



Universal Multiple Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation
Международная организация по стандартизации

L2/04-406

Doc Type: Working Group Document
Title: Progress Report on Mathematical Symbols
Source: Asmus Freytag, Murray Sargent, Barbara Beeton, David Carlisle
Status: Expert contribution
Action: For consideration by JTC1/SC2/WG2 and UTC
Related:

Unicode 3.2, and to a lesser extent 4.0 added a significant set of characters for mathematical use, for example with MathML. The source of these characters was derived on the one hand from research by the STIX consortium of scientific and technical publishers, and on the other hand by the desire to be able to map existing ISO SGML entity sets for technical publishing to Unicode and 10646 character codes or sequences.

Since then, there have been several more modest proposals to add mathematical characters to the Unicode Standard and 10646. Mathematical notation is evolving, therefore, in the strict sense, the repertoire of mathematical characters is not fixed. However, more than that, what drives the current set of proposals is the completion and double checking of both the STIX database and the formal efforts at mapping ISO entity sets to Unicode.

Recent papers from experts involved in implementing MathML for Arabic also point out the need for a few symbols plus specialized alphabetical forms of Arabic characters used in Arabic mathematical context.

This paper gives a progress report on the state of these review efforts and recommends some characters that should eventually be part of a formal proposal for additional characters.

Repertoire differences

Table 1 below gives the list of characters flagged in various efforts as not yet represented in Unicode. The majority of the entries come from the STIX project of reviewing the mathematical and technical literature, which means that both actual use can be attested and that the reviewers feel that publishers have an interest in being able to encode the character in question. The remainder of the entries, a much smaller number, is needed to

complete the mapping of entity sets. Finally, MathML has raised the issue of needing additional spacing clones of combining marks, due to issues with the interaction of markup syntax and isolated combining characters.

In some cases, reviewers have not been aware of characters pending for 4.1 (10646-Amd1) or, on occasion, other recently encoded characters. In these cases a tentative unification is shown. However, it is up to the reviewers to confirm that these unifications meet their needs.

Where known, an ISO entity name, TeX character name, font encoding name, or AFII identifier for the proposed characters are listed in the description column.

There are two systematic differences in repertoire between Unicode and the sets under review: one concerns the treatment of arrows and the other a reflect an issue with the use of combining marks in mathematical notation.

Differences in Arrow models

One category of shapes in the STIX list but not in Unicode is what's essentially an "arrow kit" – arrowheads in 8 directions and extenders (solid, dotted, dashed) for all. However, it is really unclear whether and how to address that; some would surely say that it is a graphics problem, but that is not a very precise definition of what the interoperability issues are, and what are standard recommendations to bridge them.

Why would one want spacing diacritics?

Diacritical marks in mathematical notation are to be visually combined with the character or term in the end, but, when using MathML they should be so combined by the renderer and not at the character/syntax level. The W3C/ISO Character model document explicitly warns against starting any XML entity with a combining character as it means that one gets differing results (potentially) depending on the order that unicode normalization or xml entity expansion is applied.

Having an entity be just a combining mark is a special case of starting an XML entity with one. The most "dangerous" would be the combining negation slash as if one had some MathML like

```
<mo>=</mo><mo>U+0338</mo>
```

Where U+0338 is a placeholder for the character here, a Unicode normalizer would combine the combining / with the > coming from the XML markup and make it non well formed.

But in general, even for other diacritics such as these dots, if being used in XML markup (as one must assume XML entity references are) one really wants (when looking at the XML markup rather than the processed document) to see the diacritic uncombined, inside its element rather than combined with its base which is probably in another element,

the mathml way of using tdot for example would be

```
<mover><mi>a</mi><mo>&tdot;</mo></mover>
```

One really does not want a combining character in that mo element.

The problem is particularly acute for U+226F NOT GREATER-THAN which has a **canonical** decomposition into > and U+0338. Normalization applied to XML source text will change the ‘>’ syntax character into > which would cause a syntax error.

Even where normalization does not change the characters, most text editors will not separate the > from the following combining mark, making editing difficult. This could be alleviated by using NCRs – except that many environments substitute character codes for them, if they can be represented in the file character set.

It is tempting to think that <mo>=</mo><mo>U+0338</mo> should be written as <mo>=U+0338</mo>. After all it is legal to have more than one character inside an <mo>, <mn>, or <mi> tag.

While this is legal, it is rare and it is not usually desirable to do math diacritics that way. Often the accent is going to be over an arbitrary math expression and it is more consistent to mark everything up the same way with the accent being explicitly positioned by an <mover>.

