTML makes sense. Note that the documents may need to be formatted differently for effective web and print delivery.

Microsoft Office and Macromedia Dreamweaver are available to faculty at a subsidised price from UNSW Software Distribution: [http://www.acsu.unsw.edu.au/soft\\_home.htm](http://www.acsu.unsw.edu.au/soft_home.htm).

## Formatting text

### Font sizes

It is possible in HTML to set the font size that will appear on screen. In fact using style sheets (CSS = cascading style sheets) the size can be specified to the exact pixel size, however this is not recommended. Users may be using a variety of browser settings to optimise the screen display for their needs, and setting fixed font sizes will override browser settings.

If font sizes are set in a style sheet, it is possible for the user to override the style sheet in the browser settings, however this will also remove any other formatting you have set in a style sheet. It is preferable to leave font sizes as default and use HTML heading tags to create appropriately sized headings etc.

Jakob Nielsen on font size: <http://www.useit.com/alertbox/20020819.html>

When saving Microsoft Word documents as HTML, the resulting file will often have specified font sizes. The file must be imported into a web editor such as Dreamweaver to remove the font formatting: see 'Converting Word documents' for more information.

## Style sheets

Using style sheets to control other formatting attribute such as text colour, weight, indenting etc is preferable to directly formatting the text. You will need to create and edit style sheets in a web editor such as Dreamweaver, or create a Word document with styles that translate appropriately when saving for the web (see 'Converting Word documents').

For more information about style sheets (CSS) in HTML, see 'Web Design & Construction: Style Sheets':  
[http://www.edtec.unsw.edu.au/inter/dload/webmedia/web\\_design/styles.html](http://www.edtec.unsw.edu.au/inter/dload/webmedia/web_design/styles.html).

## Using tables

Tables are often useful for laying out of text or data, and may enhance legibility if used well. Tables generally convert quite well from Word documents to HTML, although there are some points to be aware of:

- Tables generated by Word will have a fixed width - ie, will not change size when you change the size of the browser window. Columns and rows will also have fixed sizes attached. If this is not what you want, you will need to use a web editor to remove that formatting.
- Empty table cells do not display in some browsers – you will need to add an invisible character inside the cell to retain the cell borders, colour etc.
- The table should be identified as either a data or layout table, and have row and column headers identified, so that screen readers will be able to make sense of the table. This cannot be done in Word, but may be done in Dreamweaver.
- Always create tables in Word or Dreamweaver, rather than use a scanned image of a table, which is completely inaccessible to anyone using a screen reader.

## Converting Word documents

Word documents saved as HTML files are usable on the web, but have some inherent problems that it is preferable to remove if possible:

### Removing Word-specific code

Word HTML files contain information that allows you to open them and edit them in Word. This creates very bloated HTML code that is more difficult for browsers to display. It is possible to filter out these elements, depending on which version of Word you are using.

Either:

- When saving from newer versions of Word, select 'Save as web page, filtered'
- For older versions of Word, a filter may be downloaded from the Microsoft web site (PC only): <http://office.microsoft.com/Assistance/2000/htmlfilter.aspx>
- Dreamweaver has a Word filter that is used by going to 'File>Import>Word HTML'.

If you are using Word as your only web editor, save a full version of the file for editing purposes, and a filtered version for uploading to the web.

## Retaining formatting when saving Word as HTML

Depending on the version of Word you are using, and how you have formatted the document, one of several things may happen to the text formatting:

Older versions of Word:

- Formatting such as headings etc may be lost, with Word instead generating font tags to create font sizes. In this case, it is preferable to strip out font sizes and re-format headings etc using a web editor such as Dreamweaver.

Newer versions of Word:

- If you are using the text styles set up in the standard Word template, the file may be saved with text formatting intact, and the correct formats applied to headings etc.
- If you have created customised styles, Word will instead generate style sheets which approximate the Word styles you have created. This is an effective way to format HTML, but may be difficult for you to edit if you are unhappy with the results.

The most foolproof approach is to either use the standard Word template and styles, or set up your own template with a very basic set of styles, and see how the HTML file looks when you have saved and filtered it. When you have style settings that you are confident will convert satisfactorily without you needing to re-format the document, save the Word document to use as a template for other documents.

## Using Word/Dreamweaver templates

We have created basic templates that you can use for producing your Word documents, then importing them into Dreamweaver, for use in WebCT.

If using older versions of Word (pre-2000), you should:

- Use the Word document as a basis for your own, using the styles set up in it. (**Note:** pasting text in from other Word documents may import styles from the original document that you will need to remove – save your document as text only before cutting and pasting to avoid this.)
- Save your document as HTML, and import into Dreamweaver to remove Word-specific code. It can then be formatted as you wish, or pasted into the supplied Dreamweaver document.

If using newer versions of Word (post-2000), you should:

- Use the Word document as a basis for your own, using the styles set up in it. (**Note:** pasting text in from other Word documents may import styles from the original document that you will need to remove – save your document as text only before cutting and pasting to avoid this.)
- Save your document as 'Web page, filtered', and open it in Dreamweaver. It can then be formatted as you wish (to retain style sheets generated by Word), or pasted into the supplied Dreamweaver document (to remove Word style sheet formatting).

The template documents, plus a 'readme' file on using the documents can be downloaded here: [http://www.edtec.unsw.edu.au/inter/dload/webmedia/dw\\_templates/dw\\_templates.html](http://www.edtec.unsw.edu.au/inter/dload/webmedia/dw_templates/dw_templates.html).

