L2/99-045

DATE: 1999-02-04

DOC TYPE:	Expert contribution
TITLE:	Proposal to encode mathematical alphanumeric symbols
SOURCE:	Murray Sargent III, Barbara Beeton
PROJECT:	
STATUS:	Proposal
ACTION ID:	FYI
DUE DATE:	--
DISTRIBUTION:	Worldwide
MEDIUM:	Paper and html
NO. OF PAGES:	5

A. Administrative

1. Title	Proposal to encode mathematical alphanumeric symbols
2. Requester's name	Murray Sargent III, Barbara Beeton
3. Requester type	Expert request.
4. Submission date	1998-12-01
5. Requester’s reference	Scientific and Technical Information Exchange (STIX)
6a. Completion	Complete proposal
6b. More information to be provided?	If requested

B. Technical -- General

1a. New script? Name?	No.
1b. Addition of characters to existing block? Name?	No.
2. Number of characters	14 variants or 1082 new alphanumeric symbols
3. Proposed category	
4. Proposed level of implementation and rationale	Level 3 since math variant tags qualify the base letter they follow
5a. Character names included in proposal?	14 variant tags are defined. Recommended

	to reserve a block of 16. Alternatively 1082 new alphanumeric symbols
5b. Character names in accordance with guidelines?	Yes.
5c. Character shapes reviewable?	
6a. Who will provide computerized font?	None needed
6b. Font currently available?	None needed
6c. Font format?	na
7a. Are references (to other character sets, dictionaries, descriptive texts, etc.) provided?	Yes.
7b. Are published examples (such as samples from newspapers, magazines, or other sources) of use of proposed characters attached?	Not attached, but available.
8. Does the proposal address other aspects of character data processing?	No

C. Technical -- Justification

1. Contact with the user community?	Yes. Patrick Ion, Barbara Beeton, Murray Sargent III
2. Information on the user community?	Professional mathematicians, physicists, astronomers, engineers, and other scientific and technical researchers.
3a. The context of use for the proposed characters?	Used in publication of research mathematics and other hard sciences.
3b. Reference	
4a. Proposed characters in current use?	Yes.
4b. Where?	Worldwide, by scientific and technical publishers.
5a. Characters should be encoded entirely in BMP?	Yes.
5b. Rationale	Accurate publication of mathematical and scientific research on the Web is impossible without a comprehensive and accurate collection of symbols including various

	alphabetic variants in common use. Allocation in the BMP is in accordance with the Roadmap for nonalphanumeric symbols.
6. Should characters be kept in a continuous range?	Yes
7a. Can the characters be considered a presentation form of an existing character or character sequence?	No. A given alphabetic symbol has different semantics when its style is changed and should not be found by the same plain-text search string.
7b. Where?	
7c. Reference	
8a. Can any of the characters be considered to be similar (in appearance or function) to an existing character?	No
8b. Where?	
8c. Reference	
9a. Combining characters or use of composite sequences included?	Yes if variant tags are used
9b. List of composite sequences and their corresponding glyph images provided?	A list is provided below, but the corresponding glyphs are well known and are omitted.
10. Characters with any special properties such as control function, etc. included?	All the characters are modifier characters, which is a kind of control nature.

D. SC2/WG2 Administrative

To be completed by SC2/WG2

1. Relevant SC 2/WG 2 document numbers:	
2. Status (list of meeting number and corresponding action or disposition)	
3. Additional contact to user communities, liaison organizations etc.	
4. Assigned category and assigned priority/time frame	
Other Comments	

Mathematics has need for a number of Latin and Greek alphabets that on first thought appear to be font variations of one another, e.g., normal, bold, italic and script H. However in any given document, these characters have distinct mathematical semantics. For example, a normal H represents a different variable from a bold H, etc. If one drops these distinctions in plain text, one gets gibberish. Instead of the well-known Hamiltonian formula

$$\mathsfit{H} = \int \mathrm{d}\tau(\varepsilon E^2 + \mu H^2),$$

you’d get the integral equation (!)

$$H = \int \mathrm{d}\tau(\varepsilon E^2 + \mu H^2).$$

Accordingly, the STIX project requests adding normal, bold, italic, script, etc., Latin and Greek alphabets. Straight encoding would amount to many characters and would lose some useful common information, such as all variants of H might not be recognizable as H’s. But it does allow plain text to retain the proper character semantics and it allows simple (nonrich) search methods to work.

The proposal considers two possible ways to encode mathematical alphabetic symbols in ways that allow simple search algorithms to work. One encodes them all outright in one of the surrogate planes. The other uses “math variant tags”, which act in some ways like nonspacing combining marks. For example, a math script H would be encoded as H<math script>. Encountering such a combination, a rendering engine should choose some script font to render the H. Which script font is beyond the scope of plain text.

The alphabetic symbols encountered in mathematics are given in the following table:

Math style	**Characters**	**Count**
upright (Roman)	a-z, A-Z, 0-9, α-ω, A-Z (Greek)	113
italic	a-z, A-Z, 0-9, α-ω, A-Z (Greek)	113
bold	a-z, A-Z, 0-9, α-ω, A-Z (Greek)	113
bold italic	a-z, A-Z, 0-9, α-ω, A-Z (Greek)	113
calligraphic (script)	a-z, A-Z	52
bold calligraphic (script)	a-z, A-Z	52
fraktur	a-z, A-Z	52
open-face	a-z, A-Z, 0-9	62
open-face italic	a-z, A-Z, 0-9	62
sans-serif	a-z, A-Z, 0-9	62
sans-serif italic	a-z, A-Z, 0-9	62
sans-serif bold	a-z, A-Z, 0-9, α-ω, A-Z (Greek)	113
sans-serif bold italic	a-z, A-Z, 0-9, α-ω, A-Z (Greek)	113
monospace	a-z, A-Z, 0-9, α-ω, A-Z (Greek)	113

α-ω also includes a glyph variant of ϕ, ε, and θ, since both glyphs for each of these can appear in the same document with different semantics. This gives 24+3 lowercase characters, while only 24 upper-case Greek characters are used.

Math Variant Tag Approach

Math variant tags could be defined for each of the categories above. If this approach is used, the tags should have the following properties:

1. no effect at start of line
2. combining class 255
3. affect entire previous combining character sequence
4. only the first of consecutive tags is used (others are ignored)

Outright Encoding Approach

Outright encoding would use the corresponding BMP characters for upright (Roman). The remaining characters would be stored in ASCII order in higher planes for a total of 1082 characters as currently entered (some digit ranges may be deleted on further examination). No accented characters are included. Accented mathematical symbols are always represented by combining character sequences.

Discussion

Rendering mathematics requires a fairly sophisticated 2D layout engine. Compared to the complexity needed in this engine, handling either math variant tags or surrogate pairs is straightforward. The outright encoding approach is simpler conceptually, since it doesn't require any additional rules. It only requires the ability of the display engine to handle surrogate pairs. Both approaches require multicode navigation, e.g., the text caret must not end up inside a multicode sequence. Such navigation is easy to implement if combining-mark sequences are implemented. Since combining mark sequences are required in mathematics, this navigation shouldn't be difficult to implement.