

Issues in Building a Multilingual Internet By John Klensin

Summary

This article is adapted from a talk given at the ITU/UNESCO (International Telecommunication Union/United Nations Educational, Scientific and Cultural Organization) Global Symposium on Promoting the Multilingual Internet in May 2006.

Within the Internet community there is general agreement that in order to satisfy the vision of an Internet that is available to everyone everywhere, not only must information, resources, and content be made available in local languages and scripts, but it also must be possible to navigate the network in those languages and scripts. There is disagreement, however, over how to facilitate that expansion and growth, as well as over what the problems are that must be solved in order to achieve those goals. This article attempts to identify (1) the key issues that need to be addressed so as to develop a truly multilingual Internet and (2) suggestions for how those issues might be resolved.

Dr. John C. Klensin is an independent consultant following a distinguished career as the Internet Architecture Vice President at AT&T, Distinguished Engineering Fellow at MCI WorldCom, and Principal Research Scientist at MIT. He served on the Internet Architecture Board from 1996–2002 and was its Chair from 2000 until the end of his term. Earlier, he served as IETF Area Director for Applications and was Chair, Cochair, and Editor for IETF Working Groups focused on messaging and IETF process issues. Mr. Klensin was involved in the early procedural and definitional work for DNS administration and top-level domain definitions and was part of the committee that worked out the transition of DNS-related responsibilities between USC–ISI and what became ICANN. Prior to coming to MCI in mid-1994, he was INFOODS Project Coordinator for the United Nations University and, before that, was at MIT for nearly 30 years, holding Principal Research Scientist appointments in several departments including Architecture, the Center for International Studies, and the Laboratory of Architecture and Planning.

A nonprofit organisation, the Internet Society was founded in 1992 as a leader in promoting the evolution and growth of the Internet. Through our members, chapters, and partners, we are the hub of the largest international network of people and organisations that work with the Internet. We work on many levels to address the development, availability, and technology of the Internet.

The Internet is critical to advancing economic growth, community self-reliance, and social justice throughout the world. Become a member of the Internet Society, and share this vision. For more information, visit <http://www.isoc.org>.

1775 Wiehle Avenue, Suite 102
Reston, VA 20190-5108, U.S.A.
+1 703 439 2120

4, rue des Falaises
CH-1205 Geneva, Switzerland
+41 22 807 1444

The Complexities of a Multilingual Internet

While the challenges of developing a multilingual Internet are numerous, the difficulties are far more conceptual and practical than they are technical. By focusing primarily on the technical details—such as on how to change formats to make characters other than Roman-derived ones work or on the challenges of a Domain Name System that can accurately depict the Cyrillic alphabet or Japanese lettering—we miss the larger question of how best to prepare the Internet to accommodate other languages. In the best of all possible worlds, we begin first with an understanding of how language and culture function and interoperate, especially by non-English speaking populations that do not use scripts based on the Roman alphabet and by populations for which keyboards have not yet been adapted. Second, we must be reasonably certain that we are asking the right questions. Failure to do either may not only prevent progress but also actually hinder the development and worldwide deployment of the type of multilingual Internet to which we aspire.

If our goal is to make the Internet not merely accessible globally but also consistent with expectations that derive from normal use of languages within cultures, then the real issues are how to make content available in every language, how to increase levels of access for all populations, and how to generate presentation forms that are appropriate to each culture. We must also ensure that there are no significant disadvantages in terms of usability and convenience for particular cultures or particular languages relative to others.

What we have learned after years of Internet development, as well as after adoption of the Internet by millions of users, is that while technological problems may be inevitable, generally they are not insurmountable. It is not the technological challenges that will determine our success or failure, but our ability to structure the solutions around what we know about language and culture. In other words, we should first examine the fundamental cultural and linguistic issues that need to be addressed and understood. Then we can identify the technical and economic problems that must be solved in order to get us where we're going.

Balancing Local Needs with Global Demands: The Roles of Context and Content

From both cultural and technical perspectives, there are distinct differences between a collection of networks that is geared toward use on a local level and a multilingual Internet that is intended for use in a global environment. The first instance is a multilingual Internet that would accommodate any language or script, enabling all users to have access and gain benefit from the Internet regardless of the language they speak or the script they use. The second instance is one wherein individuals can communicate in one language or obtain resources from those who speak other languages. For thousands of years, populations existed solely on the local level. The use of language and the written word was rooted in familiarity with a society's unique conventions. It was only when populations and societies began to intermingle that the difficulties arose; and that's when the need to understand context became a critical element of effective communication and information exchange.

