Unicode and math, a combination whose time has come — Finally!

Barbara Beeton
American Mathematical Society
bnb@ams.org

Abstract

To technical publishers looking at ways to provide mathematical content in electronic form (Web pages, e-books, etc.), fonts are seen as an "f-word". Without an adequate complement of symbols and alphabetic type styles available for direct presentation of mathematical expressions, the possibilities are limited to such workarounds as `.gif` and `.pdf` files, either of which limits the flexibility of presentation.

The STIX project (Scientific and Technical Information eXchange), representing a consortium of scientific and technical publishers and scientific societies, has been trying to do something about filling this gap. Starting with a comprehensive list of symbols used in technical publishing, drawn from the fonts of consortium members and from other sources like the public entity sets for SGML as listed in the ISO Technical Report 9573-13, a proposal was made to the Unicode Technical Committee to add more math symbols and variant alphabets to Unicode. Negotiations have been underway since mid-1997 (the wheels of standards organizations grind exceedingly slowly), but things are beginning to happen.

This paper will share the latest information on the progress of additional math symbols in Unicode, and the plans for making fonts of these symbols freely available to anyone who needs them.

Introduction

The composition of mathematics has never been straightforward; it has always required special fonts above and beyond the alphabetic complement required for text. Even if an author makes an effort to describe mathematical concepts and relationships in words, there comes a point where symbols become necessary for both clarity and conciseness. In some fields (for example, symbolic logic), the use of notation has expanded to such a degree that it is nearly impossible to express concepts clearly in ordinary words; symbols convey the desired meaning much more directly. The situation might be compared to that of two literate Chinese from different areas meeting, and communicating by writing rather than in their different spoken dialects. There is sometimes just no reasonable substitute for a common writing system.

Although symbols form a large and important part of written mathematics, mainly indicating operations, relations, and other similar concepts, alphabets are also co-opted from their role of representing ordinary language to provide the notation for mathematical constants, variables and functions — the things operated on. The number of different alphabets used in some documents appears to be limited only by what is available or by the capacity of the typesetting system (manual, mechanized or electronic). Only numerals seem to denote more or less the same kinds of concepts in both ordinary prose and mathematical notation. Needless to say, font foundries have never been overly eager to provide an unlimited supply of new symbol shapes of arcane design and often intricate production requirements.

Complicating this situation is the fact that the audience for typeset mathematics is relatively small. If the number of mathematicians clamoring for competently printed material in their subject were anywhere near the number of readers of novels or sports magazines, or if these mathematicians had budgets matching those of major advertising agencies, font foundries could muster much greater interest in doing this sort of work.

Terminology

When one looks at a printed page, one sees that it is constructed from many small elements. There are letters, digits, punctuation, symbols, dingbats, ...

One might think to refer to all of these as *characters*. The dictionary [9] definition of *character* is, in part,

> **character** ... **1**. A sign or token placed upon an object as an indication of some special fact, as ownership or origin; a mark, brand, or stamp. **2**. Hence: **a** a graphic symbol of any sort; esp., a graphic symbol employed in recording language, as a letter. **b** Writing; printing. **c** ...

Clear enough? Well, not quite.

In standardese, a term can have only one meaning. The basic ISO[1] definition [5] is

> **character** A member of a set of elements used for the organisation, control, or representation of data.

Thus the term *character* cannot be used in an ISO standard with any other meaning.

Another relevant term is *code*; from the same dictionary [9]:

> **code** ... **3**. A system of signals for communication by telegraph, flags, etc. (...); also, a system of words or other symbols arbitrarily used to represent words; as, a secret *code*.

This is the term adopted to identify the system by which data is stored in a computer memory, and the individual elements are known as *coded characters*, or *characters* for short. Different computer coding systems use different bit patterns to represent the same character; for example, the letter A would have a different code in ASCII, BCD, EBCDIC, ISO 646, ISO 8859-1, etc., but in each of these systems, A is still considered the same character. If it is in a context that might (in print) be represented in italic or boldface, that makes no difference; the same code is used for all.

