Developing a Success Proposal for Encoding a Script in Unicode

Internationalization and Unicode Conference #42
11 September 2018

Anshuman Pandey
Script Encoding Initiative, UC-Berkeley
About me

• Product Strategist, Avant, Chicago
  • Fin-tech analytics

• Researcher, Script Encoding Initiative, UC-Berkeley
  • Develop Unicode standards for writing systems of the world

• Education
  • Post-Doctoral Researcher (Linguistics), UC-Berkeley, USA
  • PhD (History), University of Michigan, Ann Arbor, USA
About me: Unicode

• Developed encodings for
  • 24 scripts
  • 4 number systems
  • various individual characters

• Work in progress
  • 25 script proposals
  • research on 100+ scripts
About me: Unicode

• Developed encodings for
  • 24 scripts
  • 4 number systems
  • various individual characters
  • 2 emoji

• Work in progress
  • 25 script proposals
  • research on 100+ scripts
Why?: Enable Digital Humanities Foundations

• Representation of ‘visible language’, eg. script / writing system
  • Often ignored in traditional and applied linguistics and language studies
  • Key to access and presentation of textual sources in the original

• Display and input of language (keyboard, stylus, etc.)
  • Enable users to input, search, and view text uniformly on devices

• Natural language processing (NLP) and machine learning (ML)
  • Information extraction and analysis using programmatic methods

• Goal: all languages as digitally native as English
‘Full Stack’ Digital Support for Languages

- Character-encoding standard
- Locale data
- Font / typeface
- Keyboard / input methods
- Text layout and formatting
- Database support
- System and browser support
- NLP tools and algorithms
‘Full Stack’ Digital Support for Languages

• Character-encoding standard
• Locale data
• Font / typeface
• Keyboard / input methods
• Text layout and formatting
• Database support
• System and browser support
• NLP tools and algorithms
Script Encoding in the DH Stack
Script Encoding Process
Script Encoding Process: Overview

1. **Identification**: Native users, scholars, others identify a script not yet encoded in Unicode

2. **Development**: Research script and develop script proposal

3. **Review**: Unicode Technical Committee reviews proposals; may request changes; vote to approve or disapprove

4. **Publication**: Publication of script in Unicode standard

5. **Implementation**: Create fonts, keyboards, update software (Noto project; individuals)

6. **Iterative development**: Repeat for new character additions
Script Encoding Process: Focus

1. **Identification:** Identify a script not yet encoded in Unicode

2. **Development:** Techniques for researching scripts and models
Script Encoding Process

Identification
Script Encoding Process: Identification

• Existing character-encoding standards

• Request from a native user community

• Proposal from scholarly user community

• Submission from writing-systems researchers and enthusiasts

• Fieldwork and archival research
Identifying a script

<table>
<thead>
<tr>
<th>حساب جنسی سیاقی (ترتیب دهده)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ش</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
</tbody>
</table>

Planned: Persian Siyaq
Table showing planned and unplanned writing systems:

<table>
<thead>
<tr>
<th>Planned: Persian Siyaq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unplanned: Sogdian</td>
</tr>
</tbody>
</table>

Excerpt from the text:

Identification

\[
\text{ب جنسی سیاقی (ترنیب دهدی)}
\]
Identification

Planned: Persian Siyaq

Unplanned: Sogdian

Unplanned: Old Sogdian
Case Study: Old Sogdian

• Consonantal alphabet

• Used for writing Sogdian
  • now-extinct eastern Iranian language

• Derived from Imperial Aramaic
  • Variation of Achaemenid administrative script

• In use from 3rd – 7th century CE
Old Sogdian: Identification

• Known to scholars
  • 19th century to present
  • active scholarship, publication trends

• Digital humanities content
  • TITUS project: transliterations
  • UW Silk Road Project: translations

• Manuscript digitization
  • International Dunhuang Project (IDP)
Situating the Sogdians
Old Sogdian: Identification

Ancient Letter No. 3
Paragraph: 3
= Ms. Or_8212_(98) = T._XII.a._II.3
Manuscript: Or_8212_(98)

Reichelt, 22-24

read from MS, rechecked from facsimile 3.8.94: NSW
Old Sogdian: Identification

Sogdian Ancient Letter No. 3

[Verse] From (his) daughter Shayn to the noble lord Nanai-dhat.

[On another part of the verso] From (his) servant [left unfinished].

[Recto] To (my) noble lord (and) husband Nanai-dhat, blessing (and) homage on bended knee, as is offered to the gods. And (it would be) a good day for him who might see you healthy, happy (and) free from illness, together with everyone; and, sir, when I hear (news of) your (good) health, I consider myself immortal!