What about naming? After all, the Internet became popularised in large part because the Domain Name System permitted the use of names and mnemonic phrases rather than requiring users to remember cumbersome numeric Internet Protocol addresses and to track down the new addresses when resources are relocated. Contrary to popular belief, a truly multilingual Internet is not dependent on a technologically sophisticated naming scheme. Culturally and philosophically, naming is a means to an end, not a useful end result in itself. Similarly, the identification of objects—the distinguishing of one object from another and the creation of categories of objects—is important, but like naming, it is also rarely, if ever, the key issue. The key issue is context: both proper identification and adequate navigation usually require context about the user's interests, location, or culture.

If context is a necessary component of successful navigation and identification, then the systems we design need to establish and preserve context. Difficulties arise when we move beyond local terminology to descriptions, categories, and identifiers that must be understood globally or that can be translated accurately to and from normal modes of speech. Because of subtle differences in meanings or associations, translating-dictionary and multilingual-thesaurus projects are often extremely difficult to get right—and few succeed completely. When we start in-

teracting across cultures and among larger groups, we run into difficulties even about such seemingly simple issues as terminology and people's names. Therefore, to be effective, it may be necessary to minimise the number and scope of required global norms.

Equally important is the role of local content. The value of the Internet lies not simply in its provision of the ability to communicate quickly and efficiently, but also in its providing the ability to locate and access resources. Those resources, known generally as content, include information in the form of text, graphic images, audio, and video, to name a few. While language and culture influence and shape content, a truly multilingual Internet will make local content both accessible on the local level and available across languages and cultures.

If cultural preservation were the only priority, then all of the traditional advantages of isolation would apply. Local conventions that are not internationally compatible can prevent contamination to a culture, but only at the cost of other objectives. Communication across borders has a number of important benefits: it may aid in social and economic development, improve education, expand resources, and facilitate and increase international commerce. If our primary goal is a global network that maximises international communication, then some homogeneity is a good thing, but that, too, may present trade-offs. Therefore, we must attempt to achieve a balance between the key goal of expanding global interoperability and the equally important objective of strengthening cultures through the use of local languages.

The challenge, then, is neither how to build the best or most localized network possible nor how to build the best globally integrated network. The challenge is how to balance local cultural and linguistic preservation with the needs of an increasingly global economic and social environment.

If that is the case, then how do we make it possible for individuals to use the Internet to communicate easily with each other in their native languages and scripts, while making it equally effective and effortless for individuals immersed in one language and culture to access information and resources from nonnative sources? How can we layer a localized network or collection of networks on a global Internet and take advantage of that combination? These are difficult questions because they extend far beyond questions of whether or not we can develop the proper technologies. The solutions must be seen in terms of weighing one priority against another, analysing situations and consequences, and making what we hope are the correct trade-offs.

What's in a Name?

If we think carefully about names, we see that the ways they are presented to users are little more than abstractions. Global naming conventions always involve one or more of the compromises and trade-offs discussed earlier: we can optimise for culture and linguistic correctness, or we can optimise for simplicity. We can have both as long as we achieve balance. One of the ways of achieving that balance is to have local and user-appropriate names and references layered over standard global structures.

Images versus Character Coding

There are two ways to put an existing body of text onto a computer system: One involves an image of the printed form of the text—effectively a photograph of that text, usually of full pages, or the equivalent. While that approach can render typographic conventions and calligraphy more or less accurately, it is at best difficult and at worst impossible to automatically search or index the text. The other involves mapping the characters that make up the text into a system of codes—commonly referred to as a coded character set—and then placing those codes into a file that can be used to process and regenerate the text. Computerbased character-coding systems thus support the coding of the written word—in order to provide a display of the text—as well as indexing and other operations. In order to achieve text display that's relatively culturally accurate, it is usually necessary (but not sufficient) to have coding systems. In general, in order to specify formatting and other elements of correct presentation, the coding systems must be supplemented by markup (or the equivalent) or considerable information that is not stored and transmitted through the computer system or network. Good text display, with proper design, requires page description formats and imaging display, plus very specific markup. All of these either supplement or complement the coding systems, but rarely are they part of them.