But an A in a font or on a printed page is not (by this system) a character, and an italic A is different from a boldface A, and so on. The term adopted in standardese [3] for such an element is *glyph*:

> **glyph** A recognizable abstract graphic symbol which is independent of any specific design.

Thus it is clear that one code may represent many different glyphs. The reverse is also true: while the word "file" is spelled with four letters, and coded as four characters, when printed with a font that has ligatures, only three glyphs are used.

The association between characters and glyphs is referred to as *mapping*. What is important here is that in neither direction is the mapping between a coded character and a glyph one-to-one; it may be one-to-many, or many-to-one. While this is not only adequate, but even admirable, when dealing with text, for mathematics it can introduce serious ambiguity.

[1] International Organization for Standardization

Codes

The alphanumeric soup of standardized codes has already been mentioned. Consider the history of codes used for computer input.

Although quite a few different models of digital computer architecture have been devised, very few have been based on a wider range of possibilities for the smallest element other than zero and one — on and off. (This provides the rationale for naming the bit, "binary digit".) Different combinations of bits, in strings of predefined length, designate characters. The number of bits in such a string is the limiting factor in how many characters can comprise a code.

Very early codes contained six bits — 64 characters, just enough for a single-case (latin) alphabet, ten digits, five arithmetical operators (`+`, `-`, `*`, `/` and `=`), the punctuation required to format real numbers and accounting data, and a number of *control codes* to support interaction with a Teletype machine. The following symbol complement, a variant of the BCD code, was available on a punched paper tape device used in the 1970s at AMS; symbols in the second row had to be preceded by an upshift and followed by a downshift, as they were piggy-backed onto other characters.

```
. , - / * # & $
: ; + = @ ( ) < > ' "
```

This was sufficient to support Fortran, Cobol, and other antique programming languages, but not a direct visual representation of mathematical expressions.

ASCII, in its original form, had seven bits and 128 characters, which could accommodate a lowercase alphabet and more symbols. (This is the code under which TEX was first implemented.) ISO 646 is the "international" version of ASCII. A key principle was — and is — that once a character gets into a code, it is never removed, so the current 8-bit ASCII is backward compatible with the 7-bit version, at least insofar as what can be encoded.

Other codes have been promulgated by manufacturers, national standards bodies, and the ISO. Until the mid-1980s, these codes were used almost exclusively to support processing of programming languages and natural languages. Whatever symbols were included were necessary to specify programming operations, not the symbolic representation of scientific disciplines, and typically, except for

the few symbols and punctuation characters that were already present in six-bit codes, symbolic characters were typically segregated in programming language–specific codes such as the one for APL.

Some language codes already exceeded the typical eight-bit capacity of 256 elements. It is impossible, for example, to fit in all the accented and variant letters of the alphabet needed to represent all the languages based on the latin alphabet. And codes for Japanese and Chinese had to accommodate the nearly 10,000 characters used to publish newspapers, or, preferably, the 50,000 characters or more found in literary works.

The development of a multi-byte code, ISO 10646 (originally a two-byte code), undertook to combine in a single code all existing national and commercial codes. Computer manufacturers and other commercial organizations dependent on computer technology became dissatisfied with the progress of the ISO working group responsible for standardizing codes, and, in 1988, formed the Unicode Consortium for the purpose of creating a unified international code standard on which new multinational computer technology could be based. The ISO old guard was joined or replaced by the Unicode members, and since 1991 Unicode and ISO 10646 have been parallel.

The content and structure of Unicode

In the Unicode 2.0 manual [7], the section *Design goals* identifies some of the gaps in coverage by existing codes.

> When the Unicode project began in 1988, groups most affected by the lack of a consistent international character standard included the publishers of scientific and mathematical software, newspaper and book publishers, bibliographic information services, and academic researchers. ... The explosive growth of the Internet has added to the demand for a character set standard that can be used all over the world.

The first iteration of Unicode included "characters from all major international standards published before December 31, 1990". One of these was the SGML standard [2], which contained a sizeable list of mathematical and technical symbols in its original Annex A (this list was later moved to a technical report [4]). Other sources included "bibliographic standards used in libraries ..., the most prominent national standards, and various industry standards in very common use".