Behold, I am living ..., badly, not well, wretchedly, and I consider myself dead. Again and again I send you a letter, (but) I do not receive a (single) letter from you, and I have become without hope towards you. My misfortune is this, (that) I have been in Dunhuang for three years thanks (?) to you, and there was a way out a first, a second, even a fifth time, (but) he(!) refused to bring me out. I requested the leaders that support (should be given) to Farkhund for me, so that he may take me to (my) husband and I would not be stuck in Dunhuang. (for) Farkhund says: I am not Nanai-dhat’s servant, nor do I hold his capital. I also requested thus: If he refuses to take me to (my) husband, then ... such support for me that he may take me to (my) mother. The leaders say: Here in Dunhuang there is no other relative closer than Artivan, (but) Artivan [says]: Farkhund ... whatever ... to do for you. If(?) If(?) (had) no guarantee, no protection, my father ... I have become ... not ... How much more would I have ... by my father if ... a servant of the Chinese! A free man ... who found ... and ... keeps (his) clothing in good condition(?). And you write (your) bidding to me about everything in ... so that I should ... you and I should know how to think, and if I do not ... you, then you write to me so that I should know how to serve the Chinese. In my paternal abode I did not have such a restricted ... as with(?) you. I obeyed your command (lit. took your command upon my head) and came to Dunhuang and I did not observe (my) mother’s bidding nor (my) brothers’. Surely(?) the gods were angry with me on the day when I did your bidding! I would rather be a dog’s or a pig’s wife than yours! And for me ...

Sent by (your) servant Miwnay. This letter was written in the third month on the tenth day.

[Added in the margin] From (his) daughter Shayn to the noble lord Nanai-dhat, blessing (and) homage. And (it would be) a good [day] for him [who] might see [you] healthy, rested (and) happy. ... I have become ... and I watch over a flock of domestic animals. Differently to you, I had a ..., and ... went out. I am ... and I know that you do not lack twenty staters(?) to send. It is necessary to consider the whole (matter). Farkhund has run away; the Chinese seek him but do not find him. Because of Farkhund’s debts we have become the servants of the Chinese, I together with (my) mother.
Script Encoding Process: Identification
Old Sogdian: Identification

Sogdian script

In the Sogdian script used in the "Ancient Letters" (table 48.2), most of the letters are distinct and do not change shape when joined. In the "formal" and "Uyghur" Sogdian scripts, most of the letters are joined and, owing to the use of a broad pen, are frequently difficult to distinguish. In the earlier form, 'i' is still distinguished from 'n', but in the later, 'i = n.' Some scribers distinguish 'z' from 'n' by not connecting 'z' to the preceding letter, but others make no distinction. In the later, increasingly curvilinear form, other letters tend to become indistinguishable as well: γ (gamma) and δ (delta). These letters are distinguished only in final position (by some scribes), e.g., 'n-z, x-z.'

'z' is sometimes distinguished from 'n' or 'z' by a diacritical point, and the foreign sound 'b' was noted as 'b.'

SAMPLES OF SOGDIAN

ANCIENT LETTERS

<table>
<thead>
<tr>
<th>Sogdian</th>
<th>Aramaic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old Sogdian Letters</td>
<td>Sogdian</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>(β)</td>
<td>β</td>
</tr>
<tr>
<td>g</td>
<td>γ</td>
</tr>
<tr>
<td>(γ)</td>
<td>γ</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>h (h)</td>
<td>a, ø</td>
</tr>
<tr>
<td>w</td>
<td>w, b, ø, i</td>
</tr>
<tr>
<td>z</td>
<td>z</td>
</tr>
<tr>
<td>(i)</td>
<td>i</td>
</tr>
<tr>
<td>(ζ)</td>
<td>ζ</td>
</tr>
<tr>
<td>b (h)</td>
<td>γ, x, h</td>
</tr>
<tr>
<td>i</td>
<td>t</td>
</tr>
<tr>
<td>y</td>
<td>y, ξ, l</td>
</tr>
<tr>
<td>k</td>
<td>k</td>
</tr>
<tr>
<td>l (i)</td>
<td>o</td>
</tr>
<tr>
<td>m</td>
<td>m</td>
</tr>
<tr>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>(s)</td>
<td>s</td>
</tr>
<tr>
<td>s</td>
<td>ø</td>
</tr>
<tr>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>(f)</td>
<td>f</td>
</tr>
<tr>
<td>(σ)</td>
<td>σ</td>
</tr>
<tr>
<td>q</td>
<td>q</td>
</tr>
<tr>
<td>r</td>
<td>r</td>
</tr>
<tr>
<td>s</td>
<td>s</td>
</tr>
<tr>
<td>t</td>
<td>t, θ</td>
</tr>
</tbody>
</table>

