Technical Report No: 2003/16

Issues in Creating HTML Pages with Welsh or bilingual Content

John Dyke

26 November 2003

Department of Computing
Faculty of Mathematics and Computing
The Open University
Walton Hall,
Milton Keynes
MK7 6AA
United Kingdom

http://computing.open.ac.uk

The Open University

Issues in Creating HTML Pages with Welsh or bilingual Content

John Dyke

Summary

Using utf-8 encoding web pages in HTML, XHTML or XML can contain all the letters used in Welsh including all of the diacritical marks used. These pages are rendered and display correctly on Microsoft's Internet Explorer from version 4 onwards and Netscape's browser from version 4. Opera supports all of the characters correctly from version 6 albeit with font substitution occurring for some of the less frequently used diacritical marks on w and y: the acen grom, the most frequently used diacritical mark, is rendered correctly. Opera's version 5.12 provides a reasonably good coverage. It displays the vowels a,e,i,o and u correctly with all diacritical marks but renders w, W, y and Y without an acen grom but produces blanks for the other less frequently used diacritical marks used over these characters.

For wider support on older browsers, *named character entities* should be used for the characters a, e, i, o and u with diacritical marks. The remaining vowels w and y should coded either without diacritical marks or by some other representation e.g. the character followed by the diacritical mark (w^ in place of ŵ)

The language used on a page should be denoted by using the `lang` attribute in the HTML mark up. Bilingual pages should have appropriate `lang` attributes set on section of code (`div` and `span` tags can be used to host these attributes if needed)

Introduction

Over the recent years there have been significant steps take at internationalisation of the Internet. The major factor enabling widespread internationalisation was the introduction of the HTML 4 standard in Dec 1997. This report examines issues of hosting Welsh language World Wide Web pages. It examines the relevant standards as well as their implementation. The current standards for writing (marking-up) web pages are HTML 4.0 (Raggett, Le Hors & Jacobs) and XHTML 1.0 (Pemberton, S et al,), a recasting of HTML in XML.

This report examines what are the requirements of the Welsh language and what issues its orthographic system poses. The methods of implementation of these codes for use on the Internet is then examined with a survey of how effectively these methods work in practice. Finally the issues of how web pages may be authored is addressed.

This report uses the following definitions:

Letter	A member of an alphabet of a language
Character	A symbol which is used singularly (or in a combination of characters) to represent a letter of an alphabet or a letter with a diacritical mark. It has semantic meaning
Glyph	A representation of a shape which forms all or part of a character. i.e. a character may be described by one or more glyphs
Character code	A number which refers to, or represents, a particular character. In computer and communication terms this is often the value of a byte (or series of bytes).
Character Set	A set of characters which can be directly represented in a particular scheme
Character Encoding	The schema used to relate characters in a character set to their codes.
Accent	A mark, which changes the phonetic value of a character, placed above, below, or to the side of a character.
Diacritical mark	A mark placed above, below, or to the side of a character which functions either as an accent or serves some other function such as indicating change of intonation

The Welsh Language and its Orthography

The Welsh alphabet consists of the following 28 letters:- **a**, **b**, **c**, **ch**, **d**, **dd**, **e**, **f**, **ff**, **g**, **ng**, **h** ,**i**, **l**, **ll**, **m**, **n**, **o**, **p**, **ph** ,**r**, **rh**, **s** ,**t**, **th**, **u**, **w**, **y**. The letter j is often added as it is used in some borrowings or adaptations from English: there are only 15 words starting with **j** in Y Geiriadur Mawr (Meurig Evan & Thomas). Other letters viz: **k**, **v**, **x** and **z** are sometimes used in words of foreign origin. These are particularly prevalent in technical words of classic origin. For example, Kilobeit [Kilobyte](MEU), kilometr [kilometre] (CBAC) Art Nouveau and peyldr X [X ray](Prys & Jones). But these letters can be replaced if the word is writen using Welsh orthography. [typically, k=c; v=f; x=(e)cs;z=s] e.g. cilobeit *kilobyte*, cilometr *kilometre*, gwrthfiotg *antibiotic*, ffacs *fax*, sŵ *zoo* and fertig *vertex*.