User-appropriate names can be implemented in many ways. Systems that have been known in the Internet environment as “keywords” represent one such way; digital object identifiers of various types could be regarded as another, but only if they are designed to be localizable as well as easy to remember and use. The same may be true of the aids we employ in order to remember the content or results of a search. We are likely to discover still more examples as we move along. Perhaps surprisingly, what does not appear on the list of likely user-appropriate names is the domain name itself, nor is a domain name that is embedded in a Uniform Resource Locator (URL) appropriate. Even when names are internationalised, they are subject to the very exacting matching rules of the Domain Name System; and those rules are unsuitable for most scripts and orthographies.

While ample time is spent internationally discussing domain names and Uniform Resource Identifiers (URIs), we should not get too excited about names. Creating a multilingual Internet is not about fixing the Domain Name System. Domain names are good as mnemonics. When we attempt to examine questions about well-formed phrases—or even just words in any language at all—we run into problems with domain names.

What We Need (and Do Not Need) to Have a Truly Global Internet

In the most general sense, a truly global Internet means that all of us, regardless of our language or culture, have the ability to access any resource from anywhere. Having that capability does not imply that we always need it, nor does it imply that we have the ability to understand the material. But the fundamental structure for accessing any resource from anywhere is important if we are going to have (or preserve) a global network.

In doing so, we may apply certain restrictions, such as mechanisms for keeping children from inappropriate content. But the network design and implementation should not themselves become the restrictions. If the network design or the implementation becomes the restriction, we will soon discover that we no longer have a global network that can transmit information without losing some of it.

Over the years, as the Internet has grown in popularity and general use, the decisions and traditions that have influenced its development have brought both surprising benefits and unintended consequences. The benefits are obvious, but among the consequences are protocol terminologies that are visible to users, such as the File Transfer Protocol (FTP) commands, the format of URLs, and the contents of e-mail headers and envelopes. While it was convenient and useful at the time, what are, in fact, technical protocol terminologies are now seen by users as language-specific methods for accessing the facilities and information provided by those protocols. Regardless of the script, or the terminology, or the language we use, nothing can make a lengthy URL user-friendly. Trying to map or translate technical protocols between languages or scripts in an attempt to accommodate popular usage presents serious technical as well as localization issues.

The ability to find material of interest, to navigate, to search, and to consult directories for names requires that information be accessible in local languages. However, it does not mean that the information must be accessible in all languages everywhere, independent of locale. Travellers to areas outside their own vastly complicate the localization problem, but they do not fundamentally change it. Therefore, we need to make decisions about the right balance in this area as well.

Another benefit of the global Internet is that it has allowed us to increase and facilitate commerce internationally (not just domestically). However, in order to achieve a truly multilingual Internet, we need to be able to innovate without prior international agreement about each specific approach or technology. We need to avoid politicising the process through claims of sovereignty over languages, cultures, and scripts. We also should be able to move forward without the types of delays that result from the need to obtain global approvals or to construct conversion gateways. Conversion gateways that are designed to compensate for differences in protocols or representations as we move between countries or locations usually slow things down. Worse, they tend to lose information. Retention of a global Internet-like infrastructure—another one of the general properties of the Internet that have made it successful—avoids the need to require conversion gateways.

To satisfy these and other needs, we need a predictable global name space and a predictable global address space, with few or no intermediaries that introduce conversions, lookups, or remappings of their own. However, that name space and the supporting facilities may not have to be directly accessible to end users.

Understanding the Technological Challenges of a Multilingual Internet

From a technical perspective, there are important relationships and critical differences between a multilingual Internet that is primarily international and one that is primarily local. Internationalisation is both a set of tools and a principle; it is not an important end in and of itself. In fact, except as a set of tools and a principle, no one really cares about internationalisation—other than the technicians who are developing the appropriate technologies. What people want is what we call localization—in other words, systems that are adapted to the culture, language, environment, and preferences of the user. The tools we have today facilitate multicultural access, but they do not ensure multicultural access.

