

<http://www.unicode.org>



Unicode, Inc, is a California non-profit corporation that was formed for the purpose of developing and maintaining a computer standard called The Unicode Standard. Non-profit doesn't mean it's a charity – it is a 501 c(6). That is, a “business league” or consortium formed of members who pay annual dues.

What does the Unicode consortium do?

"What is Unicode?" in 60 seconds

Like the familiar ASCII: a set of characters for text, but

- ASCII is little:
128 characters
- Unicode is **big**:
94,000 characters
...and still growing

Most people who use computers regularly are familiar with ASCII. ASCII is the set of characters that are standard in American computer systems. ABC, dollar sign, at sign, and so forth. Beyond this set are various other character encodings. Some people might be familiar with "DOS code pages", IBM's EBCDIC, or know that the Windows PC has a different set of characters than Apple's Macintosh.

The world is full of different, mostly incompatible, character encodings. Unicode is distinguished from the rest of them by being universal. Unicode can be thought of as a superset of all the character sets that anyone anywhere on the planet has been using.

Unicode is like plumbing

- Character encoding is at the heart of all your data usage. WWW. XML. HTML. Office products. Maybe your cell phone.

Windows 2000 or XP

Microsoft Office 2000 or XP

Mac OS X

Internet Explorer... Netscape...

all of these products incorporate Unicode support

UTF-8

UTF-16

Unicode is a universal character encoding. A character encoding is like plumbing or electrical wiring. It isn't glamorous, but it's necessary for other things, like running water and lights. Unlike other high-profile buzzwords in the industry – XML, WWW, HTML and so forth, Unicode pretty much stays in the background. It's a feature that is most visible to the people who write software, and not so visible to people who simply use computers in their daily lives. But users benefit – even for English – by having many more characters available than ever before.

So, who uses Unicode? If you use any of the products I've listed, you're already using Unicode.

Unicode also has been known to masquerade by obscure acronyms such as UTF-8 or UTF-16. You might be familiar with those terms.

Unicode is like plumbing

- Character encoding is at the heart of all your data usage. WWW. XML. HTML. Office products. Maybe your cell phone.
- Global data exchange requires broad cooperation.
- Unicode is a cooperative effort involving countries, companies, and people

The latest generation of operating systems and office products are all starting to implement Unicode. How did this come about? Unicode had its beginnings over ten years ago, and has been a large cooperative effort. Many companies who are fierce competitors in the software and hardware markets have come together with surprising unanimity in this venture. Everyone has realized that data exchange and interoperability is everyone's problem, and can only be solved by cooperative effort.

Some of the players in the Unicode consortium over the last dozen years have included companies like IBM, HP, Sun, Microsoft, Apple, Xerox, Compaq. The list is long and includes both hardware and software vendors. Also involved is the International Organization for Standardization – usually referred to as “ISO”. The members of ISO are countries, and the standards they make are often required for compliance to national or local laws and even other standards. This also means that character encoding is not only a technical issue, but also a political and social issue.

Unicode lowers the barrier to entry into the “information society”. Already Unicode has tackled enough scripts that all of the big problems have been solved. Scripts come in a limited number of types, even though they show a great surface variability. So at this point, the addition of a new script is practically guaranteed to work within one of the models that are already in use. Therefore, if a software system already has good Unicode support, it is not usually difficult to add support for a new script.

In the beginning...
Multi-lingual data was not portable

It's just Greek: ἄνδρα μοι ἔννεπε

Or is it?: ǎíäñá ïë é ǎííåđå

Anyone in the humanities who works with a non-Latin script, such as Greek, is familiar with the problem illustrated on this slide. You take your data from one machine to another – or from one country to another – or down the hall to your colleague's computer, and suddenly it doesn't look right. This kind of problem typically results from the use of different character encodings or "hacked" encodings that depend on having someone's home-brewed fonts. In an environment that supports Unicode, this kind of thing is not supposed to happen.

Unicode is intended to solve a large-scale problem in the world: interoperability of systems across not only national boundaries, but among different linguistic communities and with software from different vendors.

Unicode encodes *scripts*

French, Indonesian	Latin
Russian, Azeri, Bulgarian	Cyrillic
Farsi, Pashto	Arabic
Hindi, Nepali, Marathi	Devanagari

Because of its universality, a common misconception is that Unicode encodes languages, and people newly acquainted with Unicode sometimes say "I can't find my language in Unicode! How could they have missed it!". In fact, many languages use the same script. This table illustrates a few languages and their scripts, from which you can gather the basic idea. There are a lot fewer scripts than languages.

