L2/06-101

Title	Feedback on N3027 "Proposal to add medievalist characters to the UCS"
Source	UK National Body
Action	For consideration by WG2
Distribution	ISO/IEC JTC 1/SC 2/WG 2

The UK has reviewed document N3027 ("Proposal to add medievalist characters to the UCS"), which we welcome as an important contribution that reflects a genuine need by medievalists for the encoding of a range of new characters. We broadly support the aims of this proposal, and hope that agreement on the repertoire for encoding can be reached at the earliest opportunity.

However, we do have serious concerns about the encoding of the proposed insular letters D, F, R. S and T. We believe that these letters are not required in plain text, and as historical glyph variants of ordinary Latin letters they should best be catered for at a higher level using appropriate fonts.

Summary

1. Insular D and Script D for Medieval Welsh

Insular D and Script D are used by different authors as phonetic notation for the same manuscript letter corresponding to modern Welsh <dd> /ð/, and should therefore be considered as glyph variants in the context of medieval Welsh. Only one of the two proposed letters needs to be encoded, and as Script D has wider usage we believe that it is the most suitable candidate for encoding.

2. Insular Letters for Old English and Old Norse

The examples provided in N3027 clearly demonstrate that the proposed insular letters D, F, R. S and T are not required for linguistic purposes as claimed by the authors of the proposal, but are being requested simply to enable textual scholars to represent manuscript letterforms in plain text. We believe that encoding these letters would be contrary to the character-glyph model, and would cause many more problems for the electronic processing of texts written in Old English and Old Norse than it would solve. We would recommend that MUFI use a higher level protocol such as XML to represent manuscript features such as the use of Insular or Carolingian letterforms.

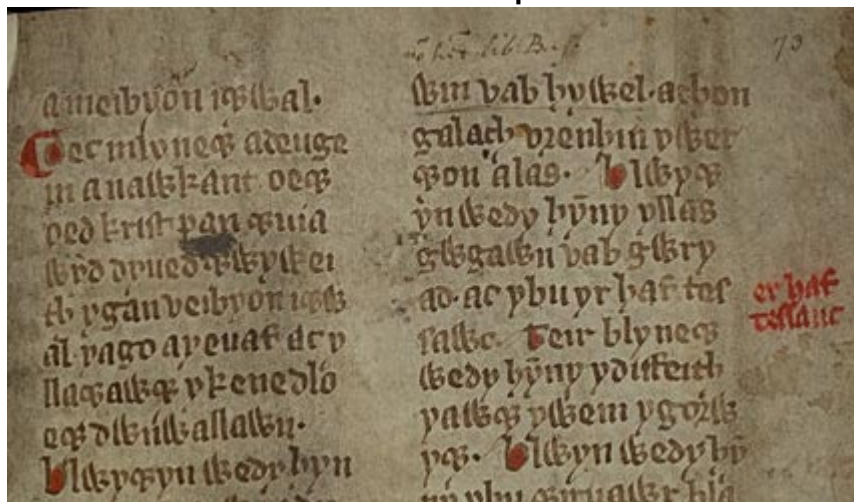## 1. Insular D and Script D for Medieval Welsh

N3027 proposes to encode both LATIN SMALL LETTER INSULAR D and LATIN SMALL LETTER SCRIPT D for use in transcribing medieval Welsh. However, they appear to be glyph variants of each other, with "script d" used by one author (J. Morris Jones, 1913) and "insular d" used by another author (Thomas Jones, 1941). In both cases the letter is used contrastively with ordinary Latin letter d, where ordinary d corresponds to modern Welsh <d> /d/ and insular or script d corresponds to modern Welsh <dd> /ð/ (this is evident from the examples given in Figs. 1, 6 and 7). As the difference between insular and script d appears to be merely typographic in these contexts, only one of the two proposed letters should be encoded. Script d seems to be the best candidate for encoding as it has wider usage as a phonetic letter (cf. N3027 Fig.41); on the other hand, in Welsh contexts, insular d appears to be the idiosyncratic usage of a single author in a single work.

One point of interest not covered in N3027 is exactly what manuscript letterform the script/insular d used by modern scholars corresponds to. In order to shed some light on this, the modern transcription of the section of "Brut y Tywysogion" (The Chronicle of the Princes), given in N3027 as an example of the use of insular d, is shown below with the corresponding portion of the manuscript on which it is based.