Twenty of the letters are represented by a single character and eight of the letters have double characters e.g. **ng** is one letter in the Welsh alphabet but is written as the character n followed by the character g. This means that their representation is easier to achieve by using two characters to represent the letter but there is an issue for collating order in using two characters to represent one letter. The word **llan [ll a n]** would come after the word **lofa [l o f a].**

Welsh has the following vowels a,e,i,o,u,w,y. These may also appear in diphthongs e.g.wy. **w** and **i** can also function as consonants. These vowels and diphthongs can have diacritical marks added to them to indicate changes in sound, stress or differentiation of words with similar letters. The use of diacritical marks is relatively low being typically less than 1% of words.

Welsh uses the following diacritical markings :-

Name	symbol	Main Function	examples
acen grom (**circumflex**)	^	Long vowel	Bûm, cânt, tîm, bôn, sêt, dŵr, tŷ
Didolnod (diaeresis, **umlaut**)	¨	Separately sounded vowel	Copïo, düwch, crëdir, fföedigaeth
Acen ddyrchafedig (**acute** accent)	´	Change of stress	Caniatáu, casét, Baróc
Acen drom (**grave** accent)	`	Short vowel	Clòs, bỳs, sièd

The words, **circumflex, umlaut acute** and **grave,** are used in the international standards to described characters with these marks. These marks can be applied to a capital or small letter.

Punctuation does not present any special problem in Welsh as usage and marks are similar to English. The collnod (’ or ') is frequently used in Welsh occurring to signify the omission of a letter or letters. It occurs typically in about 1% of words.

Character Set for Welsh

To write a web page the letters used are :- a, b, c, ch, d, dd, e, f, ff, g, ng, h ,i, l, ll, m, n, o, p, ph ,r, rh, s ,t, th, u, w, y: these can be represented by the basic Latin alphabet viz. A-Z. The following letters a,e,i,o,u,w,y with a **circumflex** (acen grom), an **umlaut** (didolnod), an **acute** accent (acen ddyrchafedig) or a **grave** accent (acen drom) are also required.

Using the notation in the Unicode standard (Unicode Consortium), the characters A-Z without diacritical marks are in the Basic Latin alphabet. The vowels a,e,i,o,u with all four diacritical marks together with Ý,ý and ÿ are in the Latin-1 alphabet. Ŵ, ŵ, Ŷ, ŷ, and Ÿ are Latin Extended-A alphabet. The remaining characters Ẁ, ẁ, Ẃ, ẃ, Ẅ, ẅ, Ỳ and ỳ are in Latin Extended Additional alphabet.

The Basic Latin alphabet has enjoyed support from the start of the Internet and has a long history in computing and communications terms. Latin-1 has been supported from the advent of HTML 2 (Berners-Lee & Connelly); ISO 8859-1 being the default character set for http transfers. Latin-A and Latin Extended Additional are more recent developments and have poorer support on older systems.

All of the characters needed to host Welsh are in UCS character set (ISO 10646), the standard character set used for HTML4.0 (Raggett, Le Hors & Jacobs). The UCS character set is based on and kept in step with the Unicode character set (Unicode Consortium).

Encoding the characters

Character encoding is how characters are represented in the file hold the text. Typically characters are encoded to a number in a single byte. A byte can represent a numeric value of 0 to 255. Such a scheme can encode up 256 characters. Usually all of these numeric values are not available as some are allocated to control function. For the ISO 8859 family of encoding, 32 are allocated to control functions. Although 256 (or 224) seems a large number, there are 49,194 in the Unicode set (Unicode Consortium).

