Script Encoding Initiative: Past and Future

Unicode Technical Workshop
November 8, 2023

Debbie Anderson
Anushah Hossain
Debbie Anderson
Researcher, Department of Linguistics
University of California, Berkeley

Anushah Hossain
Postdoctoral Scholar, Digital Civil Society Lab
Stanford University
What is the Script Encoding Initiative?

A project hosted by UC Berkeley’s Linguistics department that helps prepare proposals of historic and minority scripts for inclusion in the Unicode Standard and ISO 10646
What is the Script Encoding Initiative?

1. Has held a unique position in the Unicode Consortium as an academic member for 19 years
2. Persists despite opposing incentive structures
3. Extremely successful! Over 100 scripts already encoded through SEI’s help and many more in the pipeline
SEI’s history and unique position
Unicode’s early priorities

Wide coverage
Consistency and efficiency in a limited codespace
Buy-in from companies and governments
Unicode’s early priorities

Wide coverage
Consistency and efficiency in a limited codespace
Buy-in from companies and governments

Round-trip backwards compatibility

Characters, not glyphs
Stability policy
Unicode’s early priorities

What is needed is a new international/multilingual text encoding standard that is as workable and reliable as ASCII, but that covers all the scripts of the world.

For reference, the table below ranks the world’s writing systems roughly in order of commercial importance, as measured by the total GNP of countries using each system:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Writing System</th>
<th>Languages</th>
<th>% of World GNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Latin</td>
<td>English, German, French, Spanish, Italian, Portuguese, Indonesian/Malay, ...</td>
<td>68</td>
</tr>
<tr>
<td>2</td>
<td>CJK ideographs</td>
<td>Chinese, Japanese, (Korean)</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>Cyrillic</td>
<td>Russian, Ukrainian, ...</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>Arabic</td>
<td>Arabic, Persian, ...</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Devanāgari family</td>
<td>Hindi, Bengali, Punjabi, Marathi, ...</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Korean (Hangul)</td>
<td>Korean</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Dravidian family</td>
<td>Telugu, Tamil, ...</td>
<td>ε</td>
</tr>
<tr>
<td>8</td>
<td>Greek</td>
<td>Greek</td>
<td>ε</td>
</tr>
<tr>
<td>9</td>
<td>Khmer</td>
<td>Thai, Lao, Khmer</td>
<td>ε</td>
</tr>
<tr>
<td>10</td>
<td>Hebrew</td>
<td>Hebrew</td>
<td>ε</td>
</tr>
</tbody>
</table>

Excerpt from “Unicode 88”
Unicode’s early priorities

Excerpt from “Unicode 88”

Distinction of “modern-use” characters: Unicode gives higher priority to ensuring utility for the future than to preserving past antiquities. Unicode aims in the first instance at the characters published in modern text (e.g. in the union of all newspapers and magazines printed in the world in 1988), whose number is undoubtedly far below $2^{14} = 16,384$. Beyond those modern-use characters, all others may be defined to be obsolete or rare; these are better candidates for private-use registration than for congesting the public list of generally-useful Unicodes.

In other words, given that the limitation to 65,536 character codes genuinely does satisfy all the world’s modern communication needs with a safety factor of about four, then one can decide up-front that preserving a pure 16-bit architecture has a higher design priority than publicly encoding every extinct or obscure character form. Then the sufficiency of 16 bits for the writing technology of the future becomes a matter of our active intention, rather than passive victimization by writing systems of the past.
Script encoding process
Script encoding process

1. Ensure writing system is in active use (modern scripts) and well-documented (modern and historic scripts)
Script encoding process

2. Gather evidence to submit alongside character proposal

6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?
Script encoding process

Adlam meeting, 2014

Mongolian meeting, 2017
How SEI got started

“Scratching your own itch”
How SEI got started

Some choice meetings...
How SEI got started

...leading to regular attendance at UTC meetings and eventually membership in the Unicode Consortium
How SEI got started

Expansion from historic Indo-European scripts to modern minority languages

Old Italic

Aegean scripts (Linear B, Cypriot Syllabary)

Ol Chiki

Balinese

N’Ko

...
How SEI got started

SEI (as UC Berkeley) has been the only university on Unicode Consortium roster since it joined in 2004
Managing opposing structures
Finding funding

Private donors, UNESCO, NEH, Google grant
Finding funding
Private donors, UNESCO, NEH, Google grant

Getting university support for Unicode membership
Eventually paid from grants
Finding funding
Private donors, UNESCO, NEH, Google grant

Getting university support for Unicode membership
 Eventually paid from grants

Convincing scholars about importance of Unicode
Shifted slowly as Unicode (and the Internet) took off
Finding funding  Private donors, UNESCO, NEH, Google grant

Getting university support for Unicode membership  Eventually paid from grants

Convincing scholars about importance of Unicode  Shifted slowly as Unicode (and the Internet) took off

Convincing modern language communities of Unicode’s utility  Relayed effectively through the success of ADLaM and N’Ko
Doing lots with little; ongoing challenges
Finding and supporting experts and community members can be difficult

- Conferences and ISO meetings helpful
- Need to stay informed on ~40 scripts at a time that are somewhere in the pipeline
  - More examples needed?
  - Proposal vetted?
  - Outstanding issues resolved?
  - User community involved?
- Different parties working at different paces
Identifying long-term funding and full-stack support

- Providing assistance in getting script on devices and in software
- Ongoing goal to raise funds for this work, ideally disbursed over longer time horizons
How to address shifts in understanding and in Unicode expectations?
Takeaways

Keep engaging with communities already in Unicode

Better explain the steps after Unicode

Keep working with contacts to get information on unencoded scripts

Lobby companies to support SEI’s work and work on minority scripts

Encourage educational institutions to join Unicode and have a voice!
SEI, looking forward:

Help replenish ranks of Unicoders via coursework and research opportunities through SEI

Share past script communities’ experiences with digitization in a collective archive

Translate SEI’s story for public audiences through public writing and presentations
“I found while attending Unicode Technical Committee meetings in the early 2000s that I had found my tribe, though it took some time to follow the Unicode lingo and to understand the encoding process”
Thank you!

SEI
Script Encoding Initiative
Department of Linguistics
University of California, Berkeley

linguistics.berkeley.edu/sei

Debbie Anderson
dwanders@berkeley.edu

Anushah Hossain
ahossain@stanford.edu

Image courtesy of www.worldswritingsystems.org