Foresight of the Internet
Digital networks as structuring tools for the Knowledge Regions

On cultural diversity and multilningualism on the Internet

Borka Jerman-Blažič
ISOC-ECC, chair
Slovenia
borka@e5.ijs.si

Abstract
The paper discuss the problems of multilingualism and cultural diversity on the Internet. From 1996 the Internet community has provided technical solutions for exchanging communication of written text in all known languages on the world. However, the development and support to the multilingual services and cultural diversitytoday is still facing with several problems, starting from missing coding of less known languages through poverty of regions where these languages are spoken or with the problems of basic access to the Internet services. The way forward is discussed at the end of the paper.

1.Introduction: Internet and the multilingualism

Domination of verbal language in communication systems of the past has been an useful convention for the relationships in society where writing (strenghtened from print) was most used system. Internet has made written communication more global and accessible to everyone. However, it introduced as well the problem of multilingualism as English was dominating language for text communication. In the period of development of Internet technology when hardware and software were first designed to process English text the communication was possible with letters coded with seven-bit ASCII code that provides the repertoire of the English alphabet only. The notion of *multilingualism* goes from a simple *non-English* interpretation to quite complex conceptions of *multi-language* and *cross-language* aspects. Multilingualism or the even broader term *internationalisation* does not only cover linguistic issues. It also refers to a specific cultural behaviour of different societies which becomes visible through writing rules and in unique patterns of how to produce documents for each community (e.g. time, date, abbreviations). The missing understanding and knowledge of such differences caused by the social context may lead to major communication and understanding problems. Internet is by far not the first mass media confronted with this problem, but it is the most important media in our everyday life and as such deserve special consideration. Internet can make people to understand each other and hopefully to accept the differences in culture, heritage and the diversity of the current world languages.

After the commercialization of Internet in the beginning of nintees technical solutions required for written communication in the languages spoken all over the world were approached by the Internet community. The internationalization process of Internet services started with the e-mail standard known as MIME (1) which allows mail exchange of messages written in different languages and different scripts. However, MIME was just one Internet service and global solution was needed for the whole Internet infrastructure.

The first document that provides a global architecture for provision of the multi-lingualism in the Internet dates in 1996 when an IAB workshop was called and held on 29 February - 1 March at Information Sciences Institute (ISI) in Marina del Rey, California. This workshop had an objective to provide basic solutions to this items. The rational of the workshop was in the fact that many protocols throughout the Internet use text strings that are entered by, or are visible to, humans. As a consequence a need was recognized for anyone to be enabled to enter or read these text strings, which means that the users must be able to enter text in typical input methods and they should be displayed in any human language. Further, text containing any character from the world collection of alphabets of spoken languages should be able to be passed between Internet applications easily. That was the major challenge of internationalization considered at the IAB workshop (2). Solutions were designed and documented in RFC 2130 document. The RFC document defines the framework dedicated to the overall architecture of the Internet protocols and services required for accomodation of the world scripts used for writing the languages of the world. The framework is designed with four components: the architectural model, which specifies components necessary for on-the-wire transmission of text; recommendations for tagging of the transmitted (and stored) text; recommended defaults parameters for each level of the model; and a set of recommendations to the IAB, IANA, and the IESG for further integration of the framework into text transmission protocols. The architectural model specifies 7 layers, of which only three are required for on-the-wire transmission. The RFC 2130 report recommends the use of ISO 10646 (or its industrial version known as Unicode) as the default Coded Character Set, and UTF-8 method of coding characters as the default Character Encoding Scheme in the creation of new protocols or new version of old Internet protocols which transmit text. The specified defaults do not deprecate the use of other character sets when and where they are needed; they are simply intended to provide guidance and a specification for interoperability. This early RFC was followed with many others RFC documents and standards dealing with different applications, protocols and services offered over the Internet and as RFC2277 summarizes the main goal of these efforts was the Internet community to answer to the user requirement: "Internationalization is for humans« meaning that the protocols are not subject to internationalization; text strings are."

