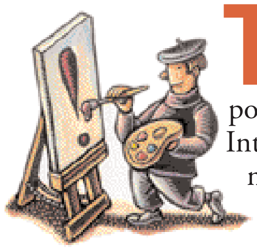
Deborah Anderson

Global Linguistic Diversity for the Internet

The Script Encoding Initiative wants to encode the world's unencoded scripts, making electronic text communication possible for millions of native speakers.



LISA HANEY

The Internet has made written communication more global, but financial support is still needed to make Internet-based written communication more accessible to everyone. The development and adoption of the international character-encoding standard known as Unicode in computer software and fonts makes it possible to send and receive—and read—text electronically for hundreds of languages, all in their original scripts. Chinese, Greek, German, and Arabic can now be included in Web pages and email without the text getting garbled or marred with square boxes as placeholders. This multilingual capability frees users from ASCII's limited repertoire. Communicating with people in their own language and script also gives the world's diverse populations an electronic presence in the global information economy.

The character-encoding standard underlying this multilingual capability is the Unicode Standard (www.unicode.org), which also has a parallel International Standard maintained by the International Organization for Standardization, called ISO/IEC 10646. With Unicode, every letter, sign, or symbol of a given writing system (or "script") receives a unique number used by every computer. Because Unicode is a single international standard, it marks a significant improvement over the earlier situation in computing where competing standards rendered the interchange of text often nearly impossible.

Unicode encodes scripts, not languages. The Latin script, for example, is used for English, Lithuanian, and Albanian, while the Arabic script is used for the Arabic language, as well as for Persian, Kashmiri, and Pashto. To date, more than 50 scripts have been included in Unicode, with space available for all the other identified scripts of the world, past and present. Most encoded scripts were selected for inclusion because they are used in languages spoken by more than five million people. Still, more than 80 scripts remain outside the standard, locking out their users from the very capabilities the Internet makes available universally.

To rectify this situation, I began (in 2002) the Script Encoding Initiative at the University of California, Berkeley, seeking to get all the missing scripts into the international standard (www.linguistics.berkeley.edu/sei). Approximately one-third of them are in active use today, most by groups in Asia and Africa. The rest is historic, including Egyptian hieroglyphics and ancient scripts of the Middle East, including Hieroglyphic Luwian.

While the popular media has focused on the effort to save biological diversity and endangered languages [2], the case for preserving the writing systems of languages is largely unnoticed. Saving scripts by including them in Unicode will help document the variety of writing systems while also enabling their study, appreciation, and use. The Rosetta Stone was inscribed more than 2,000 years ago in three scripts—Greek, Egyptian hieroglyphs, and Demotic—yet only Greek is included in Unicode. Hence, accessibility to two-thirds of the text is missing. By including all the historic scripts in Unicode, the entire original text will be available to everyone electronically.

In Bali, Indonesia, the Balinese script, which is used in many cultural and literary works, is taught in the schools. Yet students' fluency is poor and getting worse, due in part to the fact that the national language of Indonesia—Bahasa Indonesia—is written in Latin letters and predominates in schools and government offices. The Balinese community itself identifies the Balinese script as endangered and wants the script encoded in Unicode so additional learning materials and newspapers can be published in Balinese, thereby reinvigorating the study—and use and appreciation—of the script.

Being able to write and read texts in their original scripts has important practical ramifications. For example, being able to download health care materials, including those about AIDS, in one's native language could be a lifesaver, particularly in remote geographic regions where interpreters are unavailable. Likewise, being able to use the Internet to communicate with people in isolated parts of the world could be critical in times of natural disaster or war.

Access to one's own script also has the effect of empowering users, particularly minorities. Posting online documents from a person's own culture serves to affirm that person's cultural identity and pride. It can also encourage literacy efforts. The Santals of eastern India, for example, have a literacy rate of 10% to 30% (as reported by SIL Ethnologue, www.ethnologue.com/), and members of the community are eager to have their script—Ol Chiki—included in Unicode. Doing so would make it possible to create and distribute more school reading materials. Having a script in Unicode would also promote free speech and public discourse. Because the Arabic script is in Unicode, the Iranian-born blogger Hossein Derakhshan (aka "Hoder") was able to blog in the Persian language and show others how to do the same [1].

Another reason to complete the encoding of all scripts follows from the increasing use of the Internet to post and store text documents. For example, the eScholarship Repository program (begun in 2002), sponsored by the California Digital Library, is an open-archives project that requires papers be submitted in PDF format with the fonts embedded (scholarship.cdlib.org/). However, because nonstandard fonts are used for unencoded scripts, accurate searching of scripts by online readers is impossible, discouraging many from even trying. Encoding all scripts would improve search and display beyond the 50-plus scripts covered today. As documents are increasingly made available on the Internet, accurate searching is vital for any kind of online research or indexing.

One problem faced by the Script Encoding Initiative in trying to raise awareness and secure funding involves explaining the value and scientific basis of character encoding, a topic often compared to plumbing. The average user simply isn't captivated by the topic. The importance of character encoding—the Internet's backbone for text communication—still eludes many foundations and grant agencies. The academic environment, particularly at the administrative level, has also been slow to grasp the importance of actively participating in and supporting Unicode development. The study of the world's languages and the literary, historical, and cultural documents in them are often found only at major research institutions. Indeed, its future may ultimately be found only in the electronic world.

The project to encode missing scripts has a limited timeframe—10 years at most—while expertise stays focused on the problems and the economic incentive is clear. Work on missing scripts must be done now. Corporate interest and participation in the Unicode Consortium is relatively strong, and there is widespread interest in expanding the international capabilities of the major software platforms. However, as interest inevitably wanes among the larger supportive companies in the computer industry, chances are that lesser-known scripts will not work seamlessly with off-the-shelf software.

The goal is to catch the wave, encoding the scripts and making the Internet a domain that reflects the world's true linguistic diversity. C

References

1. Glaser, M. Iranian journalist credits blogs for playing key role in his release from prison. *USC Annenberg Online Journalism Review* (Jan. 9, 2004); www.ojr.org/ojr/glaser/1073610866.php; Hoder's English Web site www.hoder.com/weblog/ and the Persian Web site i.hoder.com/.
2. Knight, W. Half of all languages face extinction. *New Scientist* (Feb. 16, 2004); www.newscientist.com/news/news.jsp?id=ns99994685.

Deborah Anderson (dwanders@socrates.berkeley.edu) is a researcher in the Department of Linguistics at the University of California, Berkeley.