Multilingualism and IDNs in the Internet Age

Qiheng Hu
Vice President, Chinese Academy of Sciences

In order to build "a people-centred, inclusive and development-oriented Information Society, where everyone can create, access, utilize and share information and knowledge…", as stated in the WSIS Declaration of Principles, the issue of multilingualism on the Internet is becoming more and more important, since the Internet has evolved into the most widespread global infrastructure for the modern human society and its users have the background of unprecedented and growing diversity of languages and cultures.

Multilingual Communication

According to a September 2004 size-up (Almanac 2004), the number of the world's population now online is approaching one billion. In 2003, it was estimated that roughly two-thirds of all Internet users were non-native speakers of English (CyberAtlas, 2003). The native speakers of English no longer dominate the Internet, as they have had for many years. English is the official or dominant language in only four (the U.S., the U.K., Canada and Australia) out of fifteen top countries which accounted for 70% of Internet users in 2004.[1] Among the non-English-speaking countries, only China, Korea and Japan together accounted for nearly a fifth of the total number of individuals online. Anyhow, although the Internet is created mainly in the U.S. and is planned originally for the ASCII characters, hundreds of millions of people access and communicate over the Internet in languages other than English – or in some acceptable or even badly distorted English – because they did not learn it as a first language. Such a big benefit for non-English-speaking people is gained mainly because of the development and advancement of the Unicode, which is developed and maintained by the Unicode Consortium (http://www.unicode.org/).

[1] Brenda Danet (Yale University) and Susan C. Herring (Hebrew University of Jerusalem / Indiana University, Bloomington), "Multilingualism on the Internet"

This encoding standard provides the basis for processing, storage and interchange of text data in any language in all modern software and information technology protocols. In the foreseeable future, as Unicode continues to make progress, more and more individuals online in the world will be able to access the Internet using their own languages.

Multilingual Access to Namespace

The multilingual problems remain open in the domain name systems, mainly because the successful global deployment involves not only technical difficulties but also widespread and complicated issues of policy nature which requires extensive international coordination.

The issue of multilingual domain names has not been considered a problem at the early stage of Internet spread even in non-English-speaking countries, when the dominant users are mainly limited to the circles of so-called elites in the society – the university professors, well-educated officials, scientists etc. They are not very much concerned that the language used on the Internet is mainly English. This became a problem when great amount of people with lower levels of education come online. People realize that having a name in their native language is much more comfortable and easier to communicate online.

Perhaps this is why, while the web content is already available in different languages, the multilingualization of the DNS and email addresses has just begun. For the widest throng of people who are not prepared to study ASCII characters, the inequality in language is strengthening the inequality in the information and knowledge sharing which widens the digital divide. I think that the word "inequality" is used here without any political sense, it is merely a fact that the technology development underlying the communication capabilities cannot accomplish in an action.

To create a user-friendly environment for the address space, the topic of multilingual domain names has attracted great efforts particularly in non-English-speaking countries and regions. There are generally two main approaches to enable the use of multilingual names for DNS and email addresses. One aims to extend the protocol to handle multilingual requests by introducing in packet identifiers, while the other concerns mainly the application or the client end to convert all multilingual characters into ASCII compatible strings before releasing them to the servers. In any case, the use of multilingual names affects not only the DNS or email servers themselves but also other applications and transportation that depend on them or carry these addresses within their databases.[2]

[2] Edmon Chung, David Leung, Jim Lam, Wilson Chow and Ken Lee. "Multilingual Domain Names & Email Addresses", Neteka Inc., March 2001

The currently available methods to realize multilingual access to Internet namespace include: Internationalized Domain Name, Keyword lookup, Keyword search, and Directory services. The APRICOT (Asia Pacific Regional Internet Conference on Operational Technologies) thoroughly addresses the issues related to those methods.[3]

The solution for IDN is based on the distribution of client software. IDN has been commercialized in China, Japan, Korea and others. Due to the joint efforts of engineers from many countries the global technical standard on IDN has been established.[4] On April 14, 2004, a proposed standard that was jointly drafted by CNNIC, JPRS, KRNIC and TWNIC was finally approved and published by IETF as RFC 3743 with the title of "Joint Engineering Team (JET) Guidelines for Internationalized Domain Names (IDN) Registration and Administration for Chinese, Japanese, and Korean". The basic approach of the RFC 3743 for Internationalized Domain Names, known as "IDNA", focuses on access to domain names in a range of scripts that is broader in scope than the original ASCII. The development process made it clear that the use of characters with similar appearances and/or interpretations created potential risk for confusion, as well as difficulties in deployment and transition. The solution is that, while those issues were important, they could best be addressed administratively rather than through restrictions embedded in the protocols. The RFC 3743 defines a set of guidelines for applying restrictions of that type for Chinese, Japanese and Korean (CJK) scripts and the zones that use them and, perhaps, provides a tentative framework for thinking about other zones, languages, and scripts.

Although the IDN service started ambitiously, yet the current market response has not been adequate as expected. One of the reasons might be the lack of a built-in IDN client software in the most popular browsers. However, the major browser companies which could contribute and assist to the deployment of IDN service would update their products to meet the IDN requests only when the expected market is big enough.

The migration towards a namespace common to multilingual access is a prospective task. The impact will be critical for widening the channel of access. It needs the awareness and collaboration among all administrators and managers around the world to handle multilingual requests.

Multilingual Access to the e-Content

An extraordinary spread of English language to the non-English-speaking areas can be observed simultaneously with the pace of Internet extension from the U.S. to the whole

[3] http://www.iak.ne.kr/nativename/2005/kyoto3.htm

[4] WGIG working paper, "Multilingualization of Internet Naming System", prepared by Cheon, Beji, Seadat, Sakamaki, Al-Darrab, Hassan, MacLean, Hu, Bertola, Shaban, Pisanty; February 2005

globe. It is natural trend as far as valuable resources have been dominantly created and stored in English web-pages. There was a time when 82% of the web-pages and 90% of emails were typed in English. According to a joint study by Funredes and Union Latine, even though English is still dominant, between 1998 and 2001 English on the web decreased from 75% to 52%. The volume emptied out is mainly filled with other European languages.[5] It can yet be regarded as an effective and easy solution if everyone speaks English. But for most people it is unlikely that they can master English as well as their mother tongue in understanding and expression of the subtle things. Besides, to protect the cultural diversity is also very important for mankind. UNESCO is concerned with the protection of cultural and linguistic diversity and the promotion of language diversity on the Internet, too. UNESCO has set the goals: "to achieve worldwide access to e-contents in all languages, improve the linguistic capabilities of users and create and develop tools for multilingual access to the Internet."

Hence, the ultimate request of multilingualism over the Internet would be the cross language access of the e-content. The speed of data flow and the scale of information volume on the Internet do not allow human interpreters to play decisive role to help people in cross language access of e-content. Only computer-aided translation may probably one day bring us closer to the target when people can read in his or her native language web pages that are published, say, in English. And the real-time or semi-real-time communication can take place between people using different languages. Research on computer-aided translation has made considerable progress over the past 50 years. There are systems in application, mostly used in Japan, Canada and Europe. The machine interpreters are usually limited to bilingual and need to be heavily specialized if they are to produce translations good enough to be understood by human editors.[6]

The investigation and research in the field of multilingualism for the Internet, including computer-aided translation, is really a big challenge to science and technology. If we all work hard, and collaborate, and be creative, a multilingual Internet, easy to access, share and utilize for people speaking different languages will eventually become a reality.

[5] Funredes and Union Latine, August 22, 2001

[6] Special Report: "Multilingualism on the Internet"; by Bruno Oudet; 2001.