One of the original character encoding schemes is ASCII (American Standard Code for Information Interchange). This code has 128 values of which 33 are used for control functions. The remaining 95 characters are allocated to letters A-Z in both capital and small letters, the numbers, various punctuation marks and symbols. This character set is suitable for use with English (particularly American English) and presents difficulties for languages that use a Latin alphabet but have a heavier dependence on diacritical marks. ASCII and variants of ASCII have been in such widespread use that it has a fundamental influence on encoding schemes for characters.

Due to technical developments[1] all 256 values of a byte can now be used as a value for a code. Usually the bottom 128 values are the similar to ASCII (or Basic Latin) with the additional codes used for other characters. One such scheme is ISO 8859-1, which includes the Latin-1 characters. These give good coverage of West European languages but 128 characters cannot cover all of the possible characters. For Welsh this covers many of the vowels with diacritical marks but excludes most w and y variants.

Another encoding scheme, ISO 8859-14, was derived for the Celtic languages. This is known as the Latin 8(Celtic) alphabet. This ought to be the natural choice for Welsh. However, implementation is limited and in particular is not supported by current Microsoft Software.

There are number of proprietary encoding schemes such as Windows-1252. These are broadly similar to ISO 8859-1. These are best avoided in favour of international standards: they contain some unusual character mapping and may cause difficulties on some platforms. (For example the Euro symbol € is mapped in Win-1252 to 128_{10} instead of U+20AC this character does not exist in Latin 1)
Another family of encoding schemes is one to represent ucs or Unicode. These are multi-byte schemes i.e. a sequence of bytes is used to represent the character. Of these utf-8 is best suited for Welsh were the majority of characters are represented by a single byte and diacritical characters require two bytes with the rarer ones taking three.

[1] ASCII used 7 bits of the byte with the eighth bit being used for parity checking. Parity checking is usually done on a block of bytes and thus the 8th bit can be used for encoding characters. This increases the number available to double the number i.e. 256 characters.

Character representation

As some character encoding schemes cannot directly represent all characters we may want to include, HTML offers methods, called character references, for referring to any character.

- The syntax "&#*D*;", where *D* is a decimal number, refers to the ISO 10646 decimal character number *D*.
- The syntax "&#x*H*;" or "&#X*H*;", where *H* is a hexadecimal number, refers to the ISO 10646 hexadecimal character number *H*. Hexadecimal numbers in numeric character references are case-insensitive.
- Names entity ä É î ò

Combining Characters

A further method of adding diacritical marks is the use of combining characters. Unicode has such a scheme. For example, U+300 adds a grave accent above the preceding letter. These are not the preferred way of representing a character which exists in the character set. Thus U+0175 is a preferred way of representing ŵ rather than the sequence U+077(a "w") U+302(a combining "^"). In any event support for these characters is poor. Internet Explorer 5.5 and Netscape 6 and above render them correctly. Opera 7 makes only renders the grave accent and places it incorrectly. There is partial support in Internet Explorer 4 and Netscape 4

Candidate encoding schemes

Scheme	No of bytes	Approx. File Size	Alphabet	Notes
ISO 8859-14	1	100%	Latin-8 (Celtic)	Encoding scheme which encodes all the characters used in the Celtic languages (Welsh Irish, Scottish Gaelic, Cornish, Manx and Breton)
ISO 8859-1	1	101% - 107%	Latin-1	Widely used and available. Covers all the characters except those in Latin A and Latin Extended.
utf-8	1 to 3 (variable length)	101%	ucs / Unicode	Encodes all ucs characters. Widely available.

Browser Support

Seven sets of tests were carried out on a range of browsers using Opera 3 to Opera 7, Microsoft Internet Explorer 3 to 6 and Netscape 3 to NS 7. The tests were carried out on Windows 98 and Me for Internet Explorer 4 to 6, the Opera browsers and the Netscape browsers. Windows 95 was the operating system hosting Internet Explorer 3. Opera is particularly important as although it is lower in general usage world-wide, its usage for Welsh medium web sites is likely to be higher as a Welsh language localised version of the browser is available.

