For the World's A B C's, He Makes 1's and 0's

To Call Cyrillic, Chinese or Cherokee to the Screen, Typographer Helps Forge a Digital Lingua Franca

By MICHAEL ERARD

MICHAEL EVERSON, a 40-year-old typographer who lives in Dublin, considers himself blessed because he has found his life's work: to be an alphabetician to all the peoples of the world. Mr. Everson's largest project to date — a contribution to a new version of Unicode 4.0, an international standard for computerizing text — is cementing his reputation.

His mission has taken him to Kabul, Afghanistan, and Helsinki, Finland; to Beijing, Tokyo and Redmond, Wash. His Dublin house is a shrine to his obsession with every writing system that humans are known to have created — 148 of which Mr. Everson says he can use for writing his name. In the hallway is an icon of the saints Cyril and Methodius (Cyril is often credited with inventing the Cyrillic alphabet) and a page from a Maghreb manuscript from North Africa.

He keeps a photo of a stone inscribed with ogham, an ancient Irish alphabet that looks like hash marks, in a silver frame. His office chair, parked in front of a Macintosh G4 laptop named Cyril, is upholstered with dark blue fabric dotted with Egyptian hieroglyphics. Surrounding his desk are shelves heavy with books on the origins of cuneiform and other writing systems.

He remains fond of the Roman alphabet, however. "Of all the alphabets, it's the best one," he said in a telephone interview.

For the last 10 years, Mr. Everson, who has American and Irish citizenship, has played a crucial role in developing Unicode, which might be viewed as the computer age's Rosetta stone. Mr. Everson explains Unicode as "a big, giant font that is supposed to contain all the letters of all the alphabets of all the languages in the world."

A more technical explanation of Unicode is this: When Mr. Everson sends e-mail in ogham, his computer isn't sending ogham letters through the ether. Instead, strings of 0's and 1's are transmitted, and when they arrive on a friend's computer, they generate on its screen the same ogham letters that Mr. Everson typed. Unicode is the master list that resides in both computers and translates individual letters and symbols into strings of 0's and 1's and back again. Most current software is Unicode-compliant, which means that this master list of all the world's writing systems has been built into operating systems, browsers and software.

The code assigned to all 96,000 characters is

STANDARD SETTER Michael Everson translates alphabets into code that can be read by computers. A key to the alphabets at right, Page 6.

Translating All the World's A

handled only by programmers in its naked form, while computer users (and sometimes vendors) install the specific fonts that represent a specific alphabet. A font renders a language readable to humans; Unicode renders a font readable to computers.

Most people don't even realize Unicode is at work. "Unicode is like plumbing," said Rick McGowan, the vice president of the Unicode Consortium. "Yet it's the most far-reaching and ambitious multilingual project in history."

It is because of Unicode that bloggers can muse in Arabic and domain names can exist in Chinese, or that National Security Agency analysts can scour the Internet for reports on the latest threats in East African newspapers. "Because of Unicode," Mr. McGowan said, "you can plunk down a vanilla off-the-shelf computer into a cafe anywhere in the world and have any user in any language walk up to it and use it for accessing the Web."

Mr. McGowan was a member of the group of computer scientists and linguists who set out to create the system in 1990 to solve an emerging problem.

As a growing number of users wanted to write in their own languages on their machines, companies had developed methods for computerizing text that did not appear in the Roman alphabet. With the rise of the Internet, the problem became more complicated because there was no assurance that all those machines would be able to share text data. Without a shared standard, manufacturers and even governments were creating isolated islands of data, each with its own standard, and each computer would have to be cus-

The evolution of Unicode, a computer-age version of the Rosetta stone.

tomized to the writing system that the owner wanted to use. Many users could not write e-mail, build Web sites or search databases in their own languages and alphabets.

The solution was Unicode, an international standard for character encoding. (Character encoding is simply any system that transmits textual information; Morse code is one example.) Last month the latest version of the standard, Unicode Standard Version 4.0, was published. It contains encodings (that is, unique strings of 0's and 1's) for some 96,000 letters and symbols. Approximately 70,000 of them are Chinese characters. Unicode also contains support for 54 other writing systems, from Mongolian to Thai to Gothic to Cyrillic.

Mr. Everson said he had worked on about 5,000 of those characters. Version 4.0 includes characters for Linear B (for which he designed the font) and other ancient Mediterranean alphabets that are used mainly by scholars.

As vast as Version 4.0 seems, it is still not complete, and nearly 100 writing systems remain to be encoded. Mr. Everson is haunted by the prospect that Unicode may never be finished. "Imagine how you would feel if your name was François, but there was no ç available," Mr. Everson said. "You would be irritated that your phone bill came addressed spelling your name wrong. Now imagine if your language used a totally different alphabet and you couldn't use computers at all because of it. It's a question of human rights, really."