In the Unicode 3.0 manual [8], only one reference can be unambiguously associated with math symbols: ISO 6862, *Information and documentation — Mathematics character set for bibliographic information interchange* (no explicit references are shown in the Unicode 2.0 manual). Many of the symbols listed in the annex to the SGML standard don't appear in Unicode. More about this later.

There are several design principles especially relevant to the designation of math symbols as characters [8]:

- The Unicode Standard encodes characters, not glyphs.
- Characters have well-defined semantics.
- The Unicode Standard encodes plain text.

The implication is that the meaning of each character is distinct, so that the representation when interchanged or typeset will be unambiguous. More about this later as well.

Unicode is organized into segments of 65,536 characters called *planes*. The first of these, plane 0, is the *basic multilingual plane* (BMP). Within this plane, characters with common characteristics are grouped into blocks, usually of 256 characters. The first full block is equivalent to *Latin 1*, with the first half comprising 7-bit ASCII. The code for any character assigned to the BMP can be represented by 16 bits, a two-byte, or *two-octet* code. The formal representation of such a code is "U+*xxxx*", where *xxxx* is a string of four hexadecimal digits.

Within the BMP, these blocks are occupied by symbols:

- U+2000–206F: General punctuation
- U+2070–209F: Subscripts and superscripts
- U+20A0–20CF: Currency symbols
- U+20D0–20FF: Combining diacritical marks for symbols
- U+2100–214F: Letterlike symbols
- U+2150–218F: Number forms
- U+2190–21FF: Arrows
- U+2200–22FF: Mathematical Operators
- U+2300–23FF: Miscellaneous technical (in Unicode 2.0, U+2380–23FF are unassigned, reserved for later additions)
- U+2400–243F: Control pictures
- U+2440–245F: Optical Character Recognition
- U+2460–24FF: Enclosed alphanumerics
- U+2500–257F: Box drawing
- U+2580–259F: Block elements
- U+25A0–25FF: Geometric shapes
- U+2600–267F: Miscellaneous symbols

- U+2700–27BF: Dingbats
- U+27C0–27FF: (unassigned)
- U+2800–28FF: Braille patterns (added in Unicode 3.0)
- U+2900–2DFF: (unassigned)

Symbols that were part of earlier codes are kept with those codes in other blocks; if a code already existed, the character was not duplicated.

A segment of the BMP has been set aside for *private use*, where characters may be assigned which are not formally included in Unicode but for which an agreement exists between sending and receiving users.

Up to now, most character assignments are in the BMP, with the intention that they be easily accessed. However, for less frequently occurring characters, work has begun to populate Plane 1. This is another area with relevance that will become obvious later.

Identifying symbols required for math typesetting

Early in 1997, a group of scientific and technical societies and publishers banded together under the name STIPub — Scientific and Technical Information Publishers — to address matters of common interest. The founding members of this group were

- American Chemical Society
- American Mathematical Society
- American Institute of Physics
- American Physical Society
- Elsevier Science, Inc.
- Institute of Electrical and Electronic Engineers

One topic of growing concern to the STIPub members was how best to move into the Internet age, to make use of the World Wide Web as an adjunct, if not the new centerpiece, of their publishing efforts. A major obstacle facing Web publication was — and is — the pitifully inadequate symbol set available with HTML, and its lack of support for the two-dimensional positioning of mathematical notation. Several attempts at providing some support for this material had been brushed aside as successive releases of the HTML Recommendation[2] added features to improve the visual presentation and control of document layout.

It was understood that a future version of most browsers would include "Unicode support". Although it is unclear exactly what is meant by this, an obvious course of action was to make certain that Unicode coverage for math and technical notation is complete.

A working group for Scientific and Technical Information eXchange (STIX) was formed with the charter to identify the required symbol complement and get the missing elements incorporated into Unicode. A first step was to collect from the STIX participants and other sources lists of symbols currently in use, and to reduce this to two subcollections: symbols already in Unicode and symbols not in Unicode. Information was gathered from the following sources, in addition to the STIX members:

- the entity sets of ISO TR 9573-13 [4]; electronic files were provided by the editor, Anders Berglund.
- fonts designed to be used with TEX: Computer Modern, AMSFonts, Lucida New Math, `lasysym`, St. Mary Road, `wasysym`
- Wolfram Research (Mathematica)
- Justin Ziegler's LATEX 3 project report [10]
- Taco Hoekwater (for Kluwer Academic Publishers)
- Jörg Knappen (for Springer Verlag)
- Paul Topping, Design Science, Inc. (MathType)
- the ISO Z language standard (ISO CD 13568)
- various requests for specific symbols identified through AMS technical support and the newsgroup `comp.text.tex`

More than 2200 distinct symbols were identified.