Rather than considering dots, which often apply to single characters, consider

$$\overline{a + b} = \overline{a} + \overline{b}$$

it would be odd (and expecting a lot of the authoring environment) to use a combining macron for the two cases on the left but an <mover> construct on the right. And even with the use of an <mover> on the right hand side one needs some way to refer to the overbar without starting the element with a combining character.

Therefore the MathML authors would like to see the addition of spacing clones of combining marks used in mathematical notation.

Other repertoire differences








The STIX review also uncovered some differences in the set of phonetic symbols, but those are traditionally excluded from any mathematically or technical symbols oriented submission on the ground that linguists should submit them.










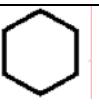


Table 1: Summary of repertoire issues for Mathematics








The ID number is simply consecutive. Shapes are relatively crude representations, however, as the shapes of these symbols in this list are simple, they are entirely sufficient for the purpose of this summary.










A **bold** character code or name refers to an existing character or character under ballot, a normal weight code and name indicate suggested values.

Where known, an ISO entity name, TeX character name, font encoding name, or AFII identifier for the proposed characters are listed in the description column.

ID	Code	Shape	Name	Description	Note
1		<no shape>		capital C with stroke	Not clear whether this can be unified with 023B, slated for 4.1, or whether it is rather a C bar, or a sans-serif C like 2201
2	213C		DOUBLE STRUCK SMALL PI	double-struck pi (lowercase) (= 3.14159..., Wolfram)	Already slated for 4.1
3	27C7		OR WITH DOT	logical or with dot inside	Contrast this to the existing U+27D1 AND WITH DOT
4	27C8		REVERSE SOLIDUS PRECEDING SUBSET	reverse solidus followed by subset = &bsolhsb; (afii DBF4)	operators are usually single characters or at best combining sequences
5	27C9		SUPERSET PRECEDING SOLIDUS	Superset followed by solidus = ⟉ (afii D95C)	operators are usually single characters or at best combining sequences
6	23xx?		STRAIGHTNESS	- "straightness" = ¯ (afii EE49)	Possibly unifiable with some existing horizontal line – but this is not a dash
7	23xx		FLATNESS	- "flatness" = ▱ (afii EE4A)	Drafting symbol. Generally not unifiable with 25B1 Parallelogram
8	23xx		AC CURRENT	- ac current = ∿ (afii DB3B)	While 223F may be used to express AC semantics, as a symbol its not unifiable with this character
9	23xx		ELECTRICAL INTERSECTION	- electrical intersection = ⏧ (afii DB4E)	

ID	Code	Shape	Name	Description	Note
10	2B14		SQUARE WITH UPPER RIGHT DIAGONAL HALF BLACK	- square, filled top right corner = &sqvarftr; (no afii)	See U+25E9 SQUARE WITH UPPER LEFT DIAGONAL HALF BLACK
11	2B15		SQUARE WITH LOWER LEFT DIAGONAL HALF BLACK	- square, filled bottom left corner = &sqvarfbl; (no afii)	See U+25EA SQUARE WITH LOWER RIGHT DIAGONAL HALF BLACK
12	2B16		DIAMOND WITH LEFT HALF BLACK	- diamond, filled left half = &diamonfl; (no afii)	
13	2B17		DIAMOND WITH RIGHT HALF BLACK	- diamond, filled right half = &diamonfr; (no afii)	
14	2B18		DIAMOND WITH BOTTOM HALF BLACK	- diamond, filled bottom half = &diamonfb; (no afii)	
15	2B19		DIAMOND WITH TOP HALF BLACK	- diamond, filled top half = &diamonft; (no afii)	
16	2B20		WHITE TRAPEZIUM	- trapezium = ⏢ (afii DBB8)	
17	2B21		WHITE PENTAGON	- open pentagon (afii DB2D)	
18	23xx		BENZENE RING WITH CIRCLE	- benzene ring [hexagon] with circle = &benzenr; (afii D8DC)	This is a variant of the benzene ring at 232C, but it should not be unified. (see note at end)
19	2B22		WHITE HEXAGON	- benzene ring [open hexagon] = &benzen; (no afii)	
20	2B23		BLACK HEXAGON	- filled hexagon	
21	2B24		HORIZONTAL BLACK HEXAGON	- horizontal filled hexagon	