1. Transliteration: OD ßyw xwert βer'kk mnyßyw'rw k'nk'kk
2. Normalization: at βar xulb ḫarא nanè-bìar kanak
3. Gloss: to lord.acc master Barak Nānā's-gift Kanak
4. tLP ßerwr ŠLM nm'cyw sp'tz'cyk AYKZY
5. (éw)-zār *hēwar *nēmcyu spátz-nīk kāb-utu
6. thousand ten.thousand greeting(?) reverence'acc. bended.knee when-that.and
7. ŽK núlw βyrt pyst MN xψyβ ġanā nyēlet
8. wēšânu ḫarēmov( ) ḫarē pišt con ġepē ġanā nanè-ğanē
9. their,obl lords,obl received written from own servant Nānā's-servant

'To the Divine Master Barak (?) Nanethwara Kanak a thousand, ten thousand greetings, reverently with bended knees when received by their divinities. Written by his own servant Nanavanta.'

—From the Old Sogdian "Ancient Letters" found in a mailbag in the Great Wall (AL II, Reichelt 1931: 12 and pl. 2).
Script Encoding Process: Identification

• Challenges
  • Charts or specimens not in published in books on writing systems
  • Studies may be outdated or very limited

• Resolution
  • Conduct research: primary sources!
Script Encoding Process

Research
### Roadmap to the SMP

<table>
<thead>
<tr>
<th>Revision</th>
<th>8.0.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authors</td>
<td>Michael Everson, Rick McGowan, Ken Whistler, V.S. Umamaheswaran</td>
</tr>
<tr>
<td>Date</td>
<td>2015-08-17</td>
</tr>
<tr>
<td>This Version</td>
<td><a href="http://www.unicode.org/roadmaps/smp/smp-8-0-1.html">http://www.unicode.org/roadmaps/smp/smp-8-0-1.html</a></td>
</tr>
<tr>
<td>Previous Version</td>
<td><a href="http://www.unicode.org/roadmaps/smp/smp-8-0-0.html">http://www.unicode.org/roadmaps/smp/smp-8-0-0.html</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Code</th>
<th>Script</th>
<th>Alternate Script</th>
<th>Code</th>
<th>Script</th>
</tr>
</thead>
<tbody>
<tr>
<td>108</td>
<td>Cypriot</td>
<td>Imp. Aramaic</td>
<td>Palmyrene</td>
<td>Nabataean</td>
</tr>
<tr>
<td>109</td>
<td>Phoenician</td>
<td>Lydian</td>
<td>???</td>
<td>???</td>
</tr>
<tr>
<td>10A</td>
<td>Kharoshthi</td>
<td>O.S. Arabian</td>
<td>O.N. Arabian</td>
<td>Meroitic H.</td>
</tr>
<tr>
<td>10C</td>
<td>Old Turkic</td>
<td>(Baburi)</td>
<td>???</td>
<td>Old Hungarian</td>
</tr>
<tr>
<td>10D</td>
<td>(Kharlingya)</td>
<td>(Garay (Wolof))</td>
<td>???</td>
<td>Old Byblos</td>
</tr>
<tr>
<td>10F</td>
<td>Any</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>Brahmi</td>
<td>Kaithi</td>
<td>Sora Sompeng</td>
<td></td>
</tr>
</tbody>
</table>
Old Sogdian: Mistaken identity
Old Sogdian: Mistaken identity
Old Sogdian: Unicode Status

14 *Aramaic* forms a rather complex family of scripts, with a number of descendants. Certainly there is a basic Aramaic, but it has many descendents (including Mongolian and possibly Brahmi) which are unique enough to merit their own encoding (see table 5.5). More research is required. However, Aramaic is expected to encompass at least:

- *Aramaic proper*
- *Middle Persian*
- *Parthian*
- *Sogdian*

Edessan is likely to be either Aramaic or Syriac. More research is required.

15 Phoenician is the catch-all for the largest group of related scripts including its ancestors, Proto-Sinaitic/Proto-Canaanite. Looking at tables 5.1, 5.3, and 5.4 (below) most of the scripts are so similar that there doesn’t seem to be any point in trying to encode them separately.