Within a script, there may also be many alphabets – which are the ordered subsets of letters from a script that are used to write particular languages. One minor point to consider here is that of ordering. There is a completely separate standard that deals with character ordering and sorting, because obviously with so many different languages using these scripts, no single ordering of the character encoding will make sense in all situations.

What languages / scripts does Unicode support?

Answer: probably all you need!

- All speech communities > 5 million
- Over 50 scripts
- Several ancient scripts
- Final target is “everything”
 - Script Encoding Initiative

Unicode already supports over fifty scripts, representing the scripts used by most of the world’s population. The ongoing project hopes to encode all living scripts as well as all significant extinct scripts. At this time, all speech communities of more than 5 million persons are supported by Unicode. That still leaves a lot of work to do, but it’s a good start.

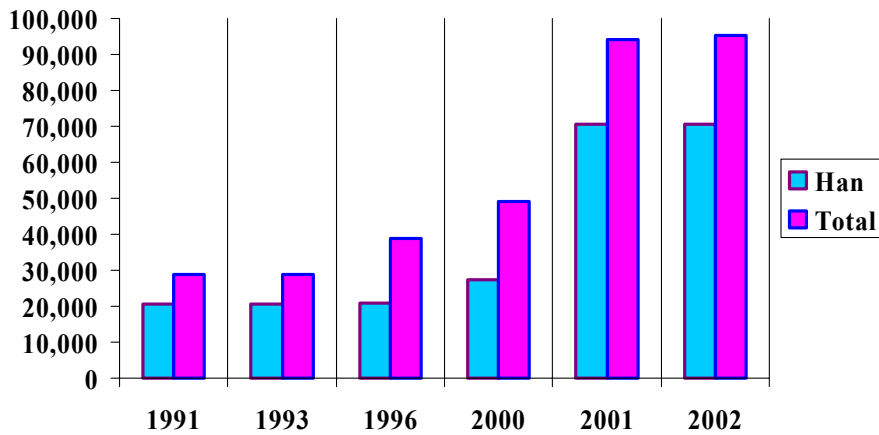
A bit later, Deborah Anderson will talk about an initiative we have started in the hope of completing the standard.

Some scripts already supported

Chinese/Japanese (or “Han”), Arabic,
Armenian, Bengali, Burmese, Cherokee,
Cyrillic, Devanagari, Ethiopic, Georgian,
Hangul, Hebrew, Kannada, Malayalam,
Mongolian, Runic, Syriac, Tamil, Telugu,
Thai, Tibetan, Yi...

As you can see from this slide, the list of scripts is pretty long. This is just a sampling.

Growth of Unicode



The Unicode Standard is over 10 years old, and there are now nearly 100,000 things in it.

Most of what needs to be encoded right now are minority scripts, because Unicode already covers all major scripts.

State of the Project 2002

- **52 scripts already encoded** ★
- **~96 scripts *not yet* encoded**
- **~25 modern minority scripts**

***Most of these will still be
un-encoded in 10 years***

We have encoded a lot of scripts: 52 of them. The figure for scripts already encoded is exact, but as you can see, the other numbers are approximate. Currently there are (I believe) 96 unencoded scripts on the list we call our “roadmap”. Many of those are extinct scripts of historical and scholarly interest. But even the number of scripts in actual use keeps changing. We sometimes learn about another minority script in India or elsewhere. So we put that on the roadmap and increment the number.

Of the twenty-five modern scripts, most will still be unencoded in ten years. You might ask why? We’ve done a lot of work, but have run into a lot of limitations. Encoding everything takes time and research.

For each new proposal to encode a script, there is a lot of research involved. Many technical details of implementation and character repertoire need to be resolved. We have reached the limit of what can be done with volunteer effort and low-level corporate backing.

Items needed for Unicode support

- Fonts
 - Many come with operating systems
- Keyboards
 - Check the vendor website first
 - Tavultesoft Keyman 5.0
- Text editors
- Browsers
- Mail programs

Someone setting out to use Unicode for producing or sharing documents with colleagues worldwide will need a few items – fonts, keyboards, text editors, browsers, mail programs. The URLs listed here, available from the Unicode home page, contain links to many resources that can help people get up and running with Unicode.

Modern operating systems such as Windows XP and Mac OS X come already equipped with fonts for handling large portions of Unicode. The vendor web sites also usually contain links to even more free fonts that are not included in the base operating systems, so the first place to look for fonts should be the platform vendor, who also usually supplies keyboard layouts.

Font and Keyboard Sources

- Fonts
 - Monotype, Microsoft, Apple
 - Code2000 by James Kass
- Keyboards
 - Check the vendor website first
 - Tavultesoft Keyman 5.0
- <http://www.alanwood.net/unicode/>
 - Alan Wood's excellent pages

Fonts are fairly easy to find, but fonts specifically useful with Unicode can be more difficult. The Unicode web site has pointers to a number of good sources for fonts.

Probably the best resource I can offer is Alan Wood's web site devoted exclusively to Unicode fonts and keyboards, along with setup tutorials.

Editing Text

- Word 2000, 2002, XP
- Apple “TextEdit”
- Yudit, BBEdit, Mined2000, etc.
- Many resources available here:

<http://www.unicode.org/onlinedat/products.html>

For most platforms these days, at least one text editor exists which is capable of editing Unicode text. Probably the most widely used is Microsoft Word on both Intel and Apple platforms. Apple also ships a text editor with Mac OS X. On Linux and Unix platforms much less Unicode-capable software is available, but the “mined2000” editor can support a variety of scripts, and runs under X-Windows.

Sending e-mail

Many mail programs support “UTF-8” format

- MS Outlook (Express)
 - Apple “Mail.app”
 - Netscape built-in mailer
 - Any mailer that supports attachments
-
- But most “free” web-based mail services don’t have Unicode support.

The most widely used mail programs on the popular platforms already support sending and receiving mail in Unicode, as long as it’s encoded in what is known as “UTF-8” format. The Netscape browser’s built-in mailer supports Unicode to some extent – including versions 4.7, 6, and 7 of Netscape. At this point, however, the web-based e-mail services typically don’t have Unicode support built-into them. However, if your mail program can do attachments, you can always attach Unicode-encoded documents.

Unicode website as a resource

- Display Problems
- Where is my Character?
- Useful Resources page

<http://www.unicode.org/onlinedat/products.html>

- Enabled Products page

<http://www.unicode.org/onlinedat/resources.html>

- The Unicode mail list

The Unicode website is full of helpful information about using Unicode. There is a page devoted to display problems. In addition to a page that lists many products with Unicode enablement, there is a useful resources page that lists fonts, keyboards, and a number of helpful files with setup instructions for various products.

The Unicode Consortium has also long supported a mail list and discussion forum, where users can go to ask questions and discuss the standard. This is a great resource for beginners, and the archives are on-line with entries dating back to 1994.

<http://www.unicode.org>



That concludes my presentation. Thank you for your time. If anyone has questions I will be happy to talk later. Now I'll turn the floor over to Deborah Anderson.

The Script Encoding Initiative at U.C. Berkeley

Deborah Anderson
Researcher
Dept. of Linguistics
UC Berkeley

I'll be speaking briefly about the Script Encoding Initiative, a project I established at UC Berkeley through the Department of Linguistics. This project is being run by myself in collaboration with Rick.

The main objective of the project is to oversee the inclusion of all those scripts missing from Unicode.

Scripts Missing From Unicode

see <http://www.unicode.org/sei/alpha-script-list.html>

Modern minority scripts:

- Pahawh Hmong (Vietnam)
- Cham (Vietnam)
- N'ko (W. Africa)
- Saurashtra (India)
- Meithei Mayek (India)

Ancient scripts:

- Egyptian hieroglyphs
- Mayan hieroglyphs
- Sumero-Akkadian cuneiform

Unicode has currently about 52 scripts, but over 90 remain outstanding.

To borrow Rick's metaphor: the underlying system of plumbing has lots of gaps in it.

This effectively means that making online teaching materials available or using email is nearly impossible in these scripts.

There are two groups missing: modern minority scripts comprise one set.

The other group is of historic scripts: Egyptian hieroglyphs, Mayan hieroglyphs, and cuneiform used by Sumerian and Akkadian.

The Script Encoding Initiative

- Goals:
 - Raise money for Unicode proposals and fonts
 - Work with individuals/groups to guide them through the international standards process
- Needs:
 - Help spread the word
 - If anyone is working on teaching materials for a language whose script is missing, he/she should include a budget item for a Unicode proposal.
 - Need greater support from the university

What this project aims to do is to provide a centralized approach to covering the missing scripts.

Currently only a few scripts/year are included into Unicode, so it will take well over 20-40 years to cover the remaining scripts.

There are two goals for the project:

(a) raise funding for graduate students, faculty, and experts to write proposals and to create fonts for the missing scripts;

and

(b) work with any individual and group to guide them through the international standards process, so the script can be included into Unicode;

My goal in speaking here is to inform you about the project and ask that if anyone on your home campus is developing teaching materials for a language whose script is missing from Unicode, he/she should consider including a line-item in the budget for a Unicode proposal. We can assist in writing a proposal and can help find an experienced proposal author or font creator.

By getting the script into Unicode, the underlying “plumbing” will be in place not only for online teaching being developed at the UC campus, but will make it accessible for anyone who wants to use the script for learning materials, online publication, communication, etc.

Contact:

Script Encoding Initiative
c/o Deborah Anderson
University of California, Berkeley
Department of Linguistics
1203 Dwinelle Hall #2650
Berkeley, CA 94720-2650

Email: dwanders@socrates.berkeley.edu

Web:

www.linguistics.berkeley.edu/~dwanders

A final comment:

I feel strongly that the push for including scripts into Unicode should come from the university, where teaching materials are developed, language courses taught, and research and publication in these scripts is carried on.

The computer industry is not particularly interested in these scripts now (or at least in contributing financially to the effort), probably because they don't reflect a major consumer market.

UC is poised to take the lead in this effort, in its support of Unicode and getting the missing, but we are in need of a stable home.

If you have any ideas for funding or thoughts on how to make the Script Encoding Initiative a more permanent presence within the UC system, I would appreciate hearing them.

If you have any questions or comments, please feel free to contact me (next slide) or Rick.

Script Encoding Initiative

Department of Linguistics
U C Berkeley

Welcome to the home of the Script Encoding Initiative

Contents of this Page:

[Announcement](#)
[List of Scripts Needing Encoding](#)
[How to Help](#)
[Who We Are](#)
[How To Contact Us](#)

Announcement of the Script Encoding Initiative (April 17, 2002)

A new Script Encoding Initiative has been set up at the Department of Linguistics of the University of California at Berkeley. The charter of this initiative is to fund proposals for those scripts missing in Unicode (and its ISO counterpart, 10646), the universal character encoding standard.

To date, Unicode has largely focused on the major modern scripts. Some minority and historic scripts have already been encoded, as well as historic characters of the major modern scripts. At least 90 scripts remain to be encoded. Minority scripts still used in parts of South and Southeast Asia, Africa, and the Middle East include Balinese, Batak, Chakma, Cham, Meithei Mayek, New Tai Lu, N'Ko, Pahawh Hmong, Pollard, Siloti Nagri, Tifinagh, and Vai. Scripts of historical significance include Aramaic, Avestan, Brahmi, Egyptian Hieroglyphics, Glagolitic, Javanese, Kitan, Lanna, Lepcha, Old Permic, Pahlavi, 'Phags-pa, Phoenician, South Arabian, Sumero-Akkadian Cuneiform, and Tangut.

Because proposals for the encoding of minority and historical scripts often entail significant research, and their user communities have little economic or political voice, script proposals have not been submitted to the Unicode Technical Committee (UTC) in any regular manner. It has been estimated that at the current slow pace of encoding, many scripts will still be unencoded in ten years.

Alphabetical List of Scripts Not Yet Encoded

The *Type* field indicates the basic script type (A) alphabets & abjads, (G) abugidas, (L) logosyllabaries, (I) ideographic systems, (S) syllabaries. The secondary values are "h" for historical scripts and "m" for living minority scripts.

The *Plane* field indicates whether the proposal is suggested for encoding on the BMP (0) or the SMP (1).

The *Chart Font* field indicates whether a font acceptable for printing the code charts in the standard has been obtained.

The *Free Font* field indicates whether a Unicode-compatible font for the proposed encoding (or PUA facsimile) is available for download by the general public at no charge.

(The *Proposal* and *Free Font* fields will contain links to the latest proposals or fonts if available.)

The *Contact* field is the name of the central contact person in charge of coordinating activities related to the script proposal.

Script Name	Type	Plane	Proposal	Chart Font	Free Font	Contact
Ahom	G m	1	-	-	-	-
Alpine scripts (see note 1)	A h	1	-	-	-	-
Aramaic (see Phoenician)						
Avestan	A m	0	Y	Y	-	-
Aztec Pictograms	I h	1	-	-	-	-
Balinese	G m	1	-	-	-	-
Balti	G m	1	Y	Y	-	-
Bamum		1	-	-	-	-
Bassa		1	-	-	-	-
Batak	G m	0	Y	Y	-	-
Blissymbols	m	1	-	-	-	-
Box Headed (see Chalukya)						

This is the website with an alphabetical list of the missing scripts and the current status.