Character Encoding and Digital Humanities in 2010 – An Insider's View

Anderson, Deborah
dwanders@sonic.net
UC Berkeley, USA

The world of character encoding in 2010 has changed significantly since TEI began in 1987, thanks to the development and adoption of Unicode (/ISO/IEC 10646) as the international character encoding standard for the electronic transmission of text. In December 2008, Unicode overtook all other encodings on the Web, surpassing ASCII and the Western European encodings (Davis 2009). As a result, Unicode's position seems to be increasingly well-established, at least on the Web, and TEI was prescient to advocate its use.

Over 100,000 characters are now defined with Unicode 5.2, including significant Latin additions for medievalists, a core set of Egyptian hieroglyphs, and characters for over 75 scripts. As such, Unicode presents a vast array of character choices for the digital humanist, so many that it can be difficult to figure out which character – if any – is the appropriate one to use. When working on a digital version of a Latin text that contains Roman numerals, should text encoder use U+2160 ROMAN NUMERAL ONE or U+0049 LATIN CAPITAL LETTER I? Should one use the duplicate ASCII characters that are located at U+FF01ff. (and why were they ever approved)? These types of questions can create confusion for text encoders.

The process of approving new characters by the Unicode Technical Committee and the relevant ISO committee is intended to be open, meaning that scholars, representatives of companies and national bodies, and other individuals may make proposals and, to a certain extent, participate in the process. Yet which characters get approved – and which don't – can still be baffling. On the TEI-list, one member wrote on 1 August 2009: "What is and isn't represented in unicode is largely a haphazard mishmash of bias, accident and brute-force normalisation. Unicode would be dreadful, if it weren't for the fact that all the alternatives are much worse."

This paper addresses the question of which characters get approved and which don't, by examining the forces at work behind the scenes, issues about which digital humanists may not be aware. This talk, by a member of the two standards committees on coded character sets, is meant to give greater insight into character encoding today, so that the character encoding standard doesn't seem like a confusing set of decisions handed down from a faceless group of non-scholars. Specific examples of the issues will be given and the talk will end with suggestions so that digital humanists, armed with such information, will feel more confident in putting forward proposals for needed, eligible characters.

The Unicode Technical Committee (UTC) is one of the two standards committees that must approve all new characters. Since it is composed primarily of industry representatives, technical discussion often predominates at meetings, including the question of whether given characters (or scripts) can be supported in current font technology and in software. For the academic, the question of whether a given character (or script) can be implemented in current fonts/software is not one commonly considered, and wouldn't necessarily be known, unless they attended the UTC meetings in person. Also, the acceptance of characters can be based on current Unicode policy or precedence of earlier encoding decisions, which again is often not known to outsiders. How to handle historical ligatures, for example, has been discussed and debated within UTC meetings, but since the public minutes of the UTC do not include the discussion surrounding a character proposal, it may appear that the UTC is blind to scholars' concerns, which is frequently not the case. In order to have a good chance at getting a proposal approved in the UTC, it is hence important for scholars to work with a current member of the UTC who can trouble-shoot proposals and act as an advocate in the meetings, as well as explain concerns of the committee and the reasoning behind their decisions.

The ISO/IEC JTC1/SC2 Working Group 2, a working group on coded character sets, is the second group that must approve characters.

This group is composed of national standards body representatives. Unlike the UTC, the WG2 is not primarily technical in nature, as it is a forum where national standards bodies can weigh in on character encoding decisions. This group is more of a "United Nations" of character encoding, with politics playing a role. Discussion can, for example, involve the names of characters and scripts, which can vary by language and country, thus causing disagreement among member bodies. Like the other International Organization for Standardization groups, decisions are primarily done by consensus (International Organization for Standardization, "My ISO Job: Guidance for delegates and experts", 2009). This means that within WG2, disagreements amongst members can stall a particular character or script proposal from being approved. For example, a proposal for an early script used in Hungary is currently held up in WG2, primarily because there is disagreement between the representatives from Austria (and Ireland) and the representative from Hungary over the name. To the scholar, accommodating national standards body positions when making encoding decisions may seem like unnecessary interference from the political realm. Still, diplomatic concerns need to be taken into account in order for consensus to be reached so proposals can be approved. Again, having the support of one's national body is a key to successfully getting a character proposal approved.

Since WG2 is a volunteer standards organization within ISO, it relies on its members to review proposals carefully, and submit feedback. Unfortunately, many scholars don't participate in ISO, partly because it involves understanding the international standard development process, as well as a long-term commitment – the entire approval process can take at least two years. Another factor that may explain the lack of regular academic involvement is that scholars participating in standards work typically do not receive professional credit. Because there is not much expert involvement in WG2 to review documents (perhaps even fewer experts than in the UTC), errors can creep in. For many of the big historic East Asian script proposals, for example, only a small handful of people are reviewing the documents, which is worrisome. The recently addition of CJK characters ("Extension C"), which has 4,149 characters, could have benefited from more scholarly review. Clearly there remains a critical need for specialists to become involved in the ISO Working Group 2, so as to prevent the inclusion of errors in future versions of the character encoding standard.

Besides the activity within each separate standards group, there are developments affecting both standards groups that may not be known to digital humanists, but which influence character encoding. New rules have recently been proposed within ISO, for example, which will slow the pace at which new characters and scripts are approved by ISO and published in *The Unicode Standard* (ISO/IEC JTC1/SC2 meeting, 2009). The new rules will severely impact new character requests. Another example of activity affecting digital projects, particularly those using East Asian characters, was the announcement in October 2009 by the Japanese National Body that it has withdrawn its request for 2,621 rare ideographs ("gaiji" characters) (Japan [National Body], "Follow-up on N3530 (Compatibility Ideographs for Government Use)", 2009), instead opting to register them in the Ideographic Variation Database, a Unicode Consortium-hosted registry of variation sequences that contain unified ideographs (Unicode Consortium, "Ideographic Variation Databse", 2009). The use of variation selectors is a different approach than that advocated in the TEI P5 for "gaiji" characters (TEI P5 Guidelines: "5. Representation of Non-standard Characters and Glyphs"), but is one that should be mentioned in future *TEI Guidelines* as an alternative.[1] In order to keep apprised of developments within the standards groups, a liaison between TEI and the Unicode Consortium (and/or ISO/IEC JTC1/SC2) would be advisable, as the activities of Unicode (/ISO) can influence TEI recommendations.

In sum, the process of character encoding is one that ultimately involves people making decisions. Being aware of the interests and backgrounds of each standard group and their members can help explain what appears to be a spotty set of characters in Unicode. Keeping up-to-date on developments within the committees can also provide insight into why

a particular character is approved or not, or why its progression has been slowed. The talk will conclude with suggestions on how digital humanists can participate more actively and effectively in the standards process.

References

Davis, Mark (May 2009). *Moving to Unicode 5.1. 5.* http://googleblog.blogspot.com/2008/05/moving-to-unicode-51.html (accessed 15 November 2009).

Unicode Consortium. *Unicode 5.2.0.* http://www.unicode.org/versions/Unicode5.2.0/ (accessed 15 November 2009).

International Organization for Standardization. *My ISO Job: Guidance for delegates and experts.* http://www.iso.org/iso/my_iso_job.pdf (accessed 15 November 2009).

ISO/IEC JTC1/SC2 meeting. Tokyo, Japan, 30 October 2009.

Japan [National Body] (16 October 2009). *Follow-up on N3530 (Compatibility Ideographs for Government Use).* http://std.dkuug.dk/jtc1/sc2/wg2/docs/N3706.doc.

Unicode Consortium. *Ideographic Variation Database.* http://www.unicode.org/ivd/ (accessed 15 November 2009).

TEI Consortium (ed.). 'P5. Representation of Non-standard Characters and Glyphs'. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.5.0.* http://www.tei-c.org/release/doc/tei-p5-doc/en/html/WD.html (accessed 15 November 2009).

Notes

1. Variation selectors have also been mentioned as being used to handle variants in other scripts, such as the historic script Tangut.