The interesting questions then become: How local does local get? Do we need to localize to a country? a village? a neighbourhood? a linguistic group? a cultural group? a tribe? The answer to any of those questions is “Sometimes yes.” They are important issues for us to sort out, but they are not technical issues.

The technical issues, particularly with regard to internationalisation, may be difficult, but they are also straightforward, particularly if the language and culture are understood and if the coding exists or can be created for writing and display of content. However, it is easy only if one does not care about simultaneously interacting with a global environment.

Within a specific language and culture, we do not need agreed-upon international conventions. Local environments do well with local terminology that is adapted to that environment. International conventions may help build on the experience of other cultures, but they are not necessary for serving local needs. To preserve and build on a global environment, we need to move beyond local terminology and to include local descriptions, categories, and identification that can be used globally, or translated accurately, to and from global norms. At the same time, it may be necessary to minimise the number and scope of required global norms. When we start interacting across cultures and among larger groups, we run into difficulties even about such seemingly simple issues as terminology or people’s names.

As described earlier, in order to have a multicultural and multilingual Internet environment, we must have content and a means for navigating to it easily. It should be obvious that it is not important to be able to navigate easily to things that do not exist. Navigation may be critical, but users are not typically interested in navigation. They want to find things, look at them, and deal with them. We need technologies that enable us to enter, transmit, and present text in every language and in every culture. Even though there have been some rough edges, the Internet elements needed have been in place for a decade. We have methods for identifying languages, scripts, character coding, and media formats and for transmitting that information across the network along with the data. But local elements are needed as well. Those elements may constitute an issue—with the issues differing depending on the language and on the scripts—but the local elements are without question the area in which most of our energy should be focused as we work to improve the ability to use a wide range of languages on the Internet.

Character coding is also crucial. Not having character coding in a single international standard—such as having a script that is not adequately represented in Unicode—creates difficulties for many reasons. However, we have survived for many years with script-specific character coding systems, and if necessary, we would continue to survive if we need such systems in the future. Keyboards, screens, printers, and other input and output devices are also important, but it is crucial to have operating systems and applications that do not get in the way.

Getting the Work Done

In closing, it is important to emphasise that we must begin the process of developing a truly multilingual Internet by understanding how language and culture operate and why context and content are key. If there is no content, then the rest of the discussion is irrelevant. Therefore, the primary concerns should be about (1) preserving local languages and culture, (2) developing navigation facilities that preserve context, and (3) the role of content. Using local content, in context, in addition to, rather than instead of, global naming and addressing will give us the freedom to develop a truly local Internet. It will also enable us to get things right culturally—without compromising in ways that interfere with the ability to communicate internationally when that is our requirement.

Finally, when we look at where particular parts of the work of internationalising the Internet should be done, we find ourselves facing an entirely different type of challenge. There are many complex problems ahead of us. Meetings are expensive. Efforts too often get duplicated. Time and resources too often get wasted. And no culturally appropriate product has ever been deployed as the result of a global meeting. In the work ahead of us, the most interesting and difficult problems will go to the very roots of language, culture, and cultural preservation. It may be comforting to note that those problems were with us long before the Internet and they will be with us long after the Internet. To be successful, we must first agree on the priorities and then work together to build a truly global Internet.

Lessons of the Past

There are lessons from the 1980s that have been forgotten but that are relevant to this discussion. One is that network protocols that are simple and that have very few options can be implemented. We learned that if there are options, transmitting information about them and their values as part of the protocol was critical to interoperability. We learned that protocols and features that are easy to implement and that interoperate well spread very quickly. Conversely, we learned that if a protocol has an extensive collection of options and requires that each system obtain information about the option choices made by the other through some external method, deployment will be slow and interoperation unlikely—at least without enormous effort.

We also learned that complex designs often do not interoperate well: They are hard to implement, and they require extensive coordination, a long time, and considerable investment to develop and deploy. Plus, it takes even more time to achieve smooth, multivendor interoperation, and people often run out of energy before that happens. We learned that if a new protocol or application has a predecessor in the field that is still being used, especially if it is simple and elegant, the newer and more complex system will rarely, if ever, displace it—no matter how many resources are invested in it or how much better it performs.

Finally, we learned that layering works. We can put localized references, local character sets, local strings, and local languages on top of global naming, addressing, and character-coding infrastructures, and things will work quite well as long as we do not let them become too complicated in the process.