The next step was to determine which were already included in Unicode. As already mentioned, not all symbols are located in the blocks designated for symbols; some have codes in other blocks, including Latin 1, Greek, and even among the CJK[3] symbols and punctuation. About half of the symbols in the collection were found in the Unicode 2.0 manual [7]; the remainder were assigned provisional identifiers in the Unicode private use area, and a table was constructed, listing the following for each symbol:

- its ID
- a possible cross-reference to an ID for another symbol of similar shape or meaning
- the AFII[4] glyph identifier
- the entity name, TEX code, or other identifying information for each contributor
- a brief description

[2] A Recommendation is the World Wide Web Consortium's (W3C) equivalent of an international standard.

[3] Chinese, Japanese and Korean

[4] Association for Font Information Interchange

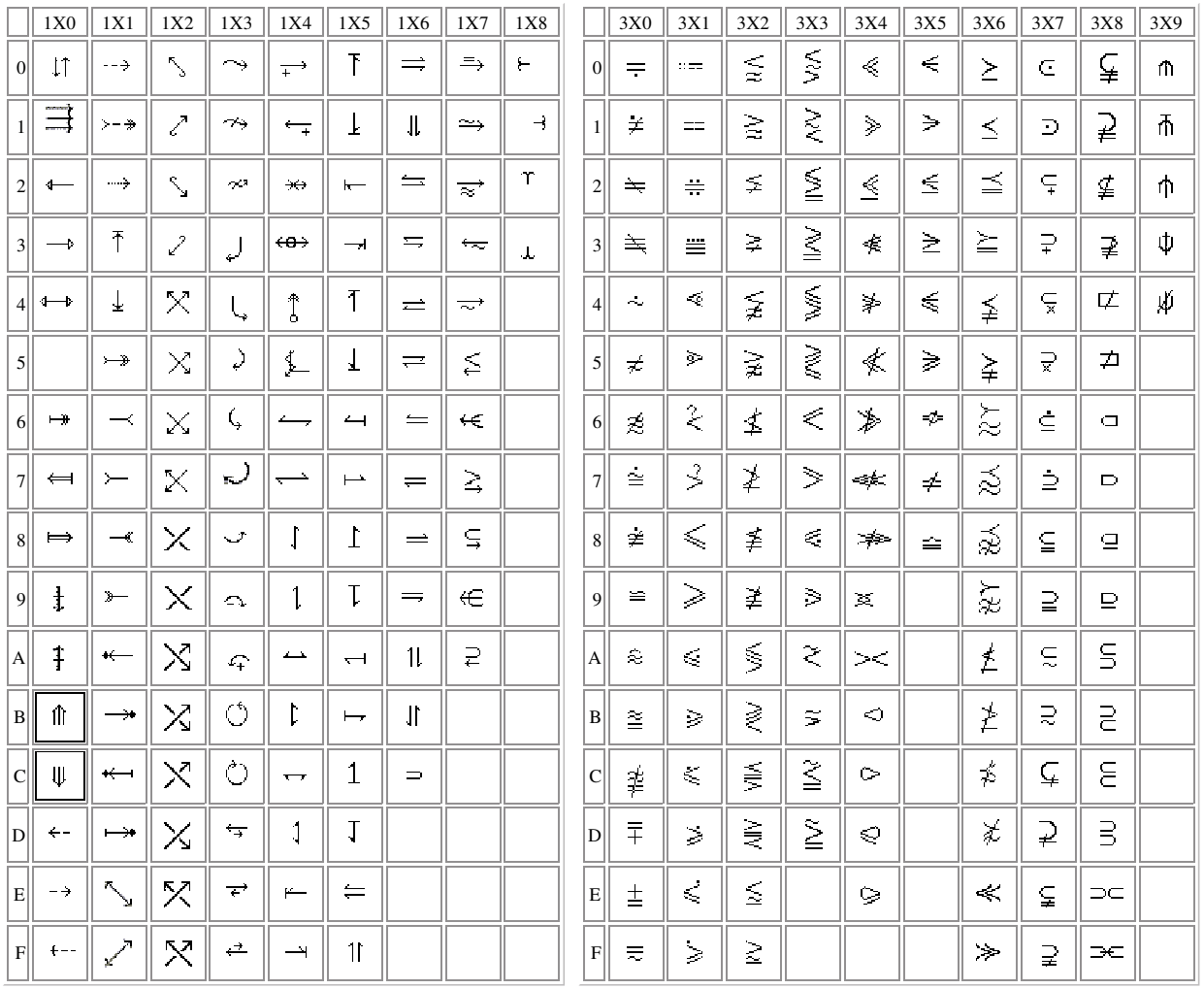

Figure 1: First (arrows) and third (binary relations) of seven symbol tables in the December 1998 version of the Unicode math proposal

This completed the first half of the project. The second, more difficult, half remained — putting the request into a form that would be acceptable to the Unicode Technical Committee (UTC).

Preparing the Unicode proposal

It is a UTC requirement that any request for encoding must be accompanied by a sample image of the requested character. This posed some problems. For many of the symbols to be submitted, no fonts were available. We solved that problem in a rudimentary way by creating GIF images at a very low resolution; for symbols we did have in fonts, this was accomplished using latex2html, and for others, bitmaps were created by hand and packaged as GIFs. The resulting images, packaged in HTML tables, were rough but recognizable. For the initial version of the proposal, the order of symbols in the tables was the same as that in our master list, by reference ID; the full collection of tables and text — about 70 files in all — was sent to the UTC in March 1998.

The arrangement of symbols was semi-random, and individual symbols were hard to find, so at the UTC's request, the proposal was reorganized and the first revision delivered in December 1998. Figure 1 shows two of the seven tables in this first revision.