N3207 Fig.1 (Jones, 1941)
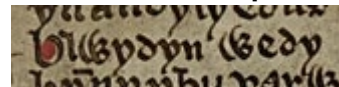

Peniarth MS 20 p.73

In the manuscript the letter corresponding to insular d in the transcription is actually the letter ▨ (this looks somewhat like the common abbreviation for Latin "que", q plus Latin Letter ET), and the letter corresponding to ordinary d in the transcription is an insular d ▨. However, elsewhere in this manuscript, as well as in other Welsh manuscripts of comparable date, <d> /d/ and <dd> /ð/ are both written with the same insular letter d, as can be seen from the words "blwydyn wedy" (modern Welsh "blwyddyn wedi") from different parts of Peniarth MS 20:
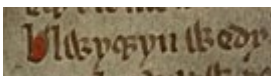
Peniarth MS 20 p.73	**Peniarth MS 20 p.264**
/d/ and /ð/ differentiated	/d/ and /ð/ not differentiated

From this it would seem that the insular/script d used by modern scholars is intended to represent the phonetic value of the manuscript letter (modern Welsh <dd>), but does not represent the physical form of the manuscript letter. This means that in Welsh usage the insular/script d is used for phonetic notation (unlike the examples of insular d used for transcribing Old Norse and Old English, which do reflect the manuscript letterform), and like other such letters (e.g. U+1D79 LATIN SMALL LETTER INSULAR G) it should be caseless.

2. Insular Letters for Old English and Old Norse

N3027 proposes to encode a number of insular letterforms (lower case insular d, f, r, s and t, and upper case insular f) for use by scholars of Old Norse and Old English. These insular letterforms are simply historical glyph variants of Latin letters, and should only be encoded if they are required for representation in plain text or are used for phonetic notation. Based upon these criteria, we do not believe that the examples provided in N3027 provide convincing evidence for the encoding of any of the proposed insular letters.

2.1 Old English Examples

Figs. 29, 30, 37, 39 and 40 in N3027 simply show Old English text set with insular style typefaces. In addition to the letters d, f, r, s and t, these examples show the use of several other uncial letterforms:

- small insular i (Figs. 29, 37, 39 and 40)
- small insular e (Figs. 29, 30 and 37)
- small insular g (Figs. 29, 30, 37, 39 and 40) -- although LATIN SMALL LETTER INSULAR G is encoded at U+1D79, it was accepted for encoding as a phonetic letter used for "older Irish phonetic notation", and is not intended for use in running text such as these examples (its use here would be comparable to using the Fraktur letters in the Mathematical Alphanumeric Symbols block to represent a German text written in Fraktur script).
- capital insular G (e.g. Gewitaþ on the 9th line of Fig.29)
- capital insular M (e.g. Mode on the 10th line of Fig.29)

From this it can be seen that encoding a selection of insular letterforms would only provide a partial solution, and that only the use of an appropriate insular style font would enable the correct rendering of the Old English examples given in N3027.

Encoding insular letterforms for purely cosmetic purposes would cause great problems for users dealing with electronic texts. For example, search operations would probably not meet user expectations (failing to find a word simply because one or more of its letters are represented by an author using an insular letterform would perplex or infuriate most users), and casing operations would be problematic. The ordinary Latin letters, as well as the letters thorn, eth, wynn and insular f would all convert from lower case to upper case and back again, but the existing insular letter g would not change case (as a phonetic letter, U+1D79 does not have any case-mapping), and it is not clear from N3027 what would happen to the proposed insular letters d, r, s and t. If insular d, r, s and t are treated like U+1D79 then they would be caseless, but if they are treated like U+017F LATIN SMALL LETTER LONG S they should have a compatibility mapping to ordinary d, r, s and t, and upper-case to ordinary D, R, S and T. Either way their behaviour would be different to insular f, for which upper and lower case forms are proposed.