2. The world today and the multilingualism

The development and adoption of the international character-encoding standard in the Internet architecture made it possible to send and receive—and read—text electronically for hundreds of languages, all in their original scripts. The last fortress regarding the usage of ASCII in the Internet protocols - the internationalization of DNS is also taken today. In DNS addresses and names it is possible to use national character and as a consequence registrars from Europe e.g. Poland are starting registration of domains that contain national - Polish characters.

Communicating with people in their own language and script over the Internet is of great importance as it gives the world's diverse populations an electronic presence in the global information economy. However, if we look today at the world after ten years of adoption of the multilingual architectural model if the multilingualism on the Internet is sufficiently supported and present – the answer would certainly not be very positive. English is still the dominating language for correspondance and communication over the Internet. Several sources are confirming that, despite the improvements presented by the latest data from year 2003 (see fig.1, source: Global reach, 2000, Fig.3 and Fig.4). The source for this domination is in the stable population of the *first-language English speakers* which is around 350 million, then in the fact that *English as a second and foreign language* in the world has grown dramatically since 1950, (according to some authors in 1992 this number was close to 750 million), and that 1-2 billion people have *some ability in English* (world population is being close to 4 billion). English is also becoming the global *lingua franca* of aviation, business, diplomacy, higher education, mathematics, science, technology, etc. English Web pages on the Internet are dominating with 68% percentage of the browser languge settings (Fig.2.).

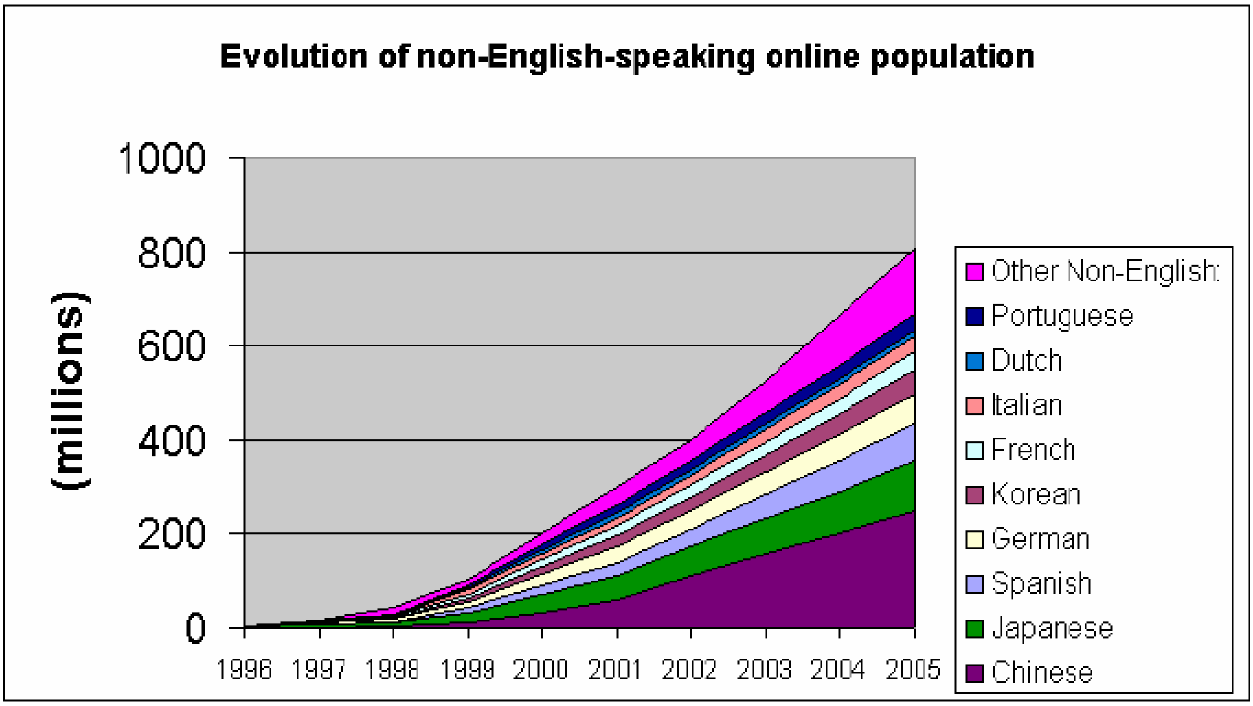
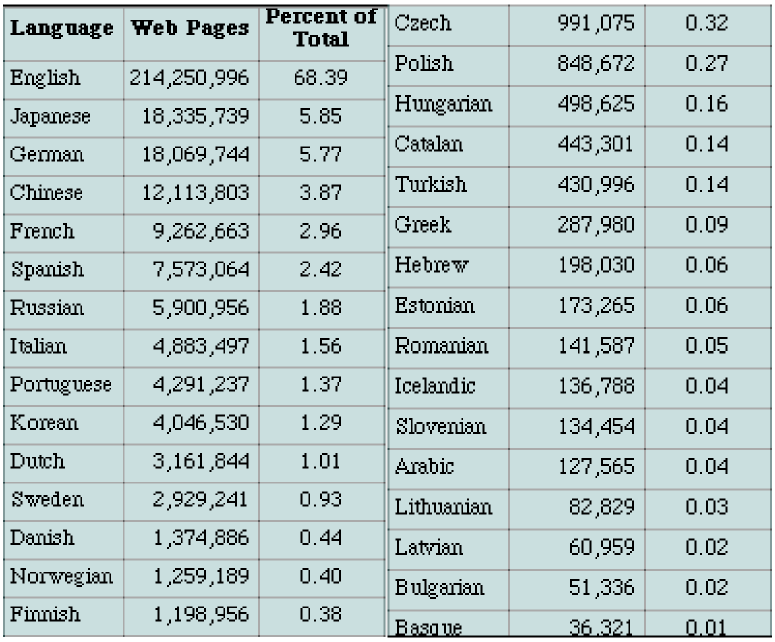