Once the symbols were rearranged, the presence of similar shapes — considered by the UTC to be possible duplicates — became obvious.

The UTC takes its job seriously [8, p. 17]:

> The Unicode Standard avoids duplicate encoding of characters by unifying them within scripts across languages; characters that are equivalent in form are given a single code.

> Common letters, punctuation marks, symbols, and diacritics are given one code each, regardless of language, ...

With respect to math symbols, this means that if two symbols look very much alike, unless there is very strong documentation to support the contention that they have different meanings, only one will be assigned a code. Thus, for example, $\leq$ and $\leqq$ might be considered equivalent — if they had not already both been accepted into Unicode. Some UTC members feel that the original inclusion of such pairs was a mistake, and they are determined not to repeat it. Knowledge of that fact guided the organization of the math proposal, and helped to determine what kind of documentation would be needed.

The first rearrangement of the symbols was into groups that roughly coincided with existing Unicode blocks: arrows, "traditional" math symbols, geometric shapes, etc. The "traditional" symbols were classified further into groups that corresponded to their functions: large operators, binary operators, binary relations, delimiters, etc. After some preliminary discussions with a member of the UTC, we decided to structure our proposal in blocks corresponding to these functional groups, with the symbols arranged in the same general order as similar ones already present in Unicode.

We were also advised to eliminate any "duplicates" of existing characters, but since slight variations do often have different meanings in mathematical exposition, we decided to keep everything for the first round, and refine on the basis of specific directives from the UTC. However, we did at this stage identify symbols with similar shapes, and possibly equivalent meanings, and flagged them in our master list to indicate the need for additional documentation.

The UTC requested at the outset that symbols used not in math, but in fields such as chemistry, astronomy, engineering and phonetics, be omitted, to be requested separately by representatives of those disciplines. This change was made, eliminating more than a hundred symbols.