Fig.39, from Ælfric's Grammar, is slightly different from the other Old English examples, as it shows Old English intermingled with Latin. However, the Old English text is all written in a typical insular script, whereas the Latin words are written with Carolingian forms of the letters d, f, r and s. Thus the insular and Carolingian letterforms are not used contrastively within the same language. Use of different script forms for vernacular and Latin text is commonplace, and no reason to encode script-specific letterforms. For example, in this late 17th century Latin work, English words are printed in a blackletter font, but no-one would suggest that because ordinary h and blackletter h are used contrastively in the same text that blackletter h needs to be encoded separately.
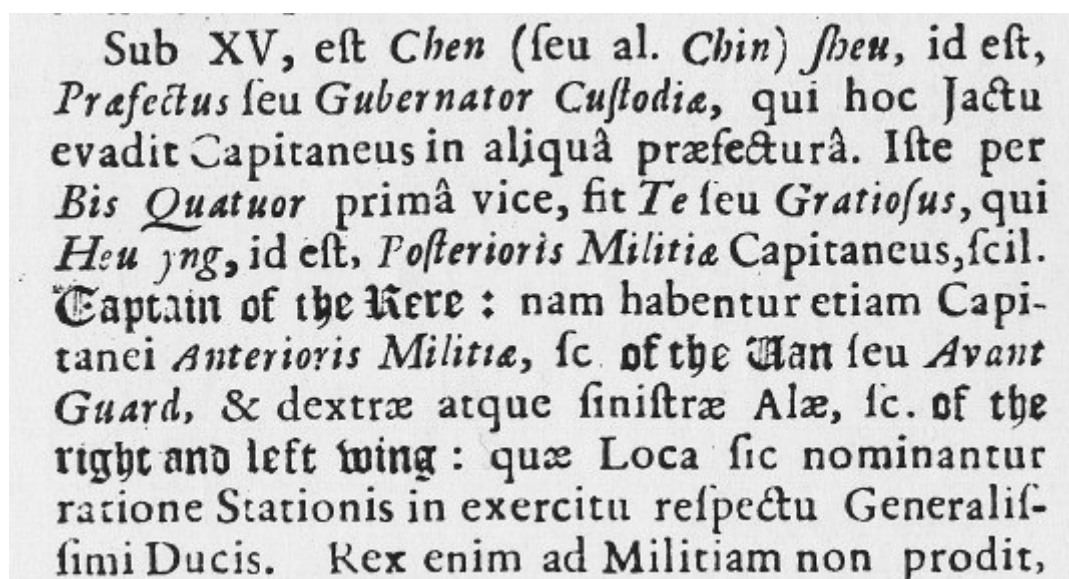

Fig.39 is essentially no different to this example, and the fact that insular letterforms are used for the Old English words and Carolingian letterforms are used for the Latin words has no bearing on the need to encode insular letterforms. The correct solution is to use different fonts for the different languages in the text.

2.2 Old Norse Examples

The Old Norse examples provided in N3027 (figs. 32, 33, 38, 42, 70, 71 and 73) are somewhat different to the Old English examples, as they only show use of one or two insular letterforms (f and/or d), and do not attempt to reproduce an insular style typeface. This is probably because Old Norse manuscripts are generally much later than Old English manuscripts, and mostly use Carolingian letterforms.

In these texts the fact that the letters f and/or d are written with insular letterforms seems to be purely cosmetic, and does not seem to add any semantic or linguistic value -- it just reflects scribal usage in the original manuscript and an overzealous desire for "authenticity" on the behalf of some modern scholars. Fig.38 shows insular f and Carolingian f used contrastively for Old Norse and Latin respectively, but this is a very weak argument in favour of encoding a separate insular letter f. The disadvantages of disunifying insular letter f from Latin letter f (such as problems with search operations) outweigh any perceived advantages. If an Old Norse scholar does want to represent insular letters f and d, then the least problematic solution is to use a font with insular letterforms for f and d.

2.3 Conclusion

N3027 claims that the proposed insular letters are "to facilitate the specific need of historical linguistic specialists to differentiate the Insular letters from the Carolingian." Analysis of the examples provided in N3027 does not support this statement, as there is no evidence that Carolingian and Insular letterforms need to be diferentiated at the plain text level, and there is certainly no indication of any need for the proposed insular letters for the purposes of historical linguistics. In all the examples given in N3027, use of an appropriate font is the correct solution.

If these insular letterforms were to be encoded, it would set a precedent for encoding other historical script variants of Latin letters. Although the proposal states that this is a closed set of letters, it is not hard to find works on paleography and epigraphy where many other historical letterforms are distinguished, and the desire of textual scholars to represent the precise glyph forms of their texts is not limited to the Latin script.

Indeed, the latest version of the MUFI character recommendation <http://gandalf.aksis.uib.no/mufi/standard/MUFI-CodeChart-2-0-f.pdf> (from which the characters proposed in N3027 are drawn) lists the following "variant letter forms [that] are not characters in the Unicode sense of the word", but which "may eventually prove to be recognised as characters according to the strict Unicode criteria." (MUFI character recommendation v.2.0 f p.171):

