

Additional Mathematical and Letterlike Characters

		ISO/IEC JTC1/SC2/WG2 N2590
Date:		August 23, 2003
Source:		The Unicode Consortium
Status:		Liaison contribution
Action:		For consideration by WG2 in preparing amendments to ISO/IEC 10646

Summary

This is a proposal for adding seven mathematical and letterlike characters that was considered and approved at a recent Unicode Technical Committee meeting and is submitted as a liaison contribution. The text below is an updated version of a proposal submitted to the UTC by the mathematical community.

Background

Unicode 3.2, but also Unicode 3.1 and to a lesser degree Unicode 4.0 added mathematical characters to support the mathematical user community. The large number of characters involved made these additions a rather complex undertaking. During a recent review of the mathematical classification and mapping to the ISO 9573-13 entity sets for addition to Unicode Technical Report #25, *Unicode support for Mathematics*, several characters were found missing. In some cases, these can more or less directly be encoded by combining sequences, and where that was possible, they were **removed** from the request before completion of this proposal. In reviewing existing character collections, some non-mathematical letterlike characters were discovered and are proposed here for addition.


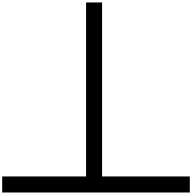
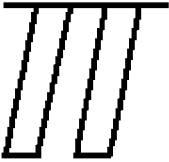

One of the goals of MathML is complete support for the SGML entity sets from ISO 9573-13. Providing this support allows existing SGML documents to be carried forward into MathML. The mapping of these entity sets has three issues




1. some entities have no reasonable character to map to
2. some entities map to a character already mapped by a different entity from the same entity set
3. some entities map to a character already mapped by a different entity from another entity set

where characters are missing or were mistakenly unified, character additions are proposed in the [list of proposed characters](#). For the other types of issues that arise in mapping ISO 9573, a final recommendation has not been made. However, the entities and characters in question are noted in http://www.unicode.org/~asmus/Notes_on_mapping_ISO_9573.html

A preponderance of existing mathematical literature is encoded in TeX format and related formats (LaTeX, etc.). TeX and its derivatives are macro languages that combine layout and glyph selection instructions directly with an entity (macro) definition. This leads to particular concerns when trying to represent existing mathematical texts in a model that is based on character encoding.

List of proposed characters

Symbol	Name / Code	Comments
	COMBINING LONG DOUBLE SOLIDUS OVERLAY suggested code: 0358	This character is requested as part of the repertoire for mathematical publications. It should look like a doubled 0338. The STIX project has the use of the following double slashed combinations attested: double-slashed: italic A, italic E, italic F
	PERPENDICULAR suggested code: 27C2	This is requested for compatibility with ISO 9573-13 as well as existing practice in TeX and LaTeX. Today, two different entities <code>perp</code> (perpendicular) and <code>bot</code> (bottom, i.e. up tack) from the <i>same</i> ISO 9573-13 entity set (ISOTECH) map to the same existing character 22A5 UP TACK. The difference between these two symbols is the way they are used: <code>Perp</code> , is an infix relation like $<$, and gets extra spacing, while <code>bot</code> is an ordinary variable. Unifying these two removes a distinction that must be expressed. LaTeX has the following definition (and plainTeX the same but less readable): <pre>\DeclareMathSymbol {\perp}{\mathrel}{symbols}{"3F} \DeclareMathSymbol {\bot}{\mathord}{symbols}{"3F}</pre> which means that <code>\perp</code> and <code>\bot</code> will by default use the same symbol, but with different white space behaviour. A " <code>\mathrel</code> " is an infix relation like $<$, and a <code>\mathord</code> is a normal letter like <code>x</code> , that gets no special spacing.
	DOUBLE-STRUCK SMALL PI suggested code: 213C	This is used by systems like Mathematica to unambiguously designate the value of pi ($= 3.14159265358979\dots$), since the ordinary Greek letter could also be used for unrelated variables. This character completes the series of double-struck Greek operators and special values found in the range U+213D..U+213F [The final glyph will be matched to the existing symbols]
	MATHEMATICAL ITALIC DOTLESS I suggested code: 1D6A4	These dotless characters are primarily intended as a compatibility character to map the ISOAMS entities <code>imath</code> and <code>jmath</code> or TeX <code>\imath</code> and <code>\jmath</code> . Most commonly, mapping these entities to the <code>mathematical italic i</code> or <code>j</code> and removing the dot when composing with math accents would result in the intended display. There are documents in which the <code>undotted i</code> and <code>j</code> are used contrastively with the dotted versions. See Additional