ID	Code	Shape	Name	Description	Note
22	Tbd	Tbd	Tbd	- some extra sizes of squares; still haven't found incontrovertible examples of missing ones in context (beeton)	Literature search ongoing, squares already have a wide range of sizes, so we may be covered.
23	Tbd	Tbd	Tbd	- some extra sizes of circles	These should be covered by medium and medium small circles that are slated for 4.1
24	Tbd	Tbd	Tbd	- one extra size of lozenge	This had been reported a while ago, need to locate the references
25	26A5 ?	<no shape available>	MALE AND FEMALE SIGN	- hermaphrodite = &hmphdite; (no afii)	This may be unifiable with 26A5 to be added in 4.1. Rename to HERMAPHRODITE?
26	26xx		NEUTER	- neuter [circle with short vertical below]	While the semantics of neuter can be represented with MEDIUM WHITE CIRCLE, this symbol cannot be unified with 26AA
27	20EC		COMBINING ANGLE ABOVE LEFT	- combining left overangle	The above right would be the ANNUITY BEND at 20E7 (presumably) . Size and style of 'angle' need to be confirmed
28	20ED		COMBINING ANGLE BELOW LEFT	- combining left underangle	Size and style of 'angle' need to be confirmed
29	20EE		COMBINING ANGLE BELOW RIGHT	- combining right underangle	Size and style of 'angle' need to be confirmed
30	20EF		COMBINING RIGHTWARDS HARPOON WITH BARB DOWNWARDS	- combining over right harpoon down	Compare U+21C1 RIGHTWARDS HARPOON WITH BARB DOWNWARDS
31	20F0		COMBINING LEFTWARDS HARPOON WITH BARB DOWNWARDS	- combining over left harpoon down	Compare U+21BD LEFTWARDS HARPOON WITH BARB DOWNWARDS
32	20EC?		COMBINING LEFT ARROW BELOW	- combining under left arrow	

ID	Code	Shape	Name	Description	Note
33	20ED?		COMBINING RIGHT ARROW BELOW	- combining under right arrow	
34	27Dx		WIDE ANGLE	dwangle	
35	23B4		TOP BRACKET	Tbrk	Already encoded as 23B4
36	25F8		UPPER LEFT TRIANGLE	Ultri	Already encoded as 25F8
37	25F9		UPPPER RIGHT TRIANGLE	Urtri	Already encoded as 25F9
38	299A ? 2307?		<VERTICAL ZIG ZAG> <WAVY LINE>	vzigzag	How does this relate to 299A Ṿ VERTICAL ZIG-ZAG LINE and 2307 Ẓ WAVY LINE
39	27Dx		EQUAL OR PARALLEL	parallel, equal; equal or parallel (AFII DB4F)	Epar
40	1D7CA?	F	MATHEMATICAL BOLD CAPITAL DIGAMMA	b.Gammad(9573-2003-isogr4)	U+03DC is mapped to Gammad (9573-2003-isogr3):
41	1D7CB?	f	MATHEMATICAL BOLD SMALL DIGAMMA	b.gammad(9573-2003-isogr4)	U+03DD is mapped to gammad(9573-2003-isogr3):
42	27Cx		EQUIVALENT BY DEFINITION	Equivalent by definition (double arrow with 225D $\stackrel{\text{def}}{=}$ def)	Proposals forthcoming in context with Arabic proposals
43	20Ex		COMBINING BY DEFINITION	Combining 'def'	Proposals forthcoming in context with Arabic proposals
44	Various		Various	Arabic symbols equivalent to 225D and the two preceding	Proposals forthcoming
45	Various		Various	Various Arabic ligatures in lieu of !, Σ etc.	Proposals forthcoming
46	Various		Various	Various special letterform for Arabic symbols	Proposals forthcoming
47	Various		Various	One or two new misc. symbols	Proposals forthcoming

ID	Code	Shape	Name	Description	Note
				used in Arabic context	
48	Various		Various	Ryummy numbers	Proposals forthcoming

Notes:

The symbol for benzene

The Kekulé structure for benzene, with its alternating single and double bonds is the reference glyph for 232C BENZENE RING. Some authors prefer it, but many others deliberately replace it by the more modern symbol, shown here and in Table 1, which shows the ring of six carbon atoms, each of which has one hydrogen attached. While hydrogen and carbon atoms are implied by the corners of the diagram in the usual manner, it is essential to include the circle as it represents the delocalized electrons. Without it, the symbol represents cyclohexane and not benzene.



Unlike the Kekulé structure, it is not possible to deduce the number of hydrogen atoms from the benzene symbol with the circle. On the other hand, the chemical bonding of Benzene is quite different from a series of alternating single and double bonds as suggested by the Kekulé structure. This is because the electrons are delocalized due to a process called resonance.

While both forms of the symbol unambiguously represent the same chemical molecule, it appears that the choice of the particular representation is often quite deliberate, as each symbol emphasizes different aspects of the structure. Even a cursory examination of the subject will lead to paper where authors give and defend opposite preferences, and almost all introductory texts indeed present both symbols, until establishing a convention.

These two forms should therefore be disunified. Unlike the differences in shape captured by variation sequences for mathematical symbols, the differences in shape and identifiable motivation in usage seem pronounced enough that there would be little benefit over adding a separate character.

References

MathML

Arabic Proposals (forthcoming)