---

Sogdian: Unicode Status – Unification?
Script Encoding Process: Research

• Analysis of sources to document repertoire and script grammar

• Compare grammar and orthography of related scripts

• Outreach to user community to understand current usage
Sogdian sources

- Kultobe: the oldest Sogdian records are stone inscriptions found at Kultobe in modern Kazakhstan

- ‘Ancient Letters’: the earliest attested Sogdian manuscripts are known as the ‘Ancient Letters’ (c.312 CE). These paper documents were found in 1907 by Aurel Stein in Dunhuang, western China

- ‘Upper Indus Graffiti’: appears on more than 600 rock carvings at Shatial and other sites in the Gilgit region of Pakistan (4th-7th c. CE)
Old Sogdian: Research
Old Sogdian: Research
Old Sogdian: Research
Script Encoding Process: ‘Script’

Define the ‘script’:

• Identify repertoire of distinctive signs and variants

• Establish representative glyphs

• Generate script grammar

• Develop encoding model
  • convert conventions for ‘writing’ into ‘typing’
  • translate qualitative data into technical metadata
Script Encoding Process: ‘Script’

Latin script for English: Distinctive signs

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Latin script for English: Variant signs

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
Script Encoding: Unicode

U+0041 LATIN CAPITAL LETTER A
Script Encoding: Unicode

Character-glyph model

Not-plain text  Plain text
Script Encoding Process: ‘Script’

• Is the ‘script’ a distinctive system or a stylistic variant?

• Character-glyph model:
  • encodes a ‘normative’ or representation form for a sign as a ‘character’
  • defines other forms as ‘glyphic’ variants
  • plain text vs. not-plain text / ‘rich text’ (font selection)

• “Unicode encodes characters, not glyphs”
Script Encoding Process: ‘Script’

• A script may have historical, scribal, and stylistic variants

• Considerable differences may exist between early and late forms
  • Old Sogdian and Sogdian

• Script boundaries unclear
  • Form of script in early phase may resemble an ancestor
  • Late form of script may resemble a descendant

• Graphically similar scripts may have structural differences
Developmental ‘Variant’: Vertical Old Sogdian
Developmental ‘Variant’: Formal Sogdian
Developmental ‘Variant’?: Cursive Sogdian
‘Script’ or ‘Variant’?: Chorasmian
‘Script’ or ‘Variant’?: Old Uyghur
‘Script’ or ‘Variant’: Mongolian
Old Sogdian: How to encode?

• Unification with Aramaic?
  • uses characters / code points from Imperial Aramaic block?
  • Palaeographically valid, but should Sogdian be considered a ‘style’ of Aramaic?

• Single encoding:
  • same character set for representing different forms of the script
  • usage of fonts to display different script styles?
  • but, which form is to be representative for plain text?

• Separate encoding:
  • two different character sets for historical stages of Sogdian
  • enables both scripts to be displayed simultaneously, without font changes
Aramaic & Old Sogdian
## Imperial Aramaic

### Central Asian Scripts
- Manichaean
- Marchen
- Mongolian
  - Mongolian Supplement
- Old Sogdian
- Old Turkic
- Phags-Pa
- Sogdian
- Soyombo
- Tibetan
- Zanabazar Square

### Middle Eastern Scripts
- Anatolian Hieroglyphs
- Arabic
  - Arabic Supplement
  - Arabic Extended-A
  - Arabic Presentation Forms-A
  - Arabic Presentation Forms-B
- Aramaic, Imperial
- Avestan
- Cuneiform (1MB)
  - Cuneiform Numbers and Punctuation
  - Early Dynastic Cuneiform
- Old Persian
- Ugaritic
- Hatran
- Hebrew
  - Hebrew Presentation Forms
- Mandaic
- Nabataean
- Old North Arabian
- Old South Arabian
- Pahlavi, Inscriptional
- Pahlavi, Psalter
- Palmyrene
- Parthian, Inscriptional
- Phoenician
- Samaritan
- Syriac
- Syriac Supplement