In the first round of the proposal, we included three alphabets — blackboard bold, script and fraktur — as well as a list of alphabets required for mathematical exposition. Some additional alphabetic inclusions were letters that occur in an unusual orientation (e.g., $\exists$) and variant forms of several Greek letters (including ϵ, the straight-backed epsilon) which were missing from the Unicode complement. Although we felt that the case for including these alphabets was strong, opposition from members of the UTC was stronger; in the next iteration, the alphabets were removed to a separate proposal.

Refining the proposal

Adjusting the content of the list of symbols to make it acceptable to the UTC took several iterations over the course of two years. Attendance at the meetings where the proposal was discussed proved to be essential, given the scope of the project. Two UTC members were assigned to help refine the proposal, cast it in the form required by the ISO working group (WG2) in charge of character coding standards, and prepare the text that will appear in the published Unicode manual.

The first task was to overcome the reluctance of some members of the UTC to believe that there could actually be more than a thousand math and technical symbols not already in Unicode. The reorganized proposal, generally following the ordering of symbols already present, clarified the situation by making it relatively easy to compare the new material to the existing Unicode.

Quite a few symbols consisted of a base symbol plus a cancellation. This could be a long or short slash or vertical stroke, a backwards slash, or a double vertical stroke. Although two lengths of the forward slash and vertical stroke appear in Unicode among the combining diacritics, it was decided that the longer versions would be designated as the proper cancellation markers for math, leaving the actual shape of the cancelled symbol as a font issue.

Special attention was paid to symbols that look similar, that differ for example in the number

Several decisions were made in order to mimimize the number of codes to be assigned. These were the most important:

- Any symbols cancelled by a vertical or slanted stroke should be constructed from a base symbol and a combining diacritic; this eliminated the entire cancelled alphabet used by physicists.
- A "variant selector" (VS) would be provided to allow for shape variants that ordinarily represent personal preference or house style and not differences in meaning. Except where needed to provide a base character for cancellation, in which case a code would be assigned, only the most common variant of such a symbol would be assigned, and the specified variant represented by the code for the base symbol plus the VS. A list of such variants appears in figure 2.

Documentation for the symbols that were most likely to be controversial was sought in published material. A request for citations was presented on

Symbol variants defined using a Variation Selector (VS)

Barbara Beeton, for STIPUB
7 February 2000

- 2268 ≨ + VS → ≨ less-than and not double equal - with vertical stroke
- 2269 ≩ + VS → ≩ greater-than and not double equal - with vertical stroke

- 22DA ⋚ + VS → ⪋ less-than above slanted equal above greater-than
- 22DB ⋛ + VS → ⪌ greater-than above slanted equal above less-than
- 2272 ≲ + VS → ⩽ less-than or similar - following the slant of the lower leg
- 2273 ≳ + VS → ⩾ greater-than or similar - following the slant of the lower leg
- 2A9D ⪝ + VS → ⪡ similar - following the slant of the upper leg - or less-than
- 2A9E ⪞ + VS → ⪢ similar - following the slant of the upper leg - or greater-than
- 2AAC ⪬ + VS → ⪯ smaller than or slanted equal
- 2AAD ⪭ + VS → ⪰ larger than or slanted equal

- 228A ⊊ + VS → ⫋ subset not equals - variant with stroke through bottom members
- 228B ⊋ + VS → ⫌ superset not equals - variant with stroke through bottom members
- 2ACB ⫋ + VS → ⫋ subset not two-line equals - variant with stroke through bottom members
- 2ACC ⫌ + VS → ⫌ superset not two-line equals - variant with stroke through bottom members

- 2A3B ⨻ + VS → ⨽ interior product - tall variant with narrow foot
- 2A3C ⨼ + VS → ⨾ righthand interior product - tall variant with narrow foot

- 2295 ⊕ + VS → ⊕ circled plus with white rim
- 2297 ⊗ + VS → ⊗ circled times with white rim
- 229C ⊜ + VS → ⊜ equal sign inside and touching a circle

- 2225 ∥ + VS → // slanted parallel
- 2225 ∥ + VS + 20E5 \ → ⫽ slanted parallel with reverse slash

** • 222A ∪ + VS → ∪ union with serifs
** • 2229 ∩ + VS → ∩ intersection with serifs
** • 2293 ⊓ + VS → ⊓ square intersection with serifs
** • 2294 ⊔ + VS → ⊔ square union with serifs

Notes:

** The shape is incorrect, owing to unavailability of a suitable font; the correct shape will be provided as soon as possible. The associated text correctly describes the desired shape.

Figure 2: Symbols constructed using the Variant Selector

the AMS Web site during the summer and early fall of 1998. Although the response was not as great as hoped for, some useful references were obtained. The ideal citation contained several similar-appearing symbols on a single page, in context, preferably with definitions of one or more of the symbols in the text. More than a hundred pages of such examples were copied, annotated, assigned reference IDs, indexed, and provided to the two UTC members responsible for advancing the symbols proposal. This mass of data proved its worth more than once, when it was possible to cite a particular sample in answer to a challenge.

Late in 1999, the content of the symbols proposal was agreed, codes were assigned, and a final version of the proposal was prepared for the spring 2000 meeting of WG2. The proposal forwarded to WG2 places material into the following blocks:

- U+2000–206F: General punctuation (9 new codes)
- U+20D0–20FF: Combining diacritical marks for symbols (4 codes)
- U+2100–214F: Letterlike symbols (16 codes)
- U+2190–21FF: Arrows (12 codes)
- U+2200–22FF: Mathematical operators (14 codes)
- U+2300–23FF: Miscellaneous technical (26 codes)
- U+2400–243F: Control pictures
- U+25A0–25FF: Geometric shapes (8 codes)
- U+2900–297F: Supplemental arrows (new; 128 codes)
- U+2980–29FF: Miscellaneous math symbols (new; 114 codes)
- U+2A00–2AFF: Supplemental math operators (new; 246 codes)

In addition, a few characters were added to other areas; in all, 584 new codes have been assigned.

After much discussion, the proposition was reluctantly accepted that the same letter from different alphabets has different meanings within a single document, and thus these different alphabets deserve to be coded *for use only in mathematical notation*. The example used to clinch the argument was the contrast between these two formulas:

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

$$H = \int d\tau(\varepsilon E^2 + \mu H^2)$$

The first is the Hamiltonian formula well known in physics; the second is an unremarkable integral equation.

These alphabets are needed for proper composition of mathematics:

- lightface upright Latin, Greek and digits
- boldface upright Latin, Greek and digits
- lightface italic Latin, Greek and digits
- boldface italic Latin, Greek and digits
- script
- fraktur
- bold fraktur
- open-face (blackboard bold) including digits
- lightface upright sans serif Latin and digits
- lightface italic sans serif Latin
- boldface upright sans serif Latin, Greek, and digits
- boldface italic sans serif Latin and Greek
- monospace Latin and digits

Except for the lightface upright letters and digits, which are to be encoded using the base Unicodes (ASCII for the Latin letters and digits), the alphanumerics are to be placed in a tightly packed block (U+D400–D7FF) in plane 1, so that they can be used for math (most likely via entity names in MathML), but will be very difficult to access for other purposes.

The math alphanumerics block has been incorporated into a larger proposal for plane 1, and its schedule is slightly behind that of the symbols proposal. A "final" version is now in preparation, and will be forwarded to WG2 for their fall meeting.

Assuming that the required three ISO ballots are favorable, the new codes should be a formal part of Unicode and ISO 10646 by

Commissioning a font

Late in 1999, even before the fate of the Unicode math proposals was known, STIPub issued a Request for Proposal to a number of font suppliers. This RFP requested bids for creating a set of fonts compatible with Times and incorporating all the symbols and alphabets identified by the STIX project, suitable both for use in Web browsers and in print. The resulting font set is to be "made available for general use under license, but free of charge, with the aim of easing and fostering the uninhibited flow, exchange, and linking of scientific information." [6]

Proposals were received from four potential suppliers, and comments from a fifth, which refrained from proposing because of a prior commitments. As of this writing, a probable supplier has been chosen, and negotiations are proceeding toward a contract.

More details will be available on this topic at the time of the Oxford TUG meeting.

Remaining

The Unicode manual contains extensive text describing the proper use of the character codes, as a guide to programmers. Particular attention is paid to processing of context dependencies, combining codes and the like. Since mathematical notation will be realized in a document as a combination of coding and markup, and not all mathematical symbols are interpreted in the same way, instructions are needed. The creation of a technical report is an open action item on the Unicode docket; the text of this report will ultimately be incorporated into the Unicode manual.

The symbols from chemistry, astronomy, engineering and phonetics that were excluded by request of the UTC have been left for consideration by the organizations that submitted them.

Mathematical notation is not static. Authors continually devise new symbols and ascribe new meanings to existing ones. The complement of symbols requested from Unicode was frozen in mid-1998; a few additions were made only to achieve consistency in the base set of symbols affected by combining codes, namely the VS and negation marker. About 50 additional symbols used in math, physics and theoretical computer science have been identified since then. Documentation must be completed for these, and the formal request presented to the UTC for their addition.

The glyph complement was frozen somewhat later than the Unicode complement, but this too remains to be addressed.

It has not been determined how these, or further additions, are to be handled. Nonetheless, I am still actively collecting citations for new notation, and welcome contributions.

Acknowledgments

Three UTC members were involved heavily in advancing this project: Murray Sargent of Microsoft, formerly a practicing physicist; Ken Whistler of Sybase, the UTC archivist and general editor of Unicode 3.0; and Asmus Freytag, the UTC font specialist, who is responsible for typesetting the Unicode manuals and ISO 10646. Their knowledge and experience of character code standards has proved invaluable, and the project's success owes much to their thoughtful assistance. Ken in particular is able to explain in non-expert terms the UTC's requirements and the historical background affecting specific decisions. From the other direction, he can quickly assess the evidence for support of an item and, if convinced that it has sufficient merit, can convince the rest of the committee that it should be included. Thus, for the math proposal, the strategy was to provide sufficient evidence for a symbol to convince Ken, and then let him persuade the committee using arguments they would find convincing.

The contributors to the symbol collection were always willing to provide additional information. Neil Soiffer of Wolfram Research attended several UTC meetings to explain the uses of several "letter-like symbols" that have special significance in Mathematica®.

Patrick Ion, co-chair of the W3C MathML working group, took my place at several UTC meetings when questions were expected that would best be answered by a practicing mathematician, such as, "Do you know for a fact that [some particular symbol] is used, and are you sure it isn't the same as [some other symbol]?" His presence and his answers successfully conveyed the importance of this project to the mathematical community.

Thanks to all for their efforts.

For more information

The history of the STIX project is recorded at the Web site `http://www.ams.org/STIX/`.

References

[1] Association for Font Information Interchange, *International Glyph Register, Volume 1: Alphabetic scripts and symbols*, Rochester, 1993.

[2] International Organization for Standardization, ISO 8879:1986, *Information Processing — Text and office systems — Standard Generalized Markup Language (SGML)*, Geneva, 1986.

[3] International Organization for Standardization, ISO 9541:1991, *Information Technology — Font Information Interchange — Part 1: Architecture*, Geneva, 1991.

[4] International Organization for Standardization, ISO 9573-13:1991, *Information Technology — SGML Support Facilities — Techniques for using SGML — Part 13: Public entity sets for mathematics and science*, Geneva, 1991.

[5] International Organization for Standardization, ISO 10646-1:1993, *Information Technology — Universal Multiple-Octet Coded Character Set (UCS) — Part 1: Architecture and Basic Multilingual Plane*, Geneva, 1992.

[6] "STIPUB Request for Proposals for Scientific and Technical Fonts for the STIX Project", November 16, 1999 (unpublished).

[7] The Unicode Consortium, *The Unicode Standard, Version 2.0*, Addison-Wesley Developers Press, Reading, MA, 1996.

[8] The Unicode Consortium, *The Unicode Standard, Version 3.0*, Addison Wesley Longman, Inc., Reading, MA, 2000.

[9] *Webster's New Collegiate Dictionary*, G. & C. Merriam, Springfield, MA, 1959.

[10] Justin Ziegler, *Technical Report on Math Font Encodings*, available from CTAN (`http://www.tug.org/tex-archive/info/ltx3pub/l3d007.tex` and supporting files in the same directory), 1994.

Barbara Beeton