Fig.1. Evolution of non-English speaking on-line population in the last ten years

The difference is obvious when the number of WEB pages in particular language are taken as a measure per native speaker of that language. There are only 1.5 people per a WEB page in English, 1.8 people in Web page in Icelandic and 175 Bolgarian or 185 Romanian people per WEB page presented in their language. The situation in the Arabic world is ten times worst as there are 1 830 arabic people per WEB page in Arabic. We do not have measurment for Africa, Tibet, Mongolia and similar undeveloped regions. There, the problems are still connected with proper recognition of their languages and the coding of their alphabets. 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 of the encoded scripts were selected for inclusion because they are used in languages spoken by more than five million people. Still, there are more than 80 scripts remaining outside the Unicode standard, locking out their users from the capabilities of the Internet Approximately one-third of them are in active use today, most by groups in Asia and Africa (3). The rest is historic, including Egyptian hieroglyphics and ancient scripts of the Middle East. While the popular media has focused on the effort to save biological diversity and endangered languages [4], the case for preserving the writing systems of languages is still largely unnoticed by them. It is known fact today that half of the known languages in the world have vanished in the last 500 years (4), by 2100, 3,000 of the remaining 6,000 languages are expected to perish and 2,400 will become near-extinct (4). The last are mostly small, indigenous languages (vs. national or international languages) that are being lost. Saving scripts by including them in Unicode will help to document the variety of writing systems in the world digital libraries accessible over Internet 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.



Language	Web Pages	Percent of Total
English	214,250,996	68.39
Japanese	18,335,739	5.85
German	18,069,744	5.77
Chinese	12,113,803	3.87
French	9,262,663	2.96
Spanish	7,573,064	2.42
Russian	5,900,956	1.88
Italian	4,883,497	1.56
Portuguese	4,291,237	1.37
Korean	4,046,530	1.29
Dutch	3,161,844	1.01
Sweden	2,929,241	0.93
Danish	1,374,886	0.44
Norwegian	1,259,189	0.40
Finnish	1,198,956	0.38
Czech	991,075	0.32
Polish	848,672	0.27
Hungarian	498,625	0.16
Catalan	443,301	0.14
Turkish	430,996	0.14
Greek	287,980	0.09
Hebrew	198,030	0.06
Estonian	173,265	0.06
Romanian	141,587	0.05
Icelandic	136,788	0.04
Slovenian	134,454	0.04
Arabic	127,565	0.04
Lithuanian	82,829	0.03
Latvian	60,959	0.02
Bulgarian	51,336	0.02
Basque	36,321	0.01

Fig.2. Percentage of the WEB pages in particular language

In Bali, Indonesia, the Balinese script, which is used in many cultural and literary works, is taught in the schools. Students' fluency is poor and is 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.