These tests check which encoding schemes and character representations are supported and note any defects in support.

Test Schedule

Test Number	Details
1	The characters **a, e, i, o and u** with all combinations of diacritical marks in both upper and lower case. Characters encoded in **utf-8**
2	The characters **w and y** with all combinations of diacritical marks in both upper and lower case. Characters encoded in **utf-8**
3	The characters **a, e, i, o and u** with all combinations of diacritical marks in both upper and lower case. Characters represented by **named character entities**
4	The characters **w and y** with all combinations of diacritical marks in both upper and lower case. Characters represented by **decimal numbered character entities** using the decimal version of their Unicode number
5	**All** letters encoded in iso 8859-14 **Latin 8 (Celtic) encoding**
6	The characters **a, e, i, o and u** with all combinations of diacritical marks in both upper and lower case. Characters encoded in **iso 8859-1 Latin 1 (Western European) encoding**
7	The characters **w and y** with all combinations of diacritical marks in both upper and lower case. Characters represented by **hexadecimal numbered character entities** using the decimal version of their Unicode number
8	All of the vowel (a,e,i,o,u,w,y) with the Unicode combing diacritical mark characters

Results of Browser Tests

Test		Op 3	Op 4.02	Op 5.12	Op 6.03	Op 7	IE 3	IE 4	IE 5	IE 5.5	IE 6	NS 3	NS 4.08	NS 4.7	NS 6.2.2	NS 7
1	utf-8 a-u	×[1]	✓	✓	✓	✓	×[2]	✓	✓	✓	✓	×[2]	✓	✓	✓	✓
2	utf-8 wy	×[1]	×[3]	×[4]	✓[6]	✓[6]	×[2]	✓	✓	✓	✓	×[2]	✓	✓	✓	✓
3	named entities a-u	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	decimal entities	×	×	×	✓[6]	✓[6]	×	✓	✓	✓	✓	×	×	×	✓	✓
5	Celtic encoding	×[5]	×[5]	×[5]	✓[6]	✓[6]	×[5]	×[5]	×[5]	×[5]	×[5]	×[5]	×[5]	×[5]	✓	✓
6	Latin-1 encoding	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7	Hex entities	×	×	×	✓[6]	✓[6]	×	×	✓	✓	✓	×	×	×	✓	✓
8	Combining Characters	×	×	×	×	×	×	×[7]	✓	✓	✓	×	×[7]	×[7]	✓	✓

Notes

1. odd incorrect characters displayed
2. Utf-8 characters are echoed as if 8859-1 encoding were used e.g. Åµ displayed instead of Ŵ
3. squares displayed
4. letters wy displayed without acen grom; w and y with other diacritical marks as squares
5. treats encoding as 8859-1 (or Windows 1252) not 8859-14 e.g. Ŵ rendered as Ð
6. Characters rendered correctly but in a different font. The tests used Verdana as the specified font but Opera 6.03 + rendered these in a serif font (Probably Times New Roman)
7. Only acute and grave accents supported

Summary of Results of Browser Tests

1. “Named Entities” are supported for all browsers tested (versions 3 and upward)
2. utf-8 is supported on all version 4 and above browsers for Internet Explorer and Netscape. Opera supports utf-8 for all characters, albeit with font substitution, from version 6. (Opera Version 5.12 provides a reasonable rendition of all letters in Latin 1 and Latin Extended A but not Latin Extended Additional.)
3. Decimal Numbered Entities are less well supported than utf-8 being supported for the version 6 and above for Netscape and Opera, and for Internet Explorer 4 and upwards.
4. Hexadecimal Numbered Entities have less support than decimal numbered identities being supported by version 6 and above for Netscape and Opera, and from version 5 of Internet Explorer i.e. the same coverage as decimal entities with the loss of support for Internet Explorer 4.
5. Celtic Encoding (ISO 8859-14 Latin 8) is supported by version 6 and above for Netscape and Opera: it is not supported by Internet Explorer.
6. ISO 8859-1 Latin-1 encoding is widely supported and works for all the browsers tested (i.e version 3 and above of Opera, Netscape and Internet Explorer) but only supports the vowel a-u.