	<p>MATHEMATICAL ITALIC DOTLESS J</p> <p>suggested code: 1D6A5</p>	<p>information on imath and jmath symbols.</p> <p>Besides mathematical use, both dotless characters can be found in other fields, such as phonetic transcriptions, but not necessarily in their italic form.</p> <p>The <code>\imath</code> and <code>\jmath</code> are by default always italic. Their appearance in TeX (and in the ISO 9573-13 entity sets) is similar to the shapes shown in the illustrations in this proposal. It is suggested to not unify the <code>\imath</code> with the existing U+0130 DOTLESS I because <code>\imath</code> is never used in situations where case mapping occurs.</p>
	<p>LATIN SMALL LETTER DOTLESS J</p> <p>suggested code: 0237</p>	<p>Many fonts contain dotless i and j glyphs, to be used to place accents on i and j. In Unicode, placing an accent on a an i or j character removes the dot, therefore there is no need for a character to represent the dotless base character <i>unless</i> it is used standalone. Just as dotless i is used in Turkish as a standalone character, a <i>dotless j</i> is used in certain orthographies and dictionaries. See Additional information on dotless j.</p>
	<p>PER SIGN</p> <p>suggested code: 214C</p>	<p>This is a character used in print as an abbreviation for the word per, in expressions such as 'per day' or 'per month'. See Additional information on the Per sign</p>

Summary of proposed characters with suggested Unicode properties

0237 SMALL LETTER DOTLESS J; Ll; 0; L; N; ; ; ; ; ;
 0358 COMBINING LONG DOUBLE SOLIDUS OVERLAY ; Mn; 1; NSM; N; ; ; ; ; ;
 27C2 PERPENDICULAR; Sm; 0; L; N; ; ; ; ; ;
 213C DOUBLE-STRUCK SMALL PI; Ll; 0; L; N; ; ; ; ; ;
 214C PER SIGN; Ll; 0; L; N; ; ; ; ; ;
 1D6A4 MATHEMATICAL ITALIC DOTLESS I; Ll; 0; L; 0131; N; ; ; ; ; ;
 1D6A5 MATHEMATICAL ITALIC DOTLESS J; Ll; 0; L; 0237; N; ; ; ; ; ;

Other properties can be assigned by comparison with existing characters in adjacent or neighboring positions.

Additional information on the imath and jmath symbols

Generally, `\imath` and `\jmath` in TeX are simply used as base forms to apply math accents to. However,

mathematical equations can have entire sub-expressions underneath a math accent, e.g. when a 'wide hat' is placed on top of $i+j$. as in this example:

```
\widehat{\imath + \jmath} = \hat{\imath} + \hat{\jmath}.
```

In such a situation a renderer can no longer rely simply on the presence of an adjacent combining character to substitute the un-dotted glyph, and whether the dots should be removed in such a situation is not 100% predictable. In TeX, this decision is left to the author, and some authors would want to use the dotted forms as in `\widehat{\imath + \jmath}`. Authors are also known to have applied `\imath` and `\jmath` explicitly without a dot. Here is one example of an electronically published journal article making use of unaccented *dotless i* and *j*.

One can search for `\imath` and `\jmath` in the TeX source here <http://ejde.math.swt.edu/Volumes/2000/21/villatex>. Or see the result in the pdf here: <http://ejde.math.swt.edu/Volumes/2000/21/villa.pdf>

See especially the last line of Hypothesis 4.2 (b) on page 8 of the pdf which comes from this TeX source:

```
\imath \in {\bf I \ /} ( resp. \jmath \in {\bf J \ /} ).
```

The article was published in *Electronic Journal of Differential Equations* Vol.2000(2000), No. 21, pp. 1{17. ISSN 1072-6691. URL <http://ejde.math.swt.edu> or <http://ejde.math.unt.edu>, or ftp ejde.math.swt.edu ftp ejde.math.unt.edu (login ftp), which according to <http://ejde.math.swt.edu> is a fully refereed journal, with articles indexed by Math Reviews etc.

Additional information on dotless j

According to people familiar with this writing system, dotless j is used in the *handsmålfabet*

It is apparently also used for the transliteration in an important early dictionary of the the Khakas language. A relevant quote from a paper describing the method:

http://home.arcor.de/marcmarti/khakas/xakvoc/xakvoc_intro.htm

...Radloff employs two additional letters, a j without dot, and a j with comma-like dot. According to his dictionary, these graphemes represent a y preceded by a soft t and by a soft d respectively;

...

The other letter mentioned in the citation can be encoded as `j + 0313`

Unlike DOTLESS I as used in Turkish, there is no case relation for *dotless j* with a *capital letter j with dot above*. In all other respects the Unicode character properties of the proposed *dotless j* should match those of the existing DOTLESS I.

Additional information on the Per sign

The character is listed on p175 of The United States Government Printing Office Style Manual 2000 <http://www.access.gpo.gov/styleman/2000/pdf/chap10.pdf> where it is listed between the number sign and the percent sign. It can also be found in the Cambridge Encyclopedia of Language, p. 190, which reproduces a list

taken from a 1916 book (L. A. Legros & J. C. Grant, /*Typographical Printing-Surfaces*/ (London:Longmans, Green 1916)), giving "the ordinary fount of 275 characters" which has "Commercial Signs" in a row

@ [per] lb / £ \$ % + - × ÷ =

In that listing it is definitely **upper** case, in the sense that it extends from the top of the l and b to below the baseline.

Modern use in print can be found a.o. in a modern printed edition of 17th- to 19th century handwritten English letters (Miller, Kerby A., Arnold Schrier, Bruce D. Boling, & David N. Doyle. 2002. *Irish immigrants in the land of Canaan letters and memoirs from colonial and revolutionary America, 1675-1815*. Oxford, Oxford UP) where it is used to abbreviate *per* in 'per day' or 'per week'.

While the origin of this character may have been a handwritten contraction, its use in print can be considered well established.

More on the origin

The *per* sign can also be found along with other symbols used in the OED at <http://dictionary.oed.com/public/help/Advanced/symbols.htm#mod1letter> (Not all these symbols are currently part of Unicode.)

It is probably the sign indicated by the editors of *The Papers of George Washington* at <http://gwpapers.virginia.edu/search/index.html>

The ampersand has been retained and the thorn transcribed as "th."
The symbol for per (\$PR) is used when it appears in the manuscript.

Unfortunately this (\$PR) does not appear in any of the transcription or facsimile examples on the website. But at <http://www.roma.unisa.edu.au/07305/symbols.htm#Percent> part of an Italian manuscript of 1684 is shown in which an early form of the percent sign is preceded by what seems to be this same per sign. The graphic can be seen more clearly at <http://www.roma.unisa.edu.au/07305/Symbolsfolder/S406.JPG> The suspected origin of the glyph for the per sign the **p** with a bar through its descender which was the standard medieval character for "per".

See <http://www.rootsweb.com/~chevaud/abbrev.htm> for a version with a single loop, seemingly a calligraphic development of the version found at <http://www.lib.umich.edu/eebo/docs/dox/instruct.htm> called "&abper".

See also both <http://www.hum.ku.dk/ami/handbook/chapter4.htm> (and search on "persarum") and <http://helmer.hit.uib.no/mufi/proposal/range9-v2.htm> (and search on "&pbardes").

Relation to the barred P

The difference between medieval and more modern glyphs is great and **p** with a bar though the descender has also been used to indicate a fricative labial or **arf**-sound in some phonetic and transliteration traditions. For example f or p with bar above the character or below the descender is generally used today in transliterating Hebrew. Accordingly it might be best to code two symbols, **p** with a bar through the descender (with corresponding uppercase) to indicate the both medieval *per* sign and modern phonetic usage

of barred *p* and a separate character *per sign* for the more modern swirly descendant of the medieval *per* sign.

This suggestion has been raised before; here are some pointers to the mail archives of relevant discussions: There is a mention of barred-p by Robert Lloyd Wheelock at <http://www.unicode.org/mail-arch/unicode-ml/y2002-m09/0019.html> though he visualizes p with a bar through the bowl, not the descender. An answer by Jim Allan is at <http://www.unicode.org/mail-arch/unicode-ml/y2002-m09/0039.html> and has links to a number of fonts with barred characters. though only Junicode He privately reported having seen p with a bar through the stem listed in Hebrew transliteration tables, and perhaps elsewhere. Of the fonts he cites, only the Junicode fonts available at <http://www.engl.virginia.edu/OE/junicode/junicode.htm> has a *p* with a bar though the descender — which he takes to be the medieval per sign— as well as the upper case, a *P* with a bar through the stem.

Acknowledgements

Jim Allan unearthed the likely history of the per symbol and relates it to the barred p. Luka Betsch traced modern printed use of the per symbol. Alistair Vining traced it to a listing from a book from 1916. Barbara Beeton requested the double slash overlay and located the attestations of their occurrence in the literature as part of the STIX project. David Carlisle, one of the editors of MathML, submitted the request for *math* and located the additional information about their use, as well as the request for perpendicular.

.. ISO/IEC JTC 1/SC 2/WG 2

PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646

(Form number: N2352-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09))

A. Administrative

1. **Title:** _____ Additional Mathematical and Letterlike Characters _____
2. Requester's name: _____ Unicode Consortium _____
3. Requester type (Member body/Liaison/Individual contribution): _____ Liaison _____
4. Submission date: _____
5. Requester's reference (if applicable): _____ L2 03/194, L2 03/201, L2 03/136R _____
6. (Choose one of the following:)
 - This is a complete proposal: _____ yes _____
 - or, More information will be provided later: _____

B. Technical - General

1. (Choose one of the following:)
 - a. This proposal is for a new script (set of characters): _____
Proposed name of script: _____
 - b. The proposal is for addition of character(s) to an existing block: _____ yes _____
Name of the existing block: _____ several _____
2. Number of characters in proposal: _____ 7 _____
3. Proposed category (see section II, Character Categories): _____ various _____

N2590 - Additional Mathematical and Letterlike Characters

4. Proposed Level of Implementation (1, 2 or 3)
(see clause 14, ISO/IEC 10646-1: 2000): 1 and 3
Is a rationale provided for the choice? no
If Yes, reference: _____
5. Is a repertoire including character names provided? yes
a. If YES, are the names in accordance with the
'character naming guidelines in Annex L of ISO/IEC 10646-1: 2000? yes
b. Are the character shapes attached in a legible form suitable for review?
yes
6. Who will provide the appropriate computerized font (ordered preference:
True Type, or PostScript format) for publishing the standard?
Unicode
If available now, identify source(s) for the font (include address,
e-mail, ftp-site, etc.) and indicate the tools used:

7. References:
a. Are references (to other character sets, dictionaries, descriptive
texts etc.) provided? yes
b. Are published examples of use (such as samples from newspapers,
magazines, or other sources) of proposed characters attached? yes
8. Special encoding issues:
Does the proposal address other aspects of character data processing
(if applicable) such as input, presentation, sorting, searching, indexing,
transliteration etc. (if yes please enclose information)?
where applicable
9. Additional Information:
See the other sections of this document.

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? no
If YES explain _____
2. Has contact been made to members of the user community (for example:
National Body, user groups of the script or characters,
other experts, etc.)? yes
If YES, with whom? mathml working group, STIX, other experts
If YES, available relevant documents: see other sections
3. Information on the user community for the proposed characters
(for example: size, demographics, information technology use, or
publishing use) is included? yes
Reference: see other sections
4. The context of use for the proposed characters (type of use;
common or rare) varies
Reference: _____
5. Are the proposed characters in current use by the user community? yes
If YES, where? Reference: see other sections
6. After giving due considerations to the principles in *Principles and
Procedures document* (a WG 2 standing document) must the proposed
characters be entirely in the BMP? not entirely
If YES, is a rationale provided? from context
If YES, reference: see other sections
7. Should the proposed characters be kept together in a contiguous range
(rather than being scattered)? isolated
8. Can any of the proposed characters be considered a presentation form of an
existing character or character sequence? yes
If YES, is a rationale for its inclusion provided? yes
If YES, reference: see other sections
9. Can any of the proposed characters be encoded using a composed character

N2590 - Additional Mathematical and Letterlike Characters

- sequence of either existing characters or other proposed characters? No
If YES, is a rationale for its inclusion provided? _____
If YES, reference: _____
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? yes
If YES, is a rationale for its inclusion provided? yes
If YES, reference: see other sections
11. Does the proposal include use of combining characters and/or use of composite sequences (see clauses 4.12 and 4.14 in ISO/IEC 10646-1: 2000)? Yes
If YES, is a rationale for such use provided? no
If YES, reference: _____
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? _____
If YES, reference: _____
12. Does the proposal contain characters with any special properties such as control function or similar semantics? none
If YES, describe in detail (include attachment if necessary) _____
13. Does the proposal contain any Ideographic compatibility character(s)? N
If YES, is the equivalent corresponding unified ideographic character(s) identified? _____
If YES, reference: _____
-