Typical language recognition problems are not present in the developing world only. There are still unsolved problems in the most world developed areas such as Europe. Euroactive (www.euroactive.org) has published recently the request of the Irish people for better status of the Gaelic language in the EU. Today Gaelic is not considered as an official European language! In joining the EU in 1973, Ireland chose English as its working language. The Irish government today says 41 percent of the country's four million people speak Gaelic. However, the Internet has given also "voice" to the minority languages as it functions as a vehicle of political empowerment (e.g., Basque, Catalan etc.). This is important from the Internet point of view as it is a media characterized with democratic access, low publication cost, seemingly limitless space enabling room for all, regardless of the viability of language. In addition to that, machine translation on the Internet insure also mutual intelligibility, however minority languages are the last to come online due to cost, lack of literacy etc. Machine translation is also difficult; it is available for major world languages only.

Being able to write and read texts in the original language scripts has important practical ramifications in many aspects. 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 and poor 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. The recent tzunami crisis in south Asia is typical example where proper alert system communicated on the Internet networks could save hundred of thousand people. The Internet community started relief action and in addition to that a standard was proposed for an alert system over the Internet. However, the problems are still in the understanding of the rich and the poor what is the most important.

In that context, I would like to refrence to the recent e-mail communicated to the ISOC delegate mailing list from the delegate of the Indonesian ISOC chapter Mr.Irwan Effendi who reacted to the proposed alert system standard and the relief actions regarding the tsunami crisis. He wrote:

»As for direction of ISOC, we in Indonesia as developing country have no objections whatever in the standards as a development activities. However, we do believe that ISOC should be moving to educate the people with the basics. However, we do believe that ISOC should be moving to educate the people with the basics, so that more people will be aware of the standards, and more people can get involved in the future. To add to your information statistic, in Indonesia less than 1 percent of internet user know about ISOC (or even hear the name, as the matter of fact). Those who know however, has accuse ISOC (and USA) for practicing a new kind of empire, technological subjugation under the pretense of standards. Though this accussation is clearly unjustified, they all point out at a single fact, that so far none of the standards being practised is developed by the members from developing or underdeveloped countries. For that, Indonesian Chapter is currently working very hard to socialize about ISOC, the standards and such, but sadly enough, we are met with the fact that NONE of the branch or representative office of the major vendors (Microsoft, IBM, HP, Oracle, Computer Associates, etc.) in Indonesia is willing to support or participate in this effort. It is as though they confirmed the accussation that people in countries such as Indonesia is being forced to remain in the user and programmer level, without needing to understand the technology a step further, even less to have anything to say about what kind of standard should be developed. The good news is: community has a strength of its own and now we are gathering the support and participation we need from local universities, user groups and small businesses. However, I do feel the need to warn ISOC that without direct involvement from the international community, Indonesian technology will most likely be developed in a waythat is similar to China, connected yet secluded from the rest, which mean another failure for the ideals of open standards. If you have any good ideas on how to resolve this issue, please do let us know«.

3. The way forward

The problem of the multilingualism and the cultural diversity is in its essence is a matter of economy and the level of development. Language is a carrier of culture, but it is rarely the driving force behind cultural domination: that is rather political, economic, religious and/or social. With absence of the dominating force, an imposed language becomes a potential resource for the advancement of its speakers, the history points to several examples e.g., French in post-Norman England or English in post-colonial India. English will certainly "dominate" on the Internet in the future – and be a vehicle for US cultural dominance but only as long as the Internet is associated with the US. This may be is already changing, as the Internet is adopted by other cultures. The changes in the language use are obvious if we compare the situation in 2001 and in 2003 among the Internet users (Fig 3. and Fig 4, source UNESCO conference).