Statistics

Browser statistics must always used with caution. The following statistics are taken from "theCounter.com" are representative of the 1 million or more web sites using their hit rate counter. The figures are for March 2003.

Browser	Hits	percent	cum percent	utf-8 support	Celtic support	numbered entities	hex entities
1. MSIE 6.x	199739964	56.21844%	56.21844%	56.21844%		56.21844%	56.21844%
2. MSIE 5.x	130209049	36.64840%	92.86684%	36.64840%		36.64840%	36.64840%
3. Netscape 5.x	6698077	1.88523%	94.75207%	1.88523%	1.88523%	1.88523%	1.88523%
4. Netscape 4.x	5650536	1.59039%	96.34246%	1.59039%			
5. MSIE 4.x	4995473	1.40602%	97.74847%	1.40602%		1.40602%	
6. Netscape comp.	3415921	0.96144%	98.70991%	0.96144%			
7. Opera x.x [1]	2254700	0.63460%	99.34452%	0.31730%	0.3173%	0.31730%	0.63460%
8. Netscape 6.x	1138778	0.32052%	99.66503%	0.32052%	0.32052%	0.32052%	0.32052%
9. Unknown	877957	0.24711%	99.91214%				
10. MSIE 2.x	115502	0.03251%	99.94465%				
11. Netscape 3.x	102264	0.02878%	99.97343%				
12. MSIE 3.x	89746	0.02526%	99.99869%				
13. Netscape 2.x	4150	0.00117%	99.99986%				
14. Netscape 1.x	312	0.00009%	99.99995%				
15. MSIE 1.x	177	0.00005%	100.00000%				
sample size=355,292,606				**99.34773%**	**2.52305%**	**96.79590%**	**95.70719%**

Source: http://www.thecounter.com/stats/

[1] The Opera figure may be lower than actual value due to Opera browsers **spoofing** other browsers. Half the Opera figure is assumed to be Op6 and above

Denoting the Language in Web pages

The language used in the page should be described by using a language attribute (lang) on the HTML tags. The values are given in ISO639-1. Thus cy for Welsh, en for english. The attribute may be applied to the root tag <html> is all or the majority of the document is in Welsh. This permits search engine robots to correctly identify the language of the page contents and any textual description in the head.

Some text readers, used by visually handicapped users of the web, can read in Welsh but need to have the language marked by these attributes in order to determine how to pronounce the text correctly. Adding the language attribute improves accessibility and may be a necessary requirement for compliance with the Disability Discrimination Act (DDA)

```
<html lang="cy">
<head> …<body>
<p>Croeso</p><p lang="en">Welcome</p><p lang="x-
klingon">nuqneH</p>
<p>… mae trigolion yn dweud <span lang="nl">Goede
Morgen</span>sy'n golygu <i>bore da</i></p>
…</html>
```

Fontage

In order to render the characters correctly on the screen the characters must be correctly encoded, the browser must support and understand the encoding **and** there must be a font available to display the character. If a font is specified on a web page and that font is not available (or does not contain a particular character) the browser will often substitute the font for another one which has those characters. The substitution is carried out on a character by character basis.

Internet Explorer is distributed with a number of fonts which include the required Unicode character blocks.