## PDF

PDF, Adobe's portable document format, is useful for providing print format documents over the web, and has the great advantage of retaining the formatting of the original document, however, when using PDF, keep the following in mind:

- PDF documents are usually designed for printing, not for reading on-screen.
- Large amounts of printing may be a financial burden for students – it may be useful to provide printed versions of documents as an alternative to downloading and printing.
- PDF documents created from text documents are searchable and can be made accessible for screen-readers, however PDF documents created from scanned pages lose these attributes.
- PDF documents created from scanned pages are usually much larger than the text-based counterpart.
- Copyright legislation applies to supplying digital material on the web. See 'Copyright issues'.

It is possible to convert a scanned PDF into a text document by using OCR (optical character recognition) software – however, this will lose the page formatting, and usually requires some editing, as unusual characters may not be recognised. It also requires that any graphical content must be re-inserted into the document. See 'OCR'.

Jakob Nielsen:

Avoid PDF for on-screen reading: <http://www.useit.com/alertbox/20010610.html>.

The differences between print design and web design:

<http://www.useit.com/alertbox/990124.html>.

For more information about the PDF format, go to;

<http://www.adobe.com/products/acrobat/adobepdf.html>

Online assistance with using Adobe Acrobat is available at Adobe Studio's Expert Center: <http://studio.adobe.com/expertcenter/acrobat/main.html> (you need to register to use this site).

## OCR

Optical character recognition software can convert a scanned image of a document into digital text, which may then be edited and reformatted. Most flatbed scanners have OCR software bundled when you purchase them; this software may or may not be adequate for your needs. If you will be doing very much OCR, it is worth investing in good quality software (such as Caere Omnipage) – it will save many hours of making corrections to scanned documents.

There are advantages and disadvantages to using OCR for document conversion:

- Page formatting is lost.
- Graphics or images will need to be re-inserted.
- The document may need considerable reformatting to be suitable for web use.
- You will need to review the document for typographical errors generated by the OCR software.
- A document that was inaccessible to screen-readers can be made accessible.
- The text in the document may now be searched, cut/pasted etc.

- The file will be considerably more compact for download than an image file.

It must be kept in mind that a document scanned and converted with OCR will probably require considerable editing and restructuring to be suitable for web delivery. It may be appropriate to provide several versions of the document:

- Edited and simplified for reading from screen
- Full-text OCR for accessibility and research
- PDF with original formatting for printing and annotating

It is possible to 'batch OCR' pages of documents with an automatic document feeder – UNSW Publishing & Printing Services has this facility: <http://www.publications.unsw.edu.au/>.

Not that copyright legislation applies to supplying digital material on the web. See 'Copyright issues'.

## Copyright Issues

On any sites with public access, full copyright clearance must be obtained to publish copyright material on the web,

For uses of copyright material in courses, under educational license, the following requirements apply:

If you wish to use copyrighted material in a course the UNSW library has a new service called the Digitisation Service.

### What is the Digitisation Service?

The University of New South Wales Library Digitisation Service provides a centralised service for processing all digital copyright material required for student course work. The material is stored on the Library's server and is accessed either via the catalogue, MyCourse@UNSW or via school or faculty Web pages. **Students can link directly through WebCT** or other teaching platforms to the material.

The centralised repository, endorsed by the VCAC earlier this year, assists the University to comply with the Copyright Act and provides the necessary record keeping by creating bibliographic records in the Library catalogue.

The Service will either receive copied material for scanning or retrieve and copy material in the Library's collection on behalf of lecturers.

### What are copyright course materials?

There are two main types:

- Printed material such as book chapters and journal articles which the Service scans into PDF format for viewing using Adobe Acrobat Reader;
- Existing digital documents from full text database or electronic journals. Initially during 2002 this material will be scanned, licence agreements permitting, with the plan in 2003 to link directly to digital documents using the new Library system. Requests to access this material must come through the Service to ensure that copyright and licence restrictions are observed.

## What material is beyond the scope of the Digitisation Service?

Lecturers' own material such as lecture notes, tutorial solutions and exercises etc. can remain on course Web sites and are outside the scope of the centralised Digitisation Service.

## What are the copyright restrictions?

The university's copyright site provides detailed information on copyright legislation and copyright restrictions at <http://www.copyright.unsw.edu.au>.

### In summary:

**Hard copy to digital copying of works:** 10% of the number of pages in a work, or one chapter (across the whole university);

**Digital to digital copying of works:** 10% of the number of words in a work, or one chapter (across the whole university);

**Copying of periodical publications:** one article per issue of a periodical publication (from either hard or digital copy). More than one article can be copied if they are on the same specific subject matter. Copying is on a course by course basis unlike book chapters and other works mentioned above, thereby providing a little more flexibility with article copying.

Licence agreements with publishers may restrict copying and communicating further, or alternatively allow us to reproduce more than the Copyright Act permits. The Digitisation Service can advise regarding the implications of the licence arrangements and copyright legislation, and suggest alternative means of providing access to course materials.

## How to submit a request for digitisation

Requests should be submitted to the Digitisation Service, located within Reserve Services, level 2 of the Library, using the Digitisation Request Form at <http://www.library.unsw.edu.au/~gsd/digireq.htm> or if already in digital form, sent to [digitisation@unsw.edu.au](mailto:digitisation@unsw.edu.au). Either way the information listed on the form is required to provide full course and bibliographic details to fulfil copyright requirements. For a journal article: author, article title, journal title, volume, issue, month, year, and pages. For a book: author/editor, chapter title, book title, place of publication, publisher, year, pages (or chapter).

Enquiries should be addressed to the above email address or by contacting either Kerrie Talmacs on extension 2622 or Anna Troiano on extension 2652.

## Additional resources

This site has some dos and don'ts for formatting web content:  
<http://www.gooddocuments.com/techniques/techniqueshome.htm>.

Effective web writing: <http://www.webtechniques.com/archives/2001/02/kilian/>