Insular Celtic Language Mark-up in WordPress

Mícheál Mac Lochlainn
Acadamh na hOllscolaíochta Gaeilge, Ollscoil na hÉireann Gaillimh,
Roisín na Mainiach, Carna, Condae na Gaillimhe, Éire
micheal.maclochlainn@oegaillimh.ie

RESUME

WordPress est une base populaire pour la création des sites Internet. Selon les statistiques actuelles, 38% des sites construits sur de tels systèmes de gestion de contenu (SGC) l'utilisent (Built With, 2016). Cependant, ses outils d'édition structurent le contenu des documents avec le balisage HTML, qui est sémantiquement compromis parce qu'il préfère le paradigme WYSISWYG (What You See Is What You Get) à l'approche WYSIWYM (…What You Mean), qui est sémantiquement significative. Músgraí WYSIWYM WP est un module d'extension WordPress qui remplace cette fonctionnalité WYSIWYG par sa propre fonctionnalité sémantiquement forte. L'éditeur rudimentaire de son module d'extension central est enrichi par des modules d'extension supplémentaires avec des fonctions spécifiques de balisage sémantique.
Cet exposé traite du développement de deux de ces modules d'extension qui facilitent l'annotation sémantique basée sur la linguistique, le balisage et le style de présentation de textes écrits dans les langues celtiques et leurs dialectes.

ABSTRACT

Insular Celtic Language Mark-up in WordPress
WordPress is a popular website creation framework. Current statistics indicate that 38% of websites built using such content management system (CMS) technologies are based on it (Built With, 2016). However, its editing tools structure document content with HTML mark-up that is semantically compromised, favouring the presentationally focussed WYSIWYG (What You See Is What You Get) paradigm over the semantically meaningful WYSIWYM (What You See Is What You Mean) approach. Músgraí WYSIWYM WP is a WordPress plug-in that replaces this WYSIWYG functionality with semantically sound WYSIWYM functionality of its own. Its plug-in core implements a basic WYSIWYM editing environment and additional plug-in modules extend this with domain-specific tools for rich semantic mark-up.

This paper discusses the development of two such plug-in modules, which facilitate linguistically-based semantic annotation, mark-up, and presentational styling of text written in the Celtic languages and in dialects thereof.

MOTS-CLÉS: WordPress, langues celtiques, WYSIWYM, balisage sémantique.
KEYWORDS: WordPress, Celtic languages, WYSIWYM, semantic mark-up.

1 WordPress, language and dialect

It should be stressed that WordPress and its graphical editor are general-purpose tools; excellent at what they do but not specifically designed for semantically rigorous linguistic mark-up. Nothing in this paper is intended, nor should it be taken, as critical of either. Regarding language, it is of course quite natural for languages and dialects to be in states of chronological and generational flux. So it should also be stressed that the emphasis here is on annotational accuracy, not linguistic purity.

2 'The Insular Celtic territories'

The plug-in extension modules discussed here reflect a design philosophy predicated on support for explicit, granular identification of any established, multi-generational L1 Insular Celtic language communities that exist or have historically existed in any geographic region. This support extends to any non-Insular Celtic languages around which overlaping L1 language communities have coalesced. For convenience, these regions are referred to here as 'the Insular Celtic territories'.

3 WordPress, semantic integrety and Músgraí WYSIWYM WP

Web pages are electronic plaintext documents, delivered to the browser in HyperText Markup Language (HTML). This language logically structures document content by marking-up the text with semantically meaningful plaintext tags: ** **this is a paragraph** **, **<q>**this is a quote**</q>**, **<cite>**this cites a creative work**</cite>** and so on. The browser uses a separate technlogy, Cascading Stylesheets (CSS), to manage the on-screen visual presentation of content elements based on their bounding tags; paragraphs are rendered in body text for instance, while quotes are placed in double-inverted commas and citations are italicised. Although the WordPress graphical editor writes technically valid HTML its semantic scope is quite limited (no means to mark-up quotes or citations, for example). Of greater concern, its **bold** and *italic* buttons work by applying the tags that signify **strong importance** and *emphatic stress* respectively, regardless of the actual semantic meaning intended by the visual styling; consider, for example, the typographical conventions regarding **snag words**, publication titles (such as *Séadna*) and taxonomic designations (such as *felis catus sapiens*). Finally, it allows the user to inappropriatly and inconsistently use visual, CSS-based styling to apply logical pseudo-structure. These things need not necessarily cause problems for sighted human readers but they can spoil the output of Braille readers and speech synthesisers as used by the blind, and can compromise the results of automated data mining. The Músgraí WYSIWYM WP plug-in core strips the WordPress graphical editor of buttons and options that facilitate this semantically compromised WYSIWYG markup and replaces them with semantically sound WYSIWYM-based alternatives. So for example, instead of italicising the publication title *Séadna* by applying inappropriate stressed emphasis (****Séadna****) it can be italicised by explicitly identifying it as a creative work (**<cite>**Séadna**</cite>**) or, with greater and therefore more useful specificity, as a book (**<cite class="leabhar">**Séadna**</cite>**). This mark-up is perfectly standards-compliant, and where the standards do not extend to the desired level of granularity it follows best current practice. Consequently, the semantic metadata are quite parsable and can be made available to the human reader simply by building the requisite functionality into the website back end. A number of plug-in extension modules exist, and more can be developed, to broaden the semantic scope of the editor by providing buttons and options to apply markup using controlled values relevant to specific domains.