Conclusions

1. utf-8 encoding is the best method. Utf8 gives a good coverage of modern and recent browsers (over 99% of browsers based on the previous statistics), with full support for all of the required characters of Welsh. This excludes version 3 browsers from correctly displaying characters with diacritical marks. In the case of Opera from version 6.03 and higher full support is provided. [version 5.12 supports a-u with all diacritical marks and Ý ý and ÿ. It gives a graceful degradation of w and y with acen grom but no support for the Latin Extended Additional characters]
2. For the widest coverage only native Latin-1 characters should be used or their equivalents in named entities. This covers all version 3 browsers tested (and above) and is probably all so the case for earlier browsers. No diacritical marks can be placed on w or y except Ý ý and ÿ. (This only add an additional 1%)
3. Celtic encoding is not suitable for general use but would be an option where only version 6 and above Opera or Netscape browsers are used. This might be an option on an Intranet site. (Suitable browsers share world wide is approx 2.5%.) Usage by Welsh speakers may be higher due to a localised Welsh version of Opera)
4. Numbered entities are less well supported than utf-8 and should be avoided unless support for Netscape 4.7 and Opera 5.12 browsers is thought unnecessary. (Gives a coverage of approx 97% with decimal entities and 96% with hexadecimal entities) [Netscape 4.7 has declined from 17% share on theCounter in 2001 to 9% in 2002 and 1.5% in 2003.]
5. Whole HTML pages should be marked up to indicate the language used. If multiple languages are used on a page then individual sections should be marked to indicate which portions are in English and which in Welsh.
6. Care needs to be taken when looking at fontage to see that a font specified is capable of supporting the Welsh alphabet including characters with diacritical marks.

Appendices

Character Codes

Capital Letters a-u

Chr	Latin 1		Latin 8(Celtic)		Unicode/ucs	
	Hex	Dec	Hex	Dec	Hex	Dec
À	C0	192	C0	192	00C0	192
Á	C1	193	C1	193	00C1	193
Â	C2	194	C2	194	00C2	194
Ä	C4	196	C4	196	00C4	195
È	C8	200	C8	200	00C8	200
É	C9	201	C9	201	00C9	201
Ê	CA	202	CA	202	00CA	202
Ë	CB	203	CB	203	00CB	203
Ì	CC	204	CC	204	00CC	204
Í	CD	205	CD	205	00CD	205
Î	CE	206	CE	206	00CE	206
Ï	CF	207	CF	207	00CF	207
Ò	D2	210	D2	210	00D2	210
Ó	D3	211	D3	211	00D3	211
Ô	D4	212	D4	212	00D4	212
Ö	D6	214	D6	214	00D6	214
Ù	D9	217	D9	217	00D9	217
Ú	DA	218	DA	218	00DA	218
Û	DB	219	DB	219	00DB	219
Ü	DC	220	DC	220	00DC	220

Lower Case Letters a-u

Chr	Latin 1		Latin 8(Celtic)		Unicode/ucs	
	Hex	Dec	Hex	Dec	Hex	Dec
à	E0	224	E0	224	00E0	224
á	E1	225	E1	225	00E1	225
â	E2	226	E2	226	00E2	226
ä	E4	228	E4	228	00E4	228
è	E8	232	E8	232	00E8	232
é	E9	233	E9	233	00E9	233
ê	EA	234	EA	234	00EA	234
ë	EB	235	EB	235	00EB	235
ì	EC	236	EC	236	00EC	236
í	ED	237	ED	237	00ED	237
î	EE	238	EE	238	00EE	238
ï	EF	239	EF	239	00EF	239
ò	F2	242	F2	242	00F2	242
ó	F3	243	F3	243	00F3	243
ô	F4	244	F4	244	00F4	244
ö	F6	246	F6	246	00F6	246
ù	F9	249	F9	249	00F9	249
ú	FA	250	FA	250	00FA	250
û	FB	251	FB	251	00FB	251
ü	FC	252	FC	252	00FC	252