Imperial Aramaic

<table>
<thead>
<tr>
<th>kāph</th>
<th>yudh</th>
<th>tēth</th>
<th>ḫēth</th>
<th>zain</th>
<th>waw</th>
<th>hē</th>
<th>dālath</th>
<th>gāmal</th>
<th>bēth</th>
<th>ālaph</th>
</tr>
</thead>
<tbody>
<tr>
<td>[k/x]</td>
<td>[j/i/e:]</td>
<td>[tʰ]</td>
<td>[h]</td>
<td>[z]</td>
<td>[w/o/u:]</td>
<td>[h]</td>
<td>[d/ð]</td>
<td>[g/γ]</td>
<td>[b/v]</td>
<td>[ʔ/ʔar/e:]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>tau</th>
<th>shin</th>
<th>rēsh</th>
<th>qoph</th>
<th>sādhē</th>
<th>pē</th>
<th>‘ē</th>
<th>semkath</th>
<th>nun</th>
<th>mim</th>
<th>lāmadh</th>
</tr>
</thead>
<tbody>
<tr>
<td>[t/θ]</td>
<td>[ʃ]</td>
<td>[r]</td>
<td>[q]</td>
<td>[sʰ]</td>
<td>[p/f]</td>
<td>[ʃ]</td>
<td>[s]</td>
<td>[n]</td>
<td>[m]</td>
<td>[l]</td>
</tr>
</tbody>
</table>
## Sogdian Scripts: Repertoire

<table>
<thead>
<tr>
<th></th>
<th>Aramaic</th>
<th>Old Sogdian</th>
<th>Sogdian</th>
</tr>
</thead>
<tbody>
<tr>
<td>aleph</td>
<td>א</td>
<td>א</td>
<td>א</td>
</tr>
<tr>
<td>beth</td>
<td>ב</td>
<td>ב</td>
<td>ב</td>
</tr>
<tr>
<td>gimel</td>
<td>ג</td>
<td>ג</td>
<td>ג</td>
</tr>
<tr>
<td>daleth</td>
<td>ד</td>
<td>ד</td>
<td>ד</td>
</tr>
<tr>
<td>he</td>
<td>ה</td>
<td>ה</td>
<td>ה</td>
</tr>
<tr>
<td>waw</td>
<td>ו</td>
<td>ו</td>
<td>ו</td>
</tr>
<tr>
<td>zayin</td>
<td>ז</td>
<td>ז</td>
<td>ז</td>
</tr>
<tr>
<td>heth</td>
<td>ח</td>
<td>ח</td>
<td>ח</td>
</tr>
<tr>
<td>teth</td>
<td>ת</td>
<td>ת</td>
<td>ת</td>
</tr>
<tr>
<td>yodh</td>
<td>י</td>
<td>י</td>
<td>י</td>
</tr>
<tr>
<td>kaph</td>
<td>ק</td>
<td>ק</td>
<td>ק</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Aramaic</th>
<th>Old Sogdian</th>
<th>Sogdian</th>
</tr>
</thead>
<tbody>
<tr>
<td>lamedh</td>
<td>ל</td>
<td>ל</td>
<td>ל</td>
</tr>
<tr>
<td>mem</td>
<td>מ</td>
<td>מ</td>
<td>מ</td>
</tr>
<tr>
<td>nun</td>
<td>נ</td>
<td>נ</td>
<td>נ</td>
</tr>
<tr>
<td>samekh</td>
<td>ס</td>
<td>ס</td>
<td>ס</td>
</tr>
<tr>
<td>ayin</td>
<td>א</td>
<td>א</td>
<td>א</td>
</tr>
<tr>
<td>pe</td>
<td>פ</td>
<td>פ</td>
<td>פ</td>
</tr>
<tr>
<td>sadhe</td>
<td>צ</td>
<td>צ</td>
<td>צ</td>
</tr>
<tr>
<td>qoph</td>
<td>ק</td>
<td>ק</td>
<td>ק</td>
</tr>
<tr>
<td>resh</td>
<td>ר</td>
<td>ר</td>
<td>ר</td>
</tr>
<tr>
<td>shin</td>
<td>ש</td>
<td>ש</td>
<td>ש</td>
</tr>
<tr>
<td>taw</td>
<td>ת</td>
<td>ת</td>
<td>ת</td>
</tr>
</tbody>
</table>
Encoding: Sogdian and Aramaic

• Differences in repertoire: Sogdian
  • Heth, teth, qoph dropped
  • Daleth, ayin, resh merged
  • Special ayin used for saluations, eg. `D ‘to’
  • Scribal innovations, eg. tail extensions for final characters
  • Special numerical sign for 100

• Cultural and scribal divergences
Old Sogdian: final letters: *aleph*
Old Sogdian: *ayin*
Conclusions

• Script encoding process
  • not art, not science, but a bit of both

• Skills to encode scripts:
  • interest in writing systems and willingness to learn
    • don’t need to be a linguist, typographer, or computer scientist
  • research and analytical skills
    • search, search, and search more, then dig deeper, and follow leads