4 The Extended Functionality (Celtic Languages) module

This module is a prototype for adding linguistic annotation functionality to WordPress. I created it after searches on keywords such as 'linguistics' in the WordPress plug-in archive returned very few results, none of which provided the desired functionality. Potential real-world applications include language learning and digital archiving of textual artefacts. The module adds options to the WordPress graphical editor, allowing the user to mark-up **<q>**quotes**</q>**, ** **paragraphs** **, **<blockquote>**blockquotes**</blockquote>**, ****spans**** (stretches of text within single paragraphs) and ** **divs** ** (document sections across multiple paragraphs) as being written in a given language. Supported Insular Celtic languages are Irish, Scottish, Manx and Canadian Gaelic; Welsh and Patagonian Welsh; Cornish; and Breton. Supported non-Insular Celtic ones are Irish, British and Canadian English; French and Canadian French; and Argentinian Spanish.

The module works by adding a **lang** attribute, which specifies the primary language of the tagged content, to the HTML tag: ** **This paragraph is in English** **, **<q lang="ga">**Is i nGaelainn athá an athfhriotal so**</q>** [this quote in Irish] et-c. The attribute's value must be a valid BCP 47 language tag, or the empty string (Faulkner, Eicholz, Leithead and Danilo, 2016). BCP 47 language tags, also called IETF tags, comprise one or more defined case-insensitive subtags, separated by hyphens. Subtags have fixed positions within the tag. Each has a maximum length of eight characters and may only include the characters A-Z, a-z and 0-9. (Phillips and Davis, 2016).

BCP 47 tags can be trivially simple, often consisting only of a **language** subtag (derived from an ISO 639 language code), as in the examples given perviously. A slightly longer form, deployed by this module, appends a **region** subtag (derived from ISO 3166-1 alpha-2 codes). For the Insular Celtic languages, the applicable values are **ga-IE** (Irish in Ireland), **gd-GB** (Scottish Gaelic in Britain), **gv-IM** (Manx Gaelic in The Isle of Man), **cy-GB** (Welsh in Britain), **kw-GB** (Cornish in Britain), **br-FR** (Breton in France), **gd-CA** (Scottish Gaelic in Canada), **cy-AR** (Welsh in Argentina). For the non-Insular Celtic ones, the values are **en-IE** (English in Ireland), **en-GB** (English in Britain), **en-IM** (English in The Isle of Man), **fr-FR** (French in France), **en-CA** (English in Canada), **fr-CA** (French in Canada) and **es-AR** (Spanish in Argentina).

The module also includes custom CSS to manage the visual presentation of these marked-up document elements in the browser, automatically highlighting languages with different colours and with shades of the same within language variations. At present, because of the questionable effects of too much conspicuous colour, this styling is only applied in the WordPress graphical editor and not in the site's public view. Uniquely and permanently styling text written in multiple dialects and languages could easily result in a particularly gross form of ransom note typography. With regard to disability and accessibility, there's also the question of how Braille readers might be expected to interpret such styling. However, the publicly presented Web document does remain fully marked-up, and is therefore parsable and linguistically searchable. It even provides speech synthesisers with a potential key for uttering strings in each language in a unique, linguistically appropriate voice.

5 The Extended Functionality (X Language annotation) module

The 'X' here is a place-holder for any of the supported Insular Celtic territory languages; extant or extinct. This module is a development of the preceding one but realised as a set of stackable sibling modules, each customised to support logically-grouped language-territory domains. Although I still consider the modules works in progress, the beta versions are stable and perfectly functional. For the

purposes of this discussion, they can be referred to as a single entity. The Irish instance will be highlighted here as it is currently the most heavily developed but commonalities will be discussed.

As with the previous module, this one works by applying the **lang** attribute, albeit with much greater specificity, and by applying presentational styling based on the attribute's values. The module's potential real-world uses are the same. Again, it was clearly necessary to work with controlled sets of attribute values but in this case some of them had to be composed ad hoc. This made it necessary to create a template for constructing them consistently. Purely for ease of reference, I have called this the **Geata Bán template**. Obviously, the starting point for constructing any such template is the full permitted BCP 47 language tag structure. This can be found in Ishida (2014):

[primary] language-extlang-script-region-variant-extension-privateuse

The **extlang** and **extension** subtags proved unsuitable for use in **Geata Bán** and were excluded. The rest were found useful, albeit to varying degrees.