Letters W and Y

Chr	Alphabet Latin	Latin 1		Latin 8(Celtic)		Unicode/ucs	
		Hex	Dec	Hex	Dec	Hex	Dec
Ŵ	Extd A	-	-	D0	208	0174	372
ŵ	Extd A	-	-	F0	240	0175	373
Ŷ	Extd A	-	-	DE	222	0176	374
ŷ	Extd A	-	-	FE	254	0177	375
Ÿ	Extd A	-	-	AF	175	0178	376
ÿ	1	FF	255	FF	255	00FF	255
Ẅ	Extd Add	-	-	BD	189	1E84	7812
ẅ	Extd Add	-	-	BE	190	1E85	7813
Ẁ	Extd Add	-	-	A8	168	1E80	7808
ẁ	Extd Add	-	-	B8	184	1E81	7809
Ẃ	Extd Add	-	-	AA	170	1E82	7810
ẃ	Extd Add	-	-	BA	186	1E83	7811
Ỳ	Extd Add	-	-	AC	172	1EF2	7922
ỳ	Extd Add	-	-	BC	188	1EF3	7923
Ý	1	DD	221	DD	221	00DD	221
ý	1	FD	253	FD	253	00FD	253

Combining Diacritical Marks

Chr	Unicode/ucs	
	Hex	Dec
◌̀	0300	768
◌́	0301	769
◌̂	0302	770
◌̈	0308	776

These combining diacritical marks cause the diacritical mark to be placed above the preceding character.
This not a preferred method.

References

Berners-Lee, T & Connelly D, HyperText Markup Language RFC 1866, Nov 1995

Bwrdd yr Iaith Gymraeg/Welsh Language Board, Canllawiau Dylunio Dwyieithog/Biligual Design Guide, 2001

CBAC, Termau Ffiseg a Mathemateg, (Physics and Mathematics Terms), CBAC (Cyd-Bwyllgor Addysg Cymru), 1983

Elwyn Hughes, J, Canllawiau Ysgrifennu Cymraeg, ISBN 1 85902 598 6, Gomer, 1998

IANA, Registered list of character sets , http://www.iana.org/assignments/character-sets

ISO639,*Codes for the representation of names of languages*, ISO 639:1988. (www.loc.gov/standards/iso639-2/langhome.html).

ISO 3166 Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes (www.iso.org or country codes from the RIPE Network Coordination Centre http://userpage.chemie.fu-berlin.de/diverse/doc/ISO_3166.html), 1997

ISO/IEC, Information technology -- 8-bit single-byte coded graphic character sets -- Part 1: Latin alphabet No. 1, ISO/IEC 8859-1, 1998 (http://www.iso.org/)

ISO/IEC, Information technology - 8-bit single-byte coded graphic character sets - Part 14: Latin alphabet No. 8 (Celtic) ISO/IEC 8859-14, 1998 (http://www.iso.org/)

ISO, ISO 10646, Information Technology -- Universal Multiple-Octet Coded Character Set (UCS) (http://www.iso.org/), 1993-2001

MEU, *Geiriadur Termau Cyfrifiadureg* (Dictionary of Computer Terms), MEU Cymru, 1995 ISBN 1-870055-51-9

Meurig Evan, H & Thomas, W O, *Y Geiriadur Mawr* (The Complete Welsh-English English-Welsh Dictionary), Christopher Davies & Gwasg Gomer, 1983

Pemberton, S et al, *XHTML™ 1.0: The Extensible HyperText Markup Language A Reformulation of HTML 4 in XML 1.0* W3C Recommendation 26 January 2000 (http://www.w3.org/TR/xhtml1)

Prys, D & Jones, J P M, Y Termiadur Ysgol (Standardised terminology for the schools of Wales), ACCAC, 1988,ISBN 1-86112-180-6

Raggett, D,. Le Hors A, Jacobs I.(Eds), *HTML 4.01 Specification* 24 Dec 1999. (http://www.w3.org/TR/REC-html40/)

Thomas, P W, *Gramadeg y Gymraeg*, Gwasg Prifysgol Cymru, 1996, ISBN 0-7083-1357-4

Unicode Consortium,The. *The Unicode Standard, Version 3.0.0,* defined by: *The Unicode Standard, Version 3.0* (Reading, MA, Addison-Wesley, 2000. ISBN 0-201-61633-5). . (http://www.unicode.org/)