- LATIN CAPITAL LETTER A SQUARE FORM
- LATIN SMALL LETTER A UNCIAL FORM
- LATIN SMALL LETTER A INSULAR FORM
- LATIN CAPITAL LETTER A INSULAR FORM
- LATIN SMALL LETTER OPEN A CAROLINE FORM
- LATIN SMALL LETTER NECKLESS A
- LATIN SMALL LETTER CLOSED A GOTHIC FORM
- LATIN CAPITAL LETTER C SQUARE FORM
- **LATIN SMALL LETTER D ROTUNDA** (i.e. LATIN SMALL LETTER INSULAR D)
- **LATIN SMALL LETTER SCRIPT D**
- LATIN CAPITAL LETTER E UNCIAL FORM
- LATIN CAPITAL LETTER CLOSED E UNCIAL FORM
- LATIN SMALL LETTER E UNCIAL FORM
- LATIN SMALL LETTER E EXTENDED BAR FORM
- LATIN SMALL LETTER E TALL FORM
- **LATIN CAPITAL LETTER INSULAR F**
- **LATIN SMALL LETTER INSULAR F**
- LATIN SMALL LETTER SEMI-CLOSED INSULAR F
- LATIN SMALL LETTER INSULAR F WITH DOTTED HOOKS
- LATIN SMALL LETTER CLOSED INSULAR F
- LATIN CAPITAL LETTER G SQUARE FORM
- LATIN CAPITAL LETTER INSULAR G
- **LATIN SMALL LETTER INSULAR G** [U+1D79]
- LATIN SMALL LETTER G WITH SEPARATE LOOPS
- LATIN SMALL LETTER CLOSED G WITH LARGE LOWER LOOP
- LATIN SMALL LETTER CLOSED G WITH SMALL LOWER LOOP
- LATIN CAPITAL LETTER UNCIAL H
- LATIN SMALL LETTER H WITH RIGHT DESCENDER
- LATIN SMALL LETTER LONG I
- LATIN SMALL LETTER J DOTLESS
- LATIN SMALL LETTER K UNCIAL FORM

- LATIN SMALL LETTER K SEMI-CLOSED FORM
- LATIN SMALL LETTER K CLOSED FORM
- LATIN SMALL LETTER M WITH RIGHT DESCENDER
- LATIN CAPITAL LETTER UNCIAL M
- LATIN CAPITAL LETTER UNCIAL M WITH RIGHT DESCENDER
- LATIN SMALL LETTER UNCIAL M
- LATIN SMALL LETTER UNCIAL M WITH RIGHT DESCENDER
- LATIN CAPITAL LETTER MARIAN M
- LATIN SMALL LETTER N WITH RIGHT DESCENDER
- LATIN LETTER CAPITAL N WITH RIGHT DESCENDER
- LATIN LETTER SMALL CAPITAL N WITH RIGHT DESCENDER
- LATIN SMALL CAPITAL N WITH LEFT DESCENDER
- LATIN CAPITAL LETTER Q WITH STEM
- **LATIN CAPITAL LETTER R ROTUNDA**
- **LATIN SMALL LETTER R ROTUNDA**
- **LATIN SMALL LETTER INSULAR R**
- LATIN CAPITAL LETTER S CLOSED FORM
- **LATIN SMALL LETTER INSULAR S**
- LATIN SMALL LETTER LONG S DESCENDING
- **LATIN SMALL LETTER T ROTUNDA** (i.e. LATIN SMALL LETTER INSULAR T)
- LATIN CAPITAL LETTER T ROTUNDA
- **LATIN CAPITAL LETTER MIDDLE-WELSH V**
- **LATIN SMALL LETTER MIDDLE-WELSH V**
- **LATIN CAPITAL LETTER INSULAR V (VEND)** (i.e. LATIN CAPITAL LETTER VEND)
- **LATIN SMALL LETTER INSULAR V (VEND)** (i.e. LATIN SMALL LETTER VEND)
- LATIN SMALL LETTER X WITH LEFT DESCENDER
- LATIN SMALL LETTER Y WITH RIGHT MAIN STROKE
- **LATIN CAPITAL LETTER Y WITH LOOP**
- **LATIN SMALL LETTER Y WITH LOOP**
- **LATIN CAPITAL LETTER VISIGOTHIC Z**
- **LATIN SMALL LETTER VISIGOTHIC Z**

It will be noticed that this list includes all the insular letters proposed in N3027, and so by MUFI's own admission these letters are variant glyph forms that "are not characters in the Unicode sense of the word."

It is also clear from this list that MUFI, at least, do not consider the insular letters proposed in N3027 to be the end of the story, and if the proposed insular letters are accepted for encoding, it is reasonable to expect that further proposals to encode some or all of the remaining variant letters recognised by MUFI will be forthcoming. Indeed, the existence of MUFI will encourage scholars to use these variant letterforms in their publications, and then these publications will no doubt be cited as evidence that these letters need to be encoded.