An incomplete Unicode is a looming possibility, however. Now that the writing systems of the major computer markets are encoded, the computer companies that once backed the Unicode project are beginning to question the expense. To ensure that the remaining writing systems are included, a project named the Script Encoding Initiative has been set up at the University of California at Berkeley to enlist scholars and apply for funds from private foundations to hire Mr. Everson full time.

One result of the dwindling interest from the private sector is to put pressure on Mr. Everson to complete large projects. "They say, 'Here, Michael, can you do Egyptian?' It's like, no. Egyptian is on my list, Egyptian is hard, and it's big."

To pay the bills, Mr. Everson works as a typesetter. He is currently setting type for "Gargantua," by Rabelais, in Irish. Other notable projects include the first publication of the entire New Testament in Cornish, as well as an English-Cornish dictionary.

But Mr. Everson admits that he is most drawn to the encoding work. "It's best for me in my life to be consumed by an obsession in writing systems, because I am extraordinarily well suited to dealing with it," he said.

TRANSLATION Michael Everson at his home in Dublin, a shrine to his obsession with writing systems.

Unlocking the Code

To include scripts and characters in the Unicode standard, each character is given a hexadecimal code position — that is, one using a base-16 notation, incorporating the digits 0 to 9 and letters A through F. The character at the top right, from Cherokee, is 13E3. That code is then distilled into 1's and 0's.

Represented above are, top row, Limbu, Linear B, Cherokee; second row, Tai Le, Cypriot, ogham; third row, Osmania, Ugaritic, Tibetan; bottom row, Tagalog, Sinhala, Gothic.

Mr. Everson was first attracted to far-off places and languages by the books of J. R. R. Tolkien, which he first read as a 13-year-old living in Tucson. (Mr. Everson said he still has a "soft spot" for Tengwar, one of the alphabets that Tolkein invented for his made-up languages of Sindarin and Quenya.) "The Lord of the Rings" led him to Anglo-Saxon and the epic poem "Beowulf," which he decided to translate from Old English at the age of 14. From his copy of the "Beowulf" manuscript, he practiced copying Anglo-Saxon letters with a calligraphy pen.

Then he graduated to designing fonts on his Macintosh, tackling Georgian and Cyrillic, then Devanagari. After feeling dissatisfied with graduate school at U.C.L.A., he moved to Ireland in 1989 and began typesetting for a living while designing exotic fonts on the side for writing systems including Cherokee, ogham and Sinhala.

In 1993, he saw a request from the Unicode Consortium for revisions involving some archaic scripts, one of which was ogham. "It was like, ooh, this is Irish, let's look into it," Mr. Everson recalled. He also sent comments on Burmese, Ethiopic, Yi and Sinhala. "I started in early," he said. "I just plunged right in."

Meanwhile, at the Unicode offices in Silicon Valley, people were impressed with the work by this relative unknown in Ireland. Mr. McGowan remembers the first proposal he received from Mr. Everson, on a particular character in ogham. The first time they met, Mr. McGowan was so captivated by Mr. Everson's charm and erudition that he saved his name tag. "Michael is a pretty special guy," he said. "Also, he wrote the month with a Roman numeral. I thought that was amusing."

Mr. Everson's knowledge of the world's writing systems has made him indispensable to Unicode. "At this point, Michael is probably the world's leading expert in the computer encoding of scripts," Mr. McGowan said. "Nobody else comes close to having his detailed knowledge about so many scripts and how they are, or should be, encoded."

Deborah Anderson, a researcher in the linguistics department at Berkeley who heads the Script Encoding Initiative, credits Mr. Everson with getting most of the lesser-known writing systems into Unicode. His 220 proposals or technical documents, she said, make him "without question the single most prolific Unicode proposal author around."

As exotic as the subject matter is, the work itself is fairly dry. It involves finding authoritative texts, assembling examples and seeking out experts and then working with them to determine how many characters there should be and how they should look.

"It's one thing to be a specialist who reads Ugaritic," Mr. Everson said. "It's another to be a person who can figure out the essential bits of the writing system in terms of the way Unicode works."

He takes a remarkably long view of the impact of his work. "There's satisfaction in knowing that the work of analyzing and encoding these languages, once done, will never need to be done again," he said. "This will be used for the next thousand years."

Mr. Everson also seems to enjoy the human interactions. He is proud of working with the grandson of Osman Yusuf Kaynandid, who invented the Osmanian script in Somalia in 1922. He also likes to tell about how he met the president of the Tibetan Calligraphy Society at a Unicode meeting in Copenhagen. Mr. Everson had helped the organization ensure that Tibetan was included in the standard. The president showed Mr. Everson how to write his name in Tibetan with a highlighter pen.

"He thanked me," Mr. Everson said with reverence. "I couldn't believe that, because his organization has been in existence for over a thousand years."