5.1 Overview of the subtags used in Geata Bán

language : For this subtag to be valid it must specify one of the ISO 639 codes permitted in the Internet Assigned Numbers Authority (IANA) Registry (Ishida 2014). Codes are available for all the Insular Celtic languages, not only in their current forms but in a large number of historical ones.

script : This subtag specifies the script in which textual content is written. For it to be valid it must specify a script supported by the ISO 15924 standard (codes for the representation of names of scripts) (Ishida 2014) and must adhere to the four-character alphabetic code assigned to that script by the standard. The **script** subtag should only be used where it adds distinguishing information however (Phillips and Davis, 2016). For most of the Insular Celtic languages, the Latin script can be assumed, making it un-necessary. But as ISO 15924 also supports the Latin (Gaelic variant) script it was logical to include the subtag in **Geata Bán**, thus enhancing Irish and Scottish Gaelic instances of the module by making it possible to mark text as being set down in either the Gaelic or the Roman script. This might be particularly useful for preparing orthographically faithful digital transcriptions of original artefacts written or printed before the general migration to the Latin script.

region : For this subtag to be valid it must have a value drawn from either the UN M.49 standard, the specification of which is incompatible with the template's requirements, or from ISO 3166-1 alpha-2 (Phillips and Davis, 2016); as used by the previous module. Although ISO 3166-1 alpha-2 was found suitable for purpose, the only Insular Celtic territory recognised by it is the Republic of Ireland. This of course causes a significant loss of specificity (one that attenuated the previous module's accuracy). As specifying any other Insular Celtic territory here would violate BCP 47, it was necessary to devise a workaround using a **privateuse** subtag sequence; discussed presently.

variant : This subtag has great potential but is, for now, rather limited with regard to the Insular Celtic languages. BCP 47 specification states that « Variant subtags are used to indicate additional, well-recognized variations that define a language or its dialects that are not covered by other available subtags » (Phillips and Davis, 2016). Initially, this reference to dialects looked promising but inspection of the sole (Ishida, 2016) authoritative reference for language subtags, the IANA Language Subtag Registry (IANA, 2016) revealed that although values representing Cornish English and no less than four different Cornish orthographic standards are currently available, along with

ones for 'Scottish Standard English' and the 'Ulster dialect of Scots' (and also, interestingly, 'Scouse') there is no mention of any other dialectal variant from within the continua of the Celtic territories.

privateuse : The BCP 47 specification states (Phillips and Davis, 2016) that « Private use subtags are used to indicate distinctions in language that are important in a given context by private agreement ». But it also states that « Private use subtags are simply useless for information exchange without prior arrangement… Private use sequences… are completely opaque to users or implementations outside of the private use agreement… » It does however conceed that « …in some cases… the choice of [whether to use them] sometimes depends on the particular domain in question ». In a similar vein, Ishida (2014) states that « Because [**privateuse**] subtags are only meaningful within private agreements and cannot be used interoperably across the Web, they should be used with great care, and avoided whenever possible ».

While **Geata Bán** was never intended to support public interoperability, it is regretable that it could not be made publically accessible, within formal standards, while still offering the desired functionality. Unfortunately, a **privateuse** subtag sequence, albeit a rigorously defined one, was the only way I could find to work around the current limitations of the **variant** and **region** subtags. I did try to adopt existing standard codes and notations so that the each subtag in the sequence would at least be intuitively recognisable to third parties but even this proved problematic.

5.1.1 The privateuse subtag sequence: data structures and values

The sequence's overall structure is fairly stable, though not yet fixed, and data and values for territories, languages and dialects other than those of present-day Ireland are still rather basic. It is intended to be specific enough to carry accurate and meaningful geographic and linguistic data but flexible enough to avoid imposing inappropriate, one-size-fits-all data structures. The sequence conforms to the rules for **privateuse** subtags as given in BCP 47.
Abstracted, its structure is: **x-geataban-AAA-Bb-Cc-D(D|d)00-1111-2222**.

the x singleton : Required by BCP 47 to mark the start of a **privateuse** sequence.

geataban : A unique identifier subtag. Required by **Geata Bán** to avoid *extremely* unlikely but still not impossible clashes in real-world use. Must always have the lowercase value **geataban**.

AAA : The Celtic territory subtag. I prepared these three-letter uppercase codes to compensate for the lack of specificity in the **region** subtag: **EIR** (Éire), **ALB** (Alba), **EVN** (Ellan Vannin), **CYM** (Cymru), **KNW** (Kernow), **BRZ** (Breizh), **ANU** (Alba Nuadh), **EAP** (Eilean a' Phrionnsa), **TAE** (Talamh an Éisc) and **PTG** (Patagonia). The subtag must always have one of these values.

5.1.1.1 First- to fifth order national subdivision subtags

These provide geographic (and therefore dialectal) specificity currently unavailable in the **variant** subtag. The slightly vague nomenclature is quite deliberate as it allows for locally appropriate linguistic and territorial categorisations. I had hoped to use ISO 3166-2 codes here (representation of names of countries and their subdivisions) but these proved unsuitable, as did other standards such as FIPS 10-4, NUTS (levels 2 and 3) and Chapman.

Bb : First order national subdivisions. In the case of Ireland, these equate to its provinces. I derived two-letter PascalCase codes for these from their names: **Co** (Connachta), **La** (Laighin), **Mu** (An Mhumha) and **Ul** (Ulaidh). The subtag must always have one of these values.

Cc : Second order national subdivisions. In the case of Ireland, these equate to its counties. I took the county abbreviations given in An Brainse Logainmneacha (2007) as codes for these, adhering to the PascalCase convention used in that publication but stripping out diacritics, which would have violated BCP 47. Sample codes : **Ao** (Aontroim), **AC** (Átha Cliath), **TA** (Tiobraid Árann), **TE** (Tír Eoghain) and **UF** (Uíbh Fhailí). The subtag must always have one of these values.

D(D|d)00, **1111** and **2222** : Third- fourth- and fifth order subdivisions. For Ireland, these equate to the historical territorial units of barony, civil parish and townland. A full set of barony codes has been prepared but codes for civil parishes and townlands are still in development at time of writing.

6 Examples

Example markup, all fully interoperable up to the **x singleton**. The **privateuse** sequence then compensates for limitations in the formal standards:

<blockquote lang="br-FR-x-geataban-BRZ">A block quotation in Breton, in the Latin script, originating in France, more specifically in Brittany.**</blockquote>**

** **A paragraph in English, containing ****a span in Welsh, in the Latin script, originating in Argentina, more specifically in Patagonia****.** **

** **A paragraph in Irish Gaelic, in the Gaelic script, originating in the Irish Republic, in Ireland, in the dialect of Munster, more specifically in the dialect of County Waterford, even more specifically in the dialect of the barony of Na Déise.** **

7 Going Forward

It is intended to increase the module's linguistic specificity, to broaden its functionality (most obviously by adding granular search mechanisms) and to explore further styling techniques. It is hoped that the module may ultimately be useful to users deploying WordPress in language learning environments, on digital archiving or transcription projects, or on any online project where there is a desire or requirement to specify the language or dialect in which text is written.

The factors that caused the need for **Geata Bán** and its **privateuse** subtag sequence have less to do with BCP 47 itself than with limitations in the external ISO standards on which it draws. This situation may improve and if it does I shall update the template, and any WordPress plug-ins based on it, to migrate metadata out of the **privateuse** sequence and into the publically interoperable part of the **lang** attribute. Space limitations meant that it was only possible in this paper to give an overview of the module and template. Documentation detailing the full set of data rules and valid data values for **Geata Bán**; the Músgraí WYSIWYM WP WordPress plug-in; and working betas of the annotation (and other WYSIWYM) modules, which are free and open software, can be found online at www.gaoluinn.com/teic/bogearra/músgraí-wysiwym-wp-2/index.en.

References

BRAINSE LOGAINMNEACHA, AN. (2007). Gasaitéar na hÉireann. An Roinn Gnóthaí Pobail, Tuaithe agus Gaeltachta, [Online]. Available at http://www.logainm.ie/eolas/Data/Brainse/gasaitear-na-heireann.pdf [Accessed 14th April 2016].

BUILT WITH. (2016). CMS Usage Statistics. Built With, [Online]. Available at http://trends.builtwith.com/cms [Accessed 14th April 2016].

IANA. (2016). Language SubtagRegistry. IANA, [Online]. Available at http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry [Accessed 14th April 2016].

ISHIDA, I. (2014). Language tags in HTML and XML. W3C, [Online]. Available at https://www.w3.org/International/articles/language-tags/ [Accessed 14th April 2016].

FAULKNER, S., EICHOLZ, A., LEITHEAD, T. AND DANILO, A. (ED). (2016). HTML 5.1 Editor's Draft. W3C, [Online]. Available at http://w3c.github.io/html/dom.html#the-lang-and-xmllang-attributes [Accessed 14th April 2016].

PHILLIPS, A., AND DAVIS, M. (ED). (2016). Htags for Identifying Languages. IETF, [Online]. Available at http://tools.ietf.org/html/bcp47 [Accessed 14th April 2016].