Script Encoding, Part 2

Working with the user community

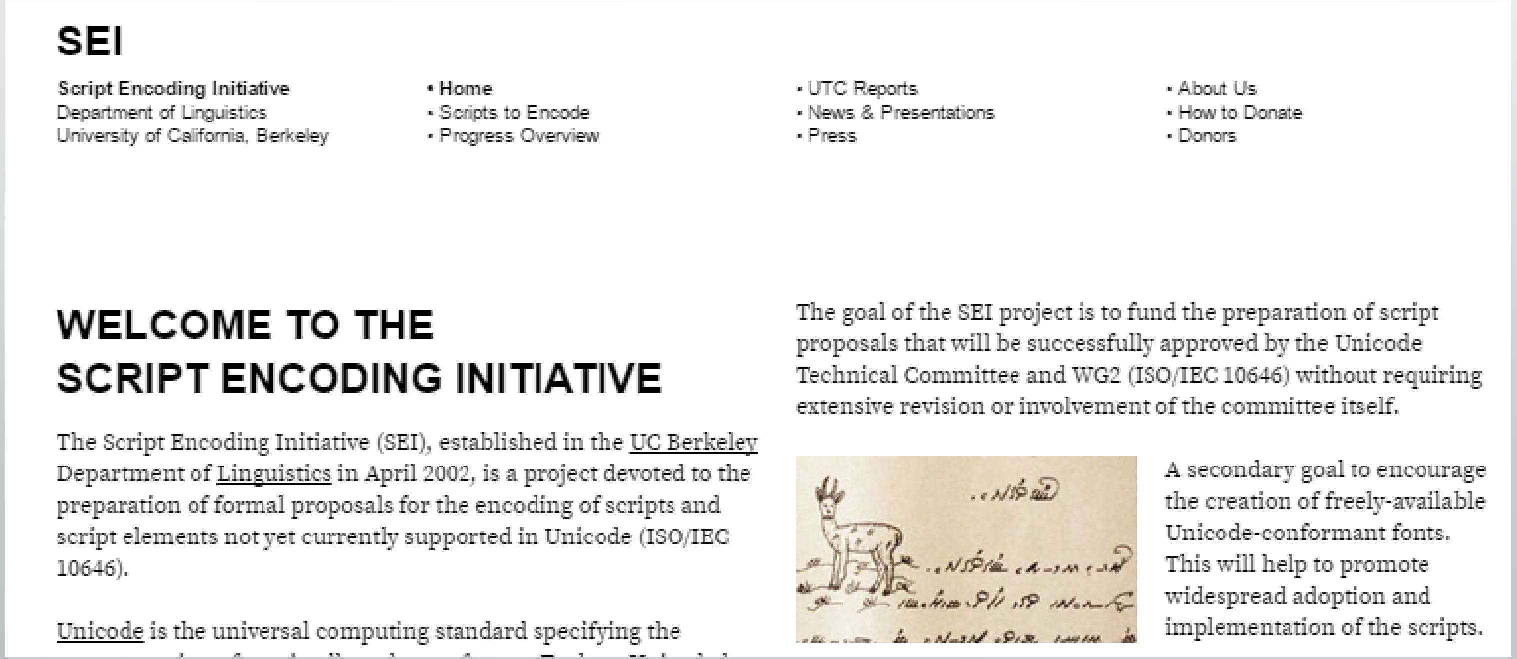
Debbie Anderson, SEI, UC Berkeley

Internationalization and Unicode Conference 40

2 November 2016

Script Encoding Initiative, UC Berkeley

- Started 2002
- Helped get over 70 scripts into Unicode
- 100+ scripts remain to be encoded



SEI

Script Encoding Initiative
Department of Linguistics
University of California, Berkeley

WELCOME TO THE SCRIPT ENCODING INITIATIVE

The Script Encoding Initiative (SEI), established in the UC Berkeley Department of Linguistics in April 2002, is a project devoted to the preparation of formal proposals for the encoding of scripts and script elements not yet currently supported in Unicode (ISO/IEC 10646).

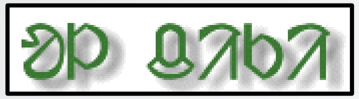
Unicode is the universal computing standard specifying the

The goal of the SEI project is to fund the preparation of script proposals that will be successfully approved by the Unicode Technical Committee and WG2 (ISO/IEC 10646) without requiring extensive revision or involvement of the committee itself.

A secondary goal to encourage the creation of freely-available Unicode-conformant fonts. This will help to promote widespread adoption and implementation of the scripts.

A few words about scripts…

- Can carry significant emotional feeling
  - Ol Chiki



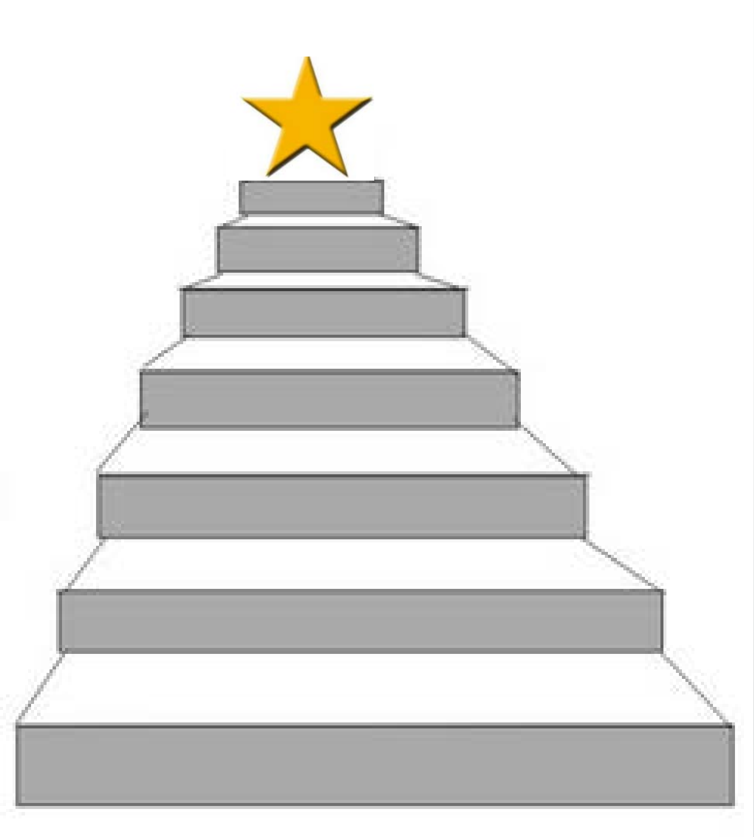
- Even if the “user” can’t read the script , script can be a symbol of identity & pride
- Can make one community different from another
  - But a new script can delay its use on devices

Bamum

Who make up the "user community"?

- Anyone with an interest in the script:
  - linguists, native users, liturgical script users, librarians, historians, script enthusiasts…
- May not be able to actively read and write the script
- To assist on Unicode proposals, should have very good working knowledge of script

Steps to Encoding a Script

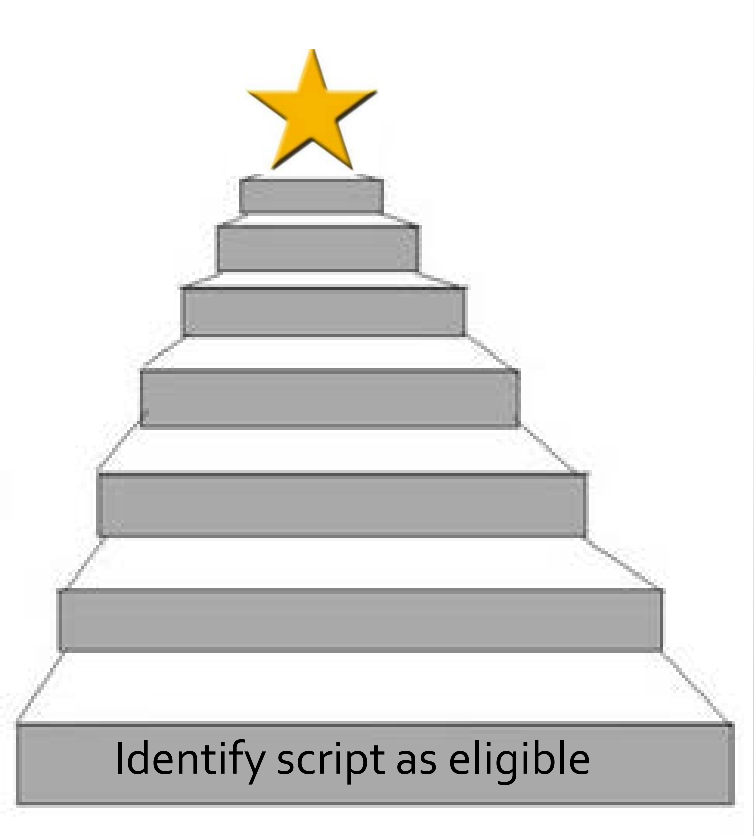


Steps to Encoding a Script:

Identify script as eligible

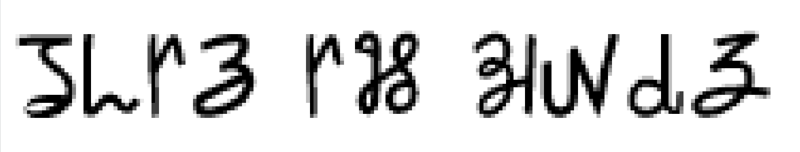
Factors:

- Users (beyond creator and few others)
- Printed materials in script
- Taught today (esp. new script)
- Script relatively stable
- Not unifiable with another encoded script

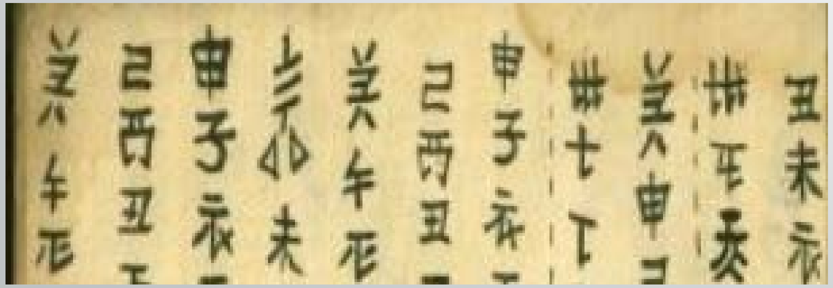


Steps to Encoding a Script:

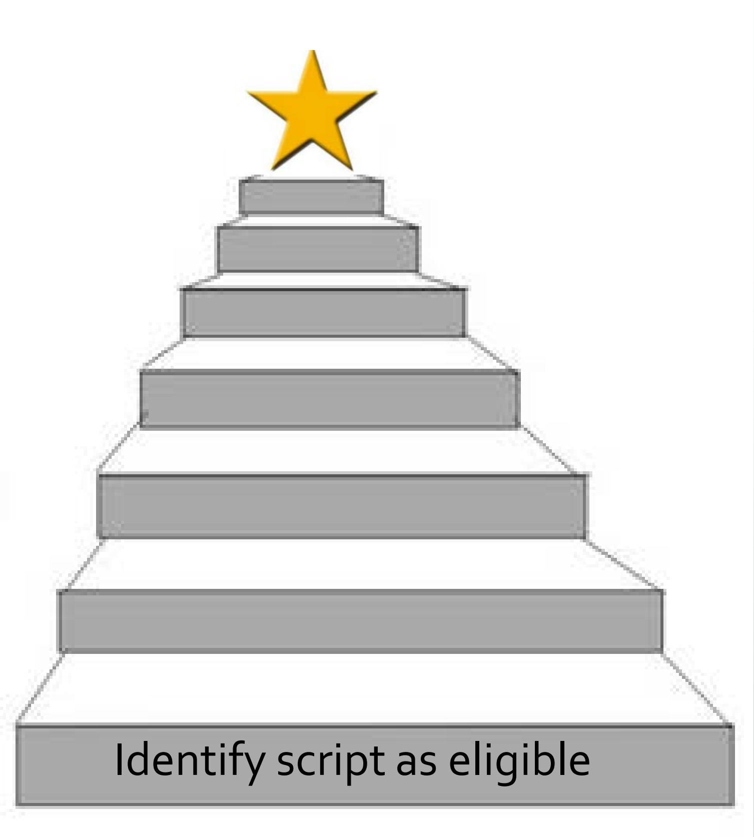
Identify script as eligible



Lakhum Mossang - Tangsa

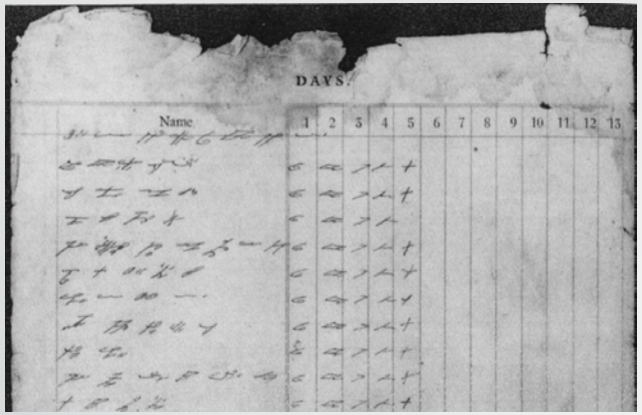


Shuishu

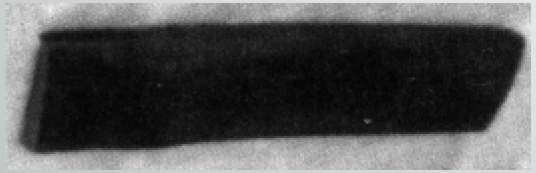


Steps to Encoding a Script:

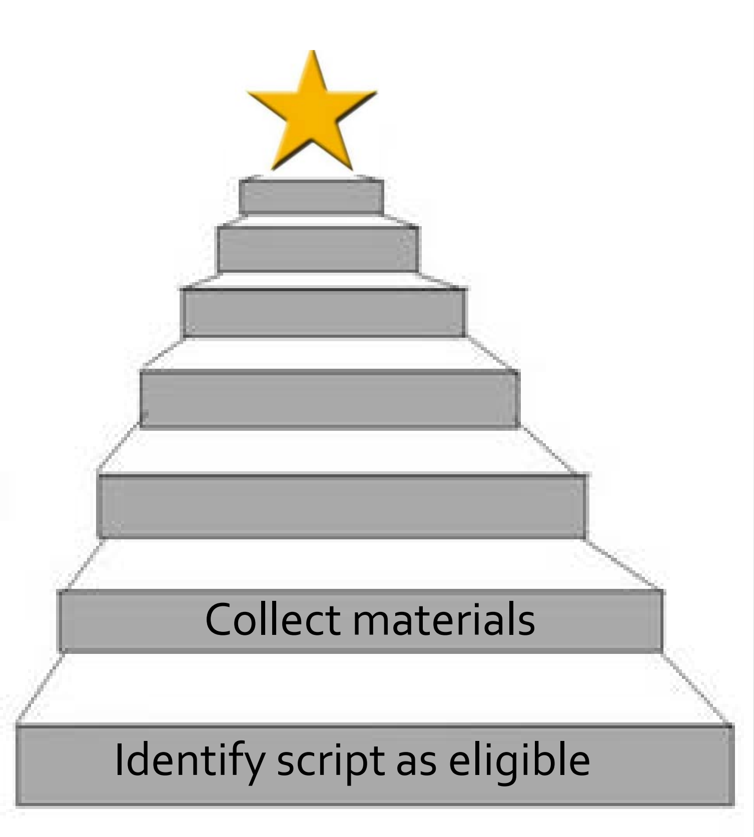
Collect materials


DAYS.
Name
1
2
3
4
5
6
7
8
9
10
11
12
13

Kpelle



Khitan Large Script

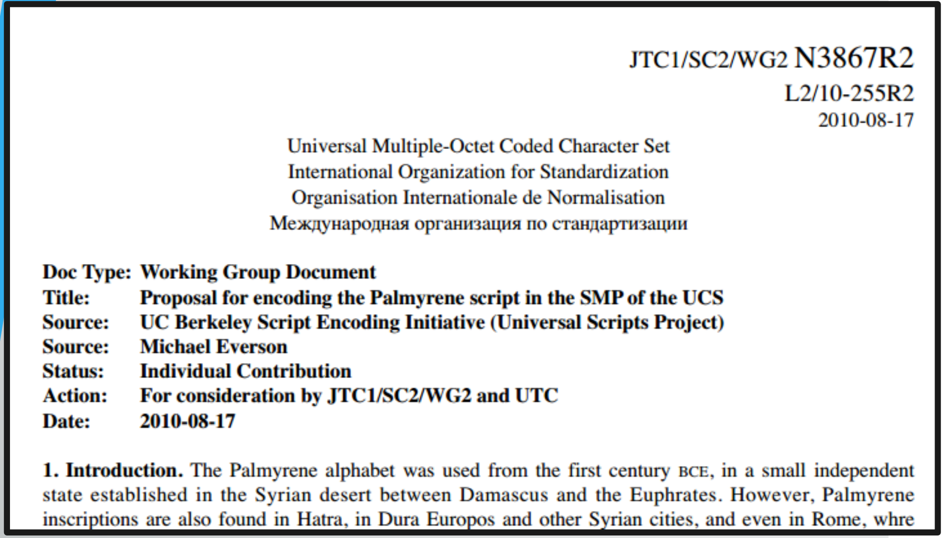


Collect materials

Identify script as eligible

Steps to Encoding a Script:

Write proposal

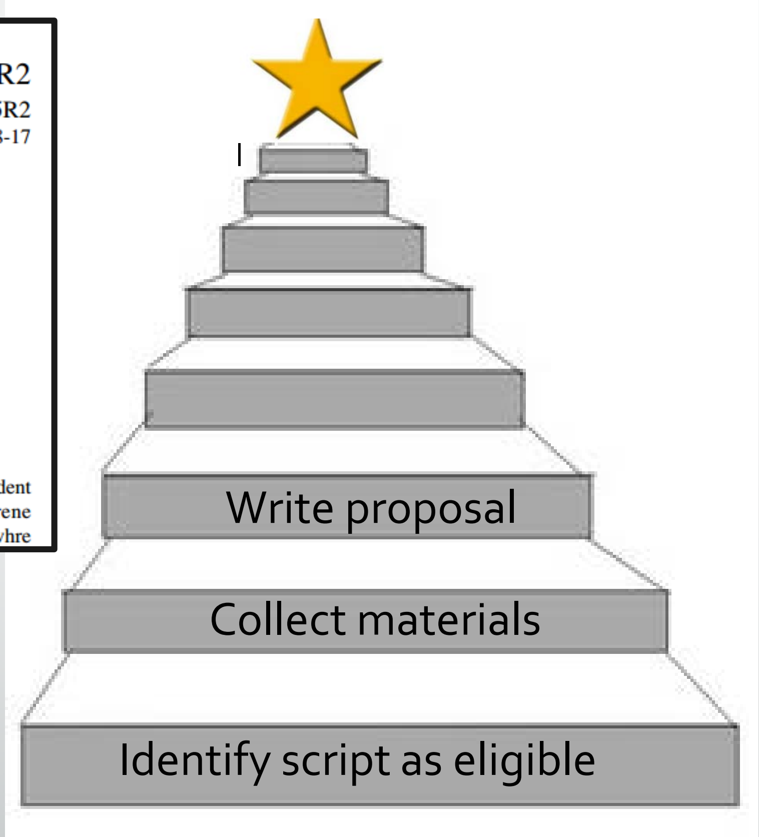


JTC1/SC2/WG2 N3867R2
L2/10-255R2
2010-08-17

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Proposal for encoding the Palmyrene script in the SMP of the UCS**
Source: **UC Berkeley Script Encoding Initiative (Universal Scripts Project)**
Source: **Michael Everson**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2010-08-17**

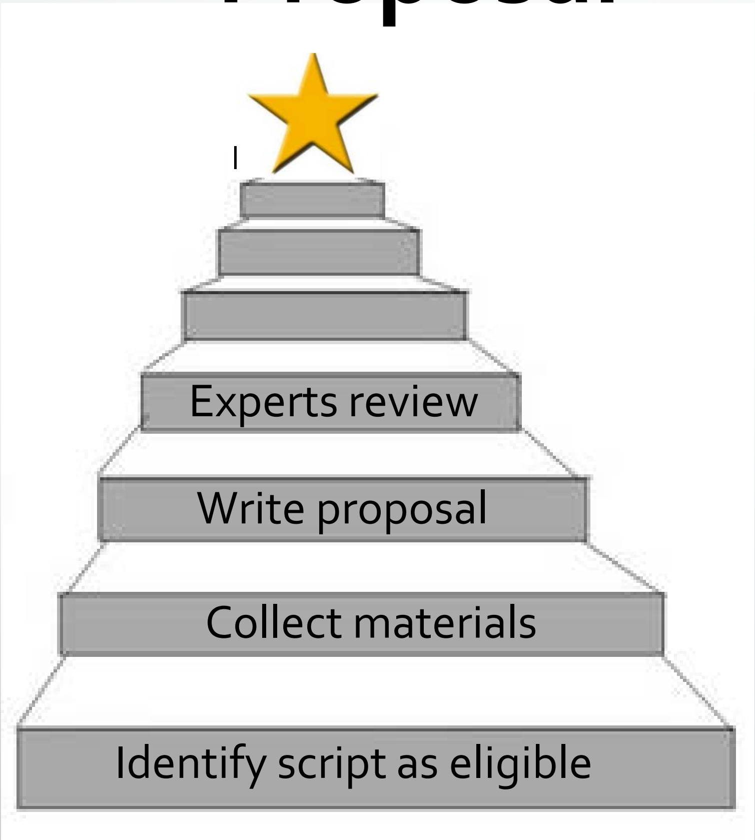
1. Introduction. The Palmyrene alphabet was used from the first century BCE, in a small independent state established in the Syrian desert between Damascus and the Euphrates. However, Palmyrene inscriptions are also found in Hatra, in Dura Europos and other Syrian cities, and even in Rome, whre



Steps to Encoding a Script:

Get Experts / User Community to Review Proposal

Consult
Users / experts



Sidebar: Locating the "user community"

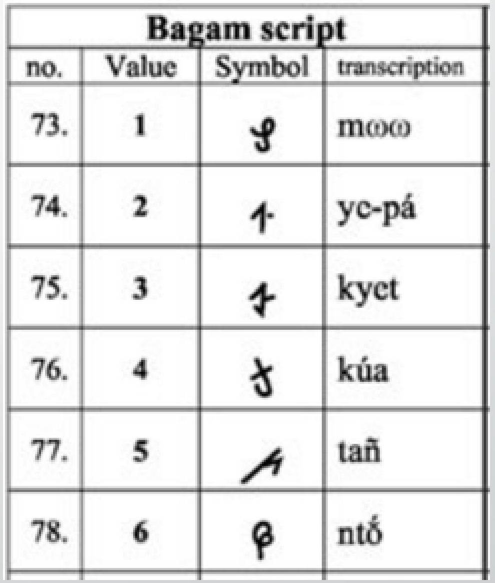
- Contacts from colleagues at Department of Linguistics, UC Berkeley
- Google – find authors of books or articles on the script
- Online projects at institutions
- Can be from outside academia

Sidebar: Screening the "user community"

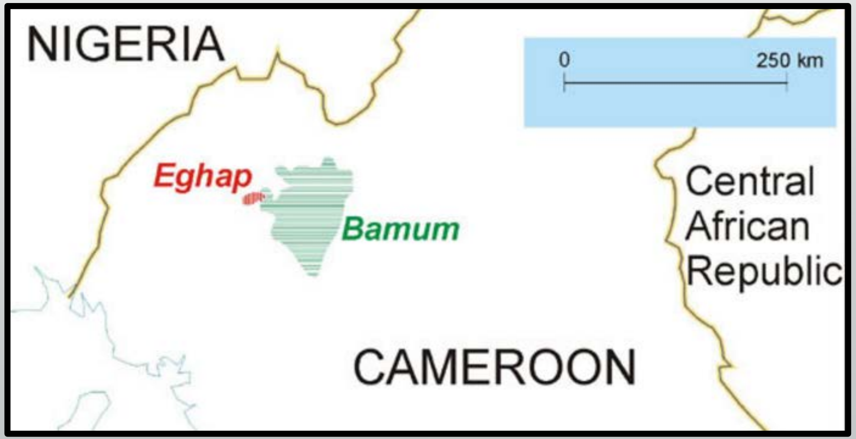
- Verify they know script, have mainstream views, no agenda
- Try to locate several people to review proposals
- For modern scripts: Find a third-party to help understand script's usage and its current social and political context

Sidebar: What if no “user community” can be found?

- Script proposal put on hold
  - Example: Bagam (Eghap) script, Cameroon
- Publicly post document summarizing info on the script
- Encourage anyone working on the script to contact proposal author



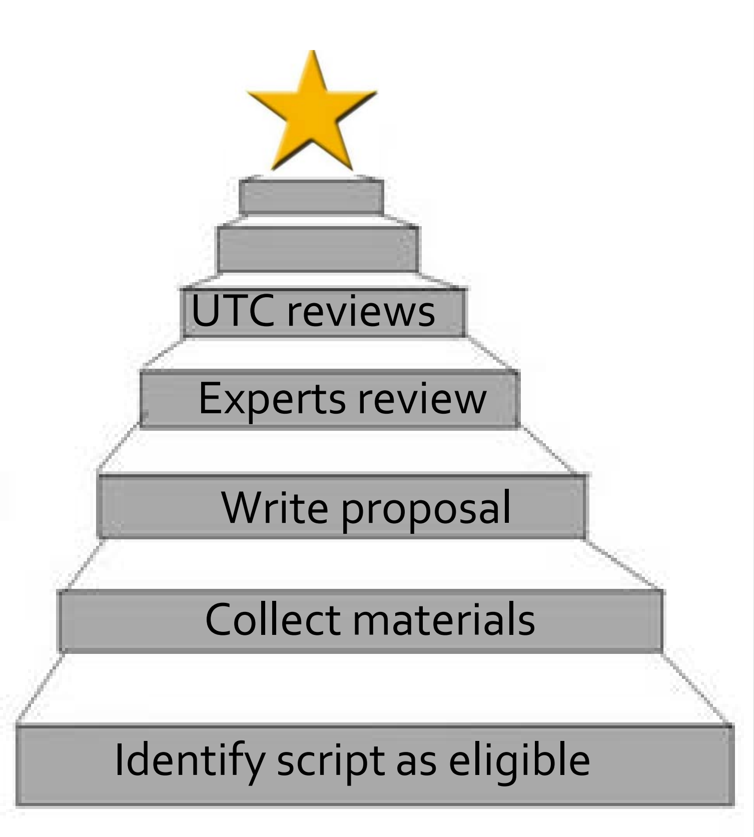
Bagam script			
no.	Value	Symbol	transcription
73.	1		mωω
74.	2		yɛ-pá
75.	3		kyɛt
76.	4		kúa
77.	5		tañ
78.	6		ntő



Steps to Encoding a Script:

Script ad hoc & UTC reviews proposal

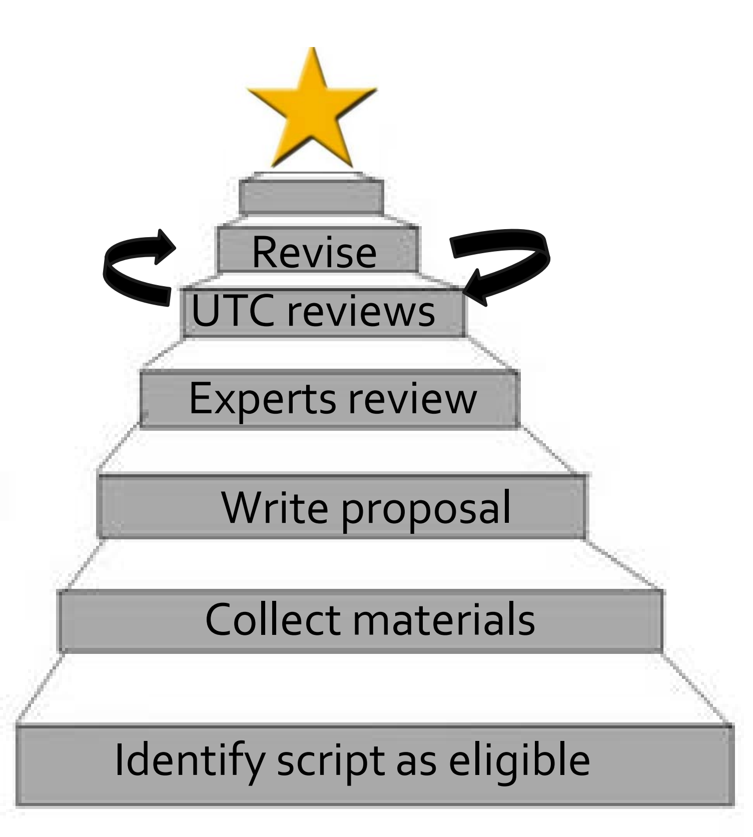
UTC meeting



Steps to Encoding a Script:

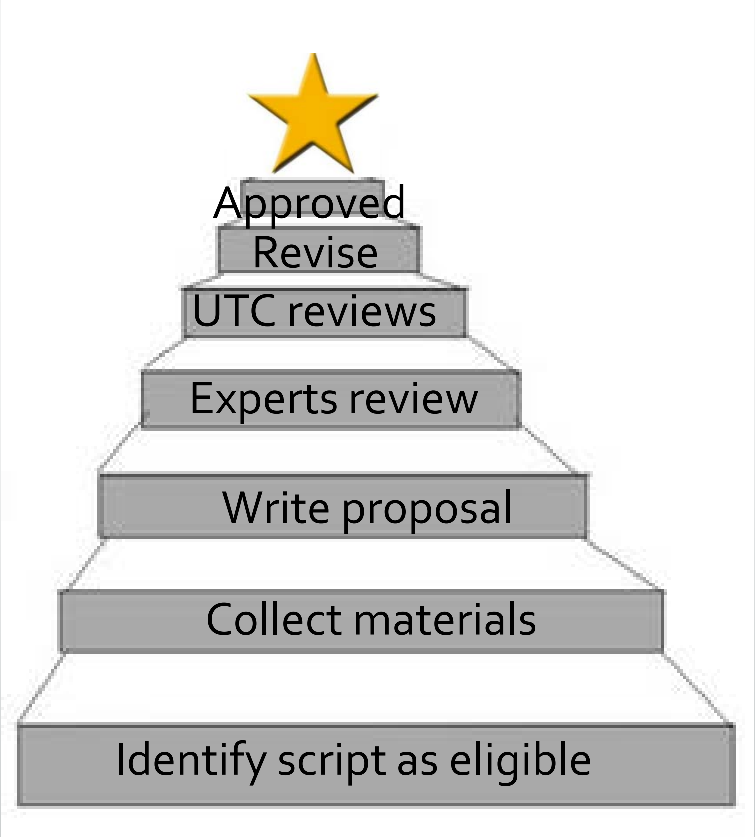
Revise Proposal / Get UTC Review

UTC meeting



Steps to Encoding a Script:

Get UTC Approval!



Approved

Revise

UTC reviews

Experts review

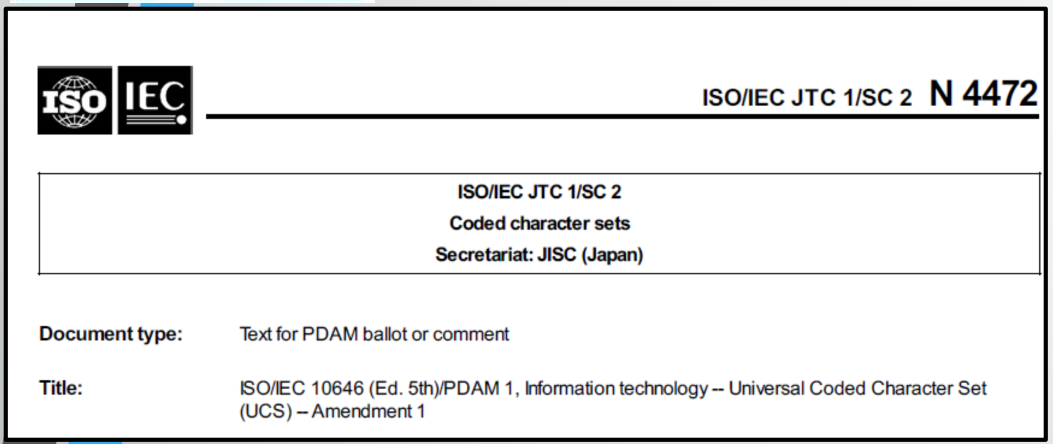
Write proposal

Collect materials

Identify script as eligible

Steps to Encoding a Script:

Get onto ISO Ballot and be subject to review by National Bodies



ISO/IEC JTC 1/SC 2 **N 4472**

ISO/IEC JTC 1/SC 2
Coded character sets
Secretariat: JISC (Japan)

Document type: Text for PDAM ballot or comment

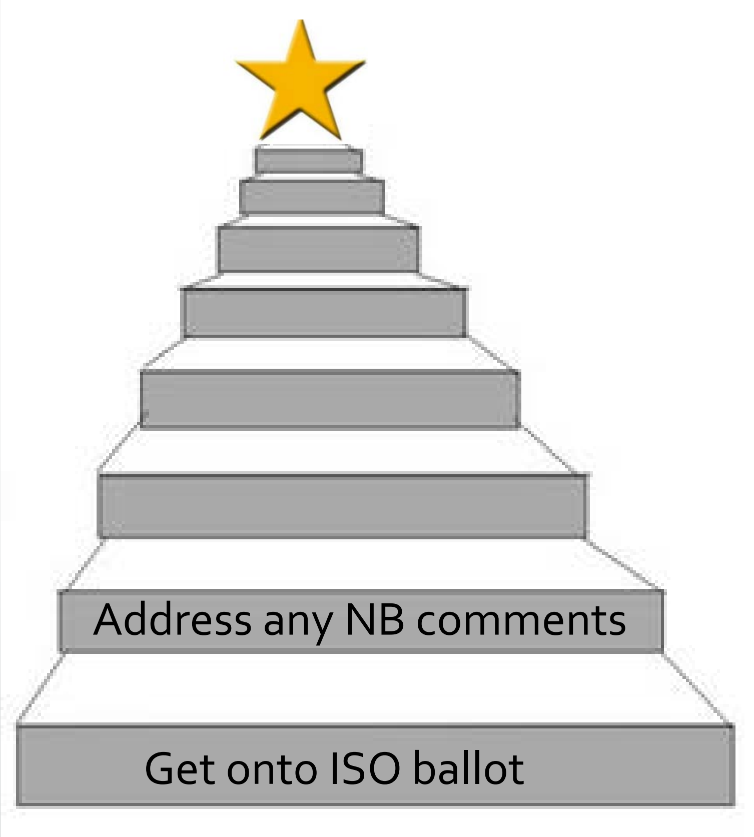
Title: ISO/IEC 10646 (Ed. 5th)/PDAM 1, Information technology -- Universal Coded Character Set (UCS) -- Amendment 1

Get onto ISO ballot

Steps to Encoding a Script:

Address National Body Comments

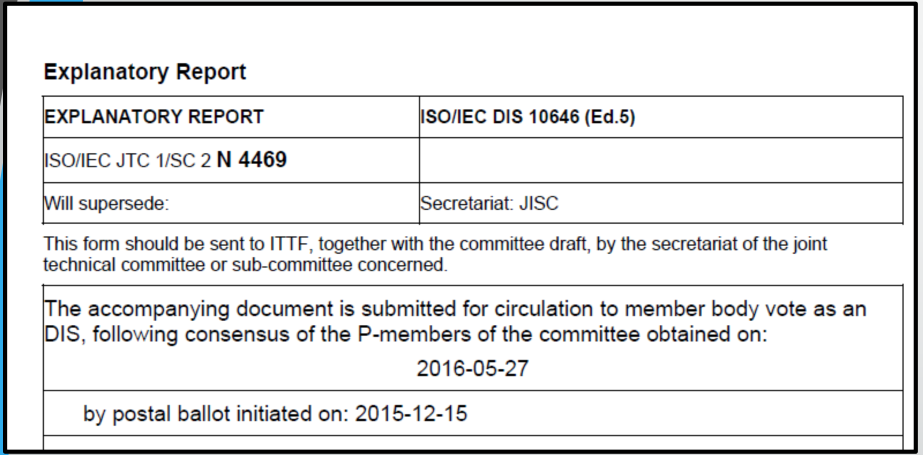
Check with users / experts



Resolve outstanding issues in face-to-face ISO meeting (SC2/WG2)

Steps to Encoding a Script:

(More ISO ballots)



Explanatory Report

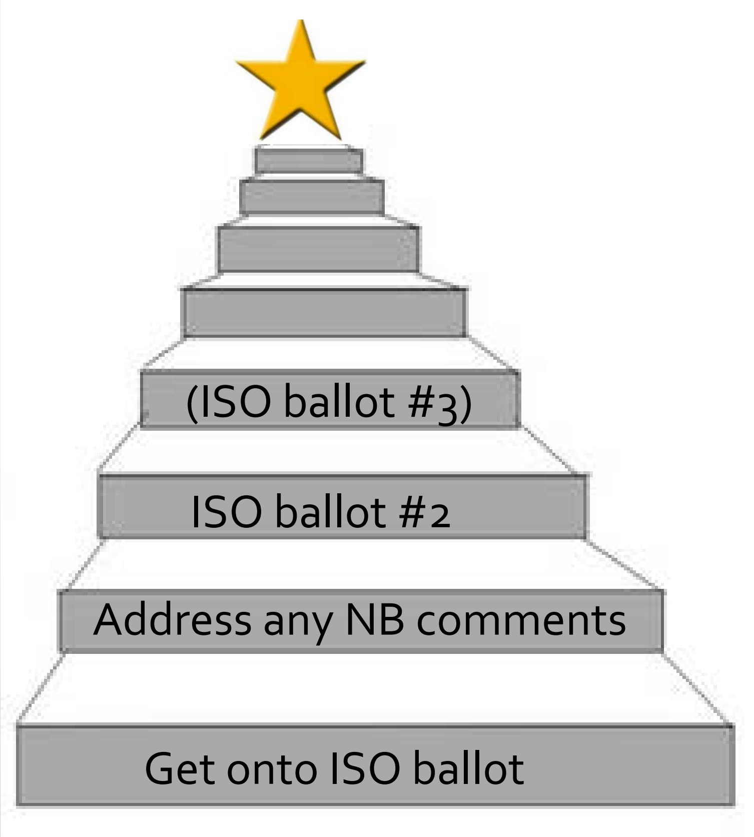
EXPLANATORY REPORT	**ISO/IEC DIS 10646 (Ed.5)**
ISO/IEC JTC 1/SC 2 **N 4469**	
Will supersede:	Secretariat: JISC

This form should be sent to ITTF, together with the committee draft, by the secretariat of the joint technical committee or sub-committee concerned.

The accompanying document is submitted for circulation to member body vote as an DIS, following consensus of the P-members of the committee obtained on:

2016-05-27

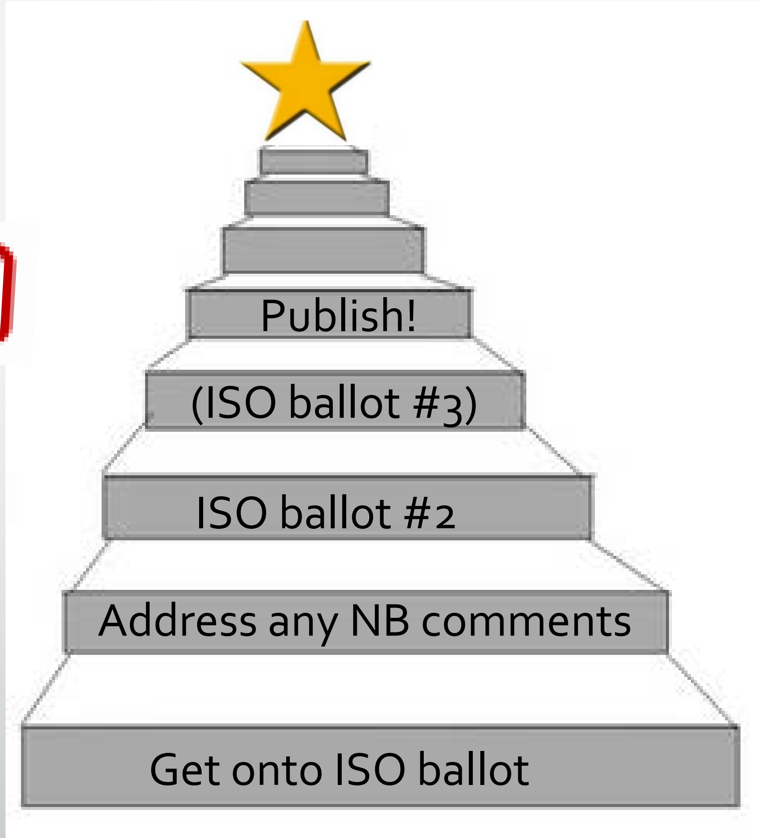
by postal ballot initiated on: 2015-12-15



Steps to Encoding a Script:

Publish in Unicode and ISO/IEC 10646

PUBLISHED



Publish!

(ISO ballot #3)

ISO ballot #2

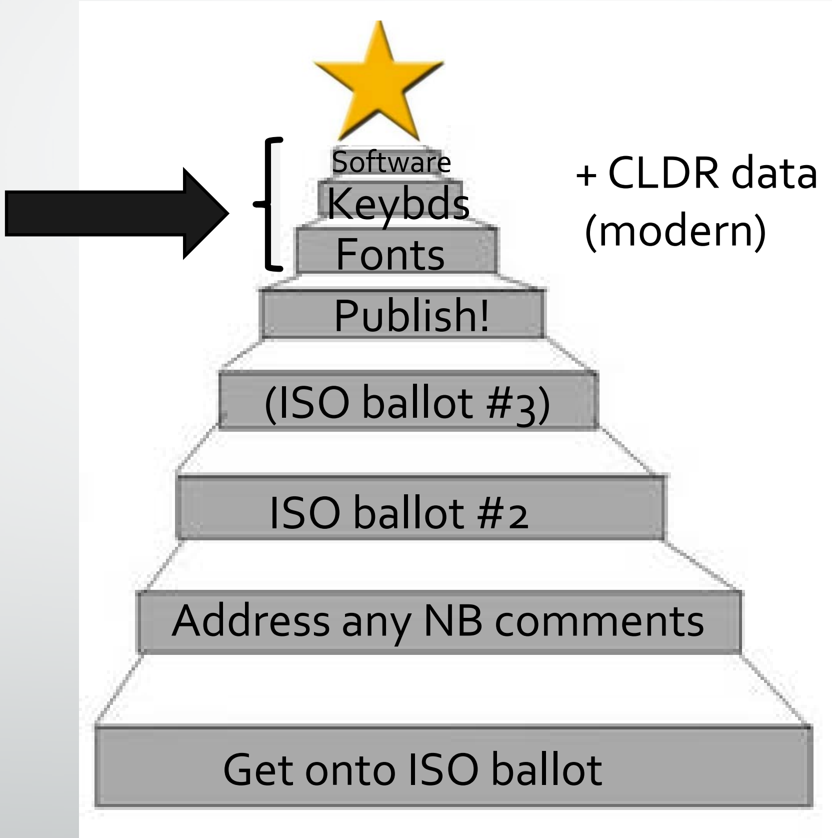
Address any NB comments

Get onto ISO ballot

After Encoding a Script:

Fonts / Keyboards / Software Updates

Users / experts



+ CLDR data (modern)

Issues: Academics

- May not answer emails in a timely manner
- Helpful to give overview of Unicode, ask specific questions
- Tendency to want to capture fine palaeographic detail (for historic scripts) when identifying characters
- View: How can Unicode be making decisions about the script?

Issues: Native users

- May want to get script into Unicode to drive orthographic change or gain political recognition (for the group using the script)
- Name of the script (cf. Lanna = Old Tai Lue = Khün => Tai Tham)
- Other problem areas: confusing language and script (Newa –Tibetan vs Indic model), letter vs. ligature
- View: How can Unicode be making decisions about my script?

When problems arise

- If email doesn’t work, arrange face-to-face meeting

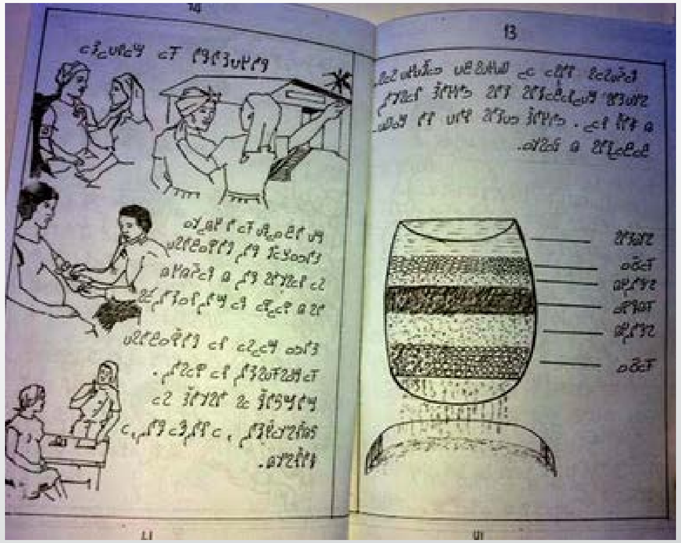
Tips:

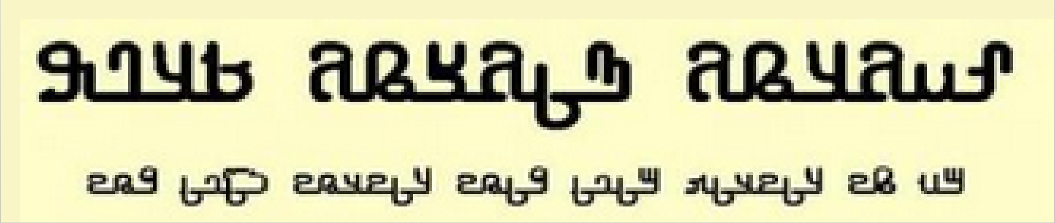
- Explain Unicode basics in easy-to-understand language
- Have translator present (if needed)
- Set aside a few days to build trust and develop relationships
- Listen to and address any concerns
- Stress disagreements will delay approval

Meeting on Tangut, Beijing, 2013

Case Studies

- Adlam
- Bamum
- Egyptian hieroglyphs
- Old Hungarian
- Mende Kikakui
- Ahom





Adlam

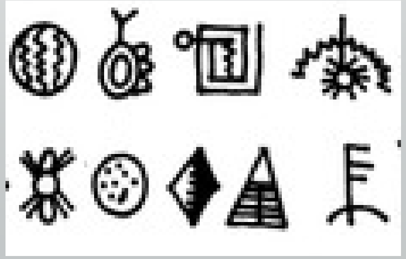
Case Study: Adlam

- Script created in Guinea in 1989 for Fulani by Barry brothers
- First contacted by Adlam creators 2012; 1st proposal 2013, revised 2014
- Creators of script and proposal author attended Oct 2014 UTC meeting
- Published June 2016, Unicode 9.0; **3 yrs**

Case Study : Bamum

- Created ca. 1896 in Cameroon for Bamum language; ideographic > syllabary
- Report on Bamum in 2006; modern syllabary 1st proposed in 2007
- 2007 modern syllabary put onto ISO ballot, then removed
- 2008 users reviewed prop., was revised (publ. Unicode 5.2, 2009) , **2 yrs**;

  2008/9 historic Bamum proposals (publ. Unicode 6.0, 2010), **2 yrs**



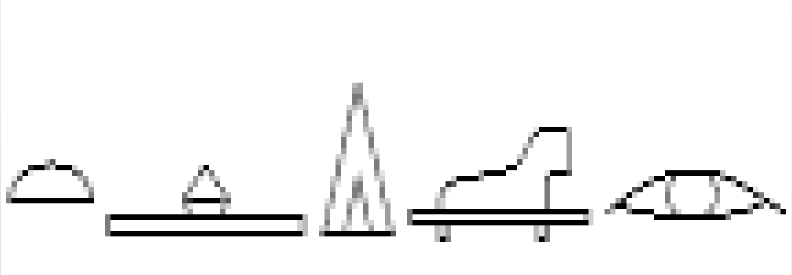
Case Study : Egyptian Hieroglyphs-1

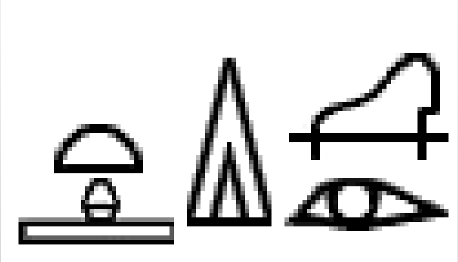
- First proposed in 1997
- Comments from users 1999, meetings 2002 and 2006
- Final proposal 2007; published Unicode 5.2, 2009; **12 years**

Case Study : Egyptian Hieroglyphs-2

Two outstanding issues:

- Current users are using images, not Unicode characters
  - Format characters proposed 2015, approved 2016, now under ISO ballot





- Only 1071 characters in Unicode, missing later period characters
  - Large set of extensions proposed in 2015/16 based on widely-used font, Hieroglyphica

Case Study: Old Hungarian

- Dates to at least 13c; recent revival 1990s
- 1st Proposal ca. 1998
- 2008 meeting in Budapest resulted proposal, later 2 alternative proposals
- 2012 Hungarian National Body divided on name, etc.
- Published Unicode 8.0, 2015; **17 years**

2 Further Case Studies (Post –encoding)

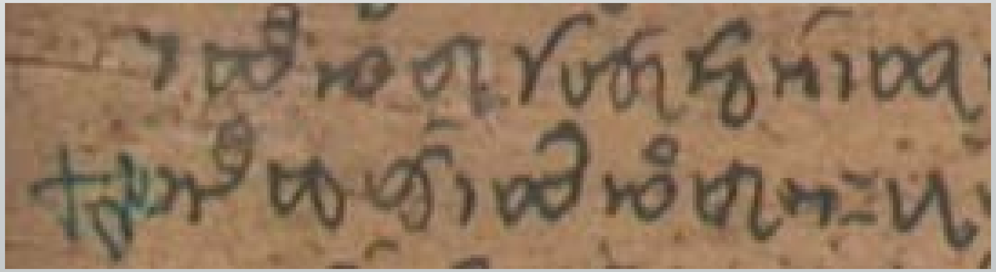
Case Study (Post –encoding): Mende Kikakui

- Originally created ca. 1917 in Sierra Leone for Mende language
- 1st proposed in 2010, revisions 2011, 2012
- Publ. Unicode 7.0, 2014
- Still lacking primers and written materials

Guinea
Sierra Leone
Liberia
Atlantic Ocean

Case Study (Post –encoding): Ahom

- Dates to 15-16c; appears in inscriptions and manuscripts
- 1st proposed in 2010, revisions 2012; approved by UTC 2012
- Publ. Unicode 8.0, 2015
- Widely used legacy font; needs Unicode font in style of legacy font (and keyboard and converter)



In sum -1

- Listen to users and address their concerns
- Engage with user communities early, if possible, and keep them in the loop
- Be inclusive, so needs of all users are taken into consideration
- Stress overall goal –get script approved, published, and implemented
- Remember script belongs to the user community

In sum-2

- Very important to work with the user community
- Without their input
  - Script may be encoded with errors
  - Difficulty in getting script passed by standards committees
  - Fonts and software may not be relevant or widely accepted

Thank you!
Questions?

Email: dwanders@berkeley.edu

Website: http://linguistics.berkeley.edu/sei

Support for the project has come from NEH Preservation and Access grant PR-50205-15 (and donations)