L2/04-303

Elaine Keown
Tucson, Arizona 85712
Email: k_isoetc@yahoo.com
July 17, 2004

Ken Whistler
Rick McGowan
Michael Everson

RE: Proposal for a first allocation or re-allocation of
Section 08 of the Roadmap to the BMP

Gentlemen:

As you may know, I spent much of the last 5.5 years researching the complete character set for ‘Extended Hebrew’ and ‘Extended Aramaic,’ in linear alphabetic script.

My current online character set list, ‘The Aramaic and Hebrew Character Sets, Revised List, (6-22-2004), http://www.lashonkodesh.org/hprelist.doc , gives an estimate of ~271 characters needed for alphabetic Hebrew and Aramaic in ‘linear Canaanite’ when all glyphs are **maximally unified**. Slightly >100 characters are already in the UCS.

In my list I **unified** square script with Samaritan, with all Hebrew and Aramaic written in ancient linear Canaanite since 1,150 B.C.E., and with all possible Judean Desert --- Qumran+ --- sets of glyphs (Cryptic A etc.). I did not count glyphs to write Hebrew or Aramaic in scripts such as Arabic, Syriac, cuneiform, Egyptian demotic, Cyrillic, Roman, etc.

Since 1987, international computer codes for Hebrew have always been **misleadingly** short. ISO 8859-8 contained only Hebrew consonants and two dots for sin and shin. It did not include even standard Tiberian vowels, despite having space for them.

The current Unicode Hebrew block inherited 8859-8 and is still missing some standard Tiberian items.

However, the June 15, 2004, additions voted in at Markham have brought the main Unicode Hebrew block closer to completion, thanks to the work of John Hudson, Michael Everson, Peter Kirk, and Mark Shoulson. We are currently discussing what else should be added to the main Hebrew block. Research is proceeding in New Jersey, East Anglia, British Columbia, and elsewhere.

This letter requests that the UTC / WG2 consider building a ‘Hebrew Extended’ section in the Roadmap area 08 (from 0800-08FF, contains 256 code points). Part of line 08 has never been allocated, part was allocated as I am asking, and part would be allocated differently.

If built, this new block would parallel the BMP extension blocks for Arabic, Ethiopic, Latin, etc. It would complete alphabetic Hebrew and Aramaic within Unicode, except for one or more number systems and epigraphy. Hebrew/Aramaic epigraphic material overlaps with other Canaanite languages and with some early Mediterranean material in Greek and so forth.

Normally a researcher would not suggest a BMP allocation, even of a section which was 40% empty and contained two allocated Semitic blocks.

But I wished to inform the committee of other global technical problems with alphabetic Hebrew/Aramaic, and of the effects of separately encoding Samaritan. This allocation relieves problems with such separation for scholars who use some Samaritan, without denying the Samaritan community a separate encoding.

Proposed New Roadmap for 0800—08FF:

	0 1 2 3 4 5	6 7 8	9 A	B C D E F
08	Hebrew Extended		Samaritan	Mandaic

Details of ‘Hebrew Extended’:

	0	1 2	3 4 5	6 7 8
08	Judeo-Arabic etc.	Judean Desert (Qumran)	Babylonian pointing	Palestinian pointing

The material proposed in ‘Hebrew Extended’ and in Samaritan has the following proposal status:

1. Judeo-Arabic etc.: to be proposed, 10 suggested characters, font in preparation.
2. Judean Desert (Qumran and related material): has 24 characters unless ‘Cryptic A’ or other esoteric Qumran scripts are disunified. Disunification of ‘Cryptic A’ would require 22 more points, obviously. I have list of characters, sample proposal, and letters in to a couple of Qumran scholars. The TLG did some Qumran characters. Their work may or may not be completely relevant to Qumran text representation.
3. Babylonian pointing: Preliminary proposal (http://www.lashonkodesh.org/bavelpro.pdf). Proposal is waiting on font and on further research to find variants in targum literature. So far found a variant rafe.
4. Palestinian pointing: font in preparation, proposal also.
5. Samaritan pointing: Very preliminary Samaritan proposal (http://www.lashonkodesh.org/samarpro.pdf). Proposal has vowels, manuscript scan, brief discussion. Font in preparation.

Technical Notes on Samaritan:
First, it appears that Samaritan writing has developed case, so I gave it more space than in your current Roadmap. However, I don’t know when case developed, only that it is probably used in contemporary Samaritan.

Earlier Samaritan manuscripts, which are highly valued by scholars, probably don’t have case.

Intra-Scriptal Hebrew Collation and Separated Samaritan

Recently it has come to me that intra-scriptally **Hebrew has 13 collation levels** (when fully unified). These 13 collation levels will be utilized by Aramaists and by Jewish studies scholars who work with many different materials. In addition, the Karaite Hebrew material in Arabic script will probably be the first ‘interleaving’ level, by default.

The Samaritan vowels, accents, and punctuation will probably be the only Samaritan sub-blocks used by many scholars. This would allow them to stay at 13 collation levels with zero or one interleaving level. However, if they work with the earlier Samaritan manuscripts and choose to have another ‘interleaving’ level, then they would also use the Samaritan upper case letters (I assume that Samaritan developed lower-case, which is what Roman did, but I don’t know yet).

It’s helpful to Semitists who use Unicode (but not necessarily XML Schemas) to have another contiguous block of Semitic material. Within Semitics, the most sophisticated technology used so far is information retrieval. In Israel, Hebrew information retrieval for unpointed Hebrew has been widely studied since the 1960s, first at Weizmann and then at Bar Ilan. Contiguous blocks are easier to test and program with this type of technology.

Thank you for considering this. I will be asking several individuals, the Unicode Hebrew list and several other electronic discussion lists to respond via the usual Unicode online response form.

Elaine Keown

REFERENCES:

1. Current Roadmap for 0800-08FF:

08	(Avestan and Pahlavi)	¿Mandaic?	¿Samaritan?	???	???	???	???	???

“Roadmap to the BMP,” Michael Everson, Rick McGowan, Ken Whistler; 2004-06-24
http://www.unicode.org/roadmaps/bmp/bmp-4-8.html .

2. Keown, E.R. “Hebrew alphabets, symbols and computer codes: history and preliminary tabulation.” Revue des Études Juives, 161 (1-2), pp. 235-240. N.B. this article is an earlier and less interesting version of my research and has miscellananeous errors, e.g., I have never had an affiliation with UPenn. There is also one chronological error.

cc: Mike Ksar, Cathy Wissink, Arnold Winkler, Lisa Rajchel, Deborah Anderson,
Kent Richards, Patrick Durusau, Kirk Lowery, Alan Groves, Dean Snyder, Peter Kirk,
unicode@unicode.org, hebrew@unicode.org, et al.