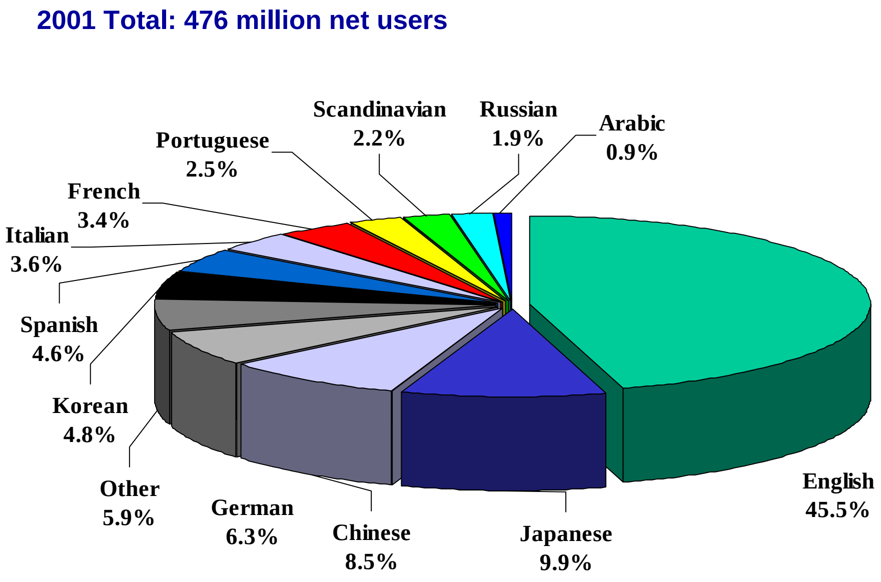


Fig.3. Languages on the Internet 2001

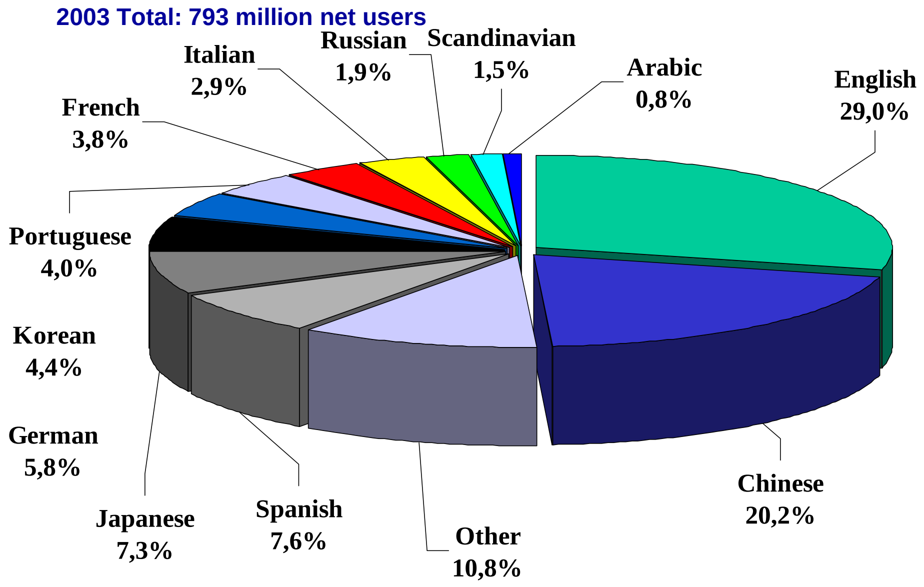


Fig.4.Languages on the Internet in 2003

If we are optimistic we may say that the Internet will promote and will reflect the linguistic diversity, and be a potential source of empowerment of minority language groups. Somewhere they are getting help due to their relevance for scientific research. In US, the the Script Encoding Initiative has a goal to populate the Unicode standard with missing scripts. The Initiative is run by the Californian scholars. They are trying to raise awareness and secure funding by explaining the value and scientific basis of character encoding for some »exhotic« languages. The importance of character encoding—the Internet's backbone for text communication is still poorly understood by foundations and grant agencies responsible for research funding. World's languages and the literary study, historical, and cultural documents of minority languages are often found mainly at major american research institutions where the relevant and important cultural heritage material is being collected. In the other part of the world where some of the languages without place for their scripts in the Unicode are still spoken and the population is struggling with basic poverty and getting Internet access. The development planners are convinced that the language use is an essential component of development and human rights meaning that without the ability of concerned people to communicate what their real needs are, they face with solutions imposed on them and they are denied access to information that can assist them to take their own decisions

On the international level there are several initiatives dealing with cultural diversity, human rights and the multilingualism over the Internet. The international organisation that is approaching systematically the problem of the multilingualism and universal access to cyber space in the same time is certainly UNESCO. In year 2001 a Declaration on Cultural Diversity was adopted and the General UNESCO Conference in year 2002 reiterated "its conviction that this organization should play a leading international role in promoting access to information in the public domain, especially by encouraging multilingualism and cultural diversity on global information networks". Director-General was invited to submit "a draft recommendation on the promotion and use of multilingualism and universal access to cyberspace" (30 C/Resolution 37). Later on, in 2004 the Resolution and the Reccomendation were adopted and published. UNESCO recognize the need for capacity-building, particularly for developing countries, in acquisition and application of the new technologies for the information-poor. UNESCO aknowledges that basic education and literacy are prerequisites for universal access to cyberspace. In other this to be achieved development of multilingual content and systems is required and this should be provided jointly by the private and public sector. In that context, UNESCO recommends to the national, regional and international levels to work together to provide the necessary resources and take the necessary measures to alleviate language barriers and promote human interaction on the Internet by encouraging the creation and processing of, and access to, educational, cultural and scientific content in digital form, so as to ensure that all cultures can express themselves and have access to cyberspace in all languages, including indigenous ones. Member States of UNESCO and international organizations are expected to encourage and support capacity building for the production of local and indigenous content on the Internet.

UNESCO is urging as well Member States to formulate appropriate national policies on the crucial issue of language survival in cyberspace, designed to promote the teaching of languages, including mother tongues, in cyberspace. UNESCO recommends to its members international support and assistance to developing countries to be strengthened and extended to facilitate the development of freely accessible materials on language education in electronic form and to the enhancement of human capital skills in this area. Member States, international organizations and information and communication technology industries are expected also to encourage collaborative participatory research and development on, and local adaptation of, operating systems, search engines and web browsers with extensive multilingual capabilities, online dictionaries and terminologies. They should support international cooperative efforts with regard to automated translation services accessible to all, as well as intelligent linguistic systems such as those performing multilingual information retrieval, summarizing/abstracting and speech understanding, while fully respecting the right of translation of authors. UNESCO, in cooperation with other international organizations, recommends to establish a collaborative online observatory on existing policies, regulations, technical recommendations, and best practices relating to multilingualism and multilingual resources and applications, including innovations in language computerization. Who else will join this initiative?

5.Conclusion

In the course of UNESCO Recommendation preparation 42 international organization among them ISOC, the European Council and the Commission of the Europena Union were contacted. This is a good start as political consenus regarding what is important for the cyberspace is in place. However, this is just a start. Much more has to be done for the support of the universal access to the Internet as an instrument for promoting the realization of the human rights and enabling expression of . The access to the Internet as a service of public interest should be promoted through adoption of appropriate policies that are enhancing the process of empowering citizenship and civil society, and are encouraging proper implementation of such policies in developing countries, with consideration of the needs of the rural communities. Most important is still the adoption of policies and mechanisms all over the world that will facilitate the universal access to the Internet through affordable telecommunications and Internet costs with special consideration given to the needs of public service and educational institutions, and to all others that are somehow disadvantaged by the poverty or are disabled. If this is in place then multilingualism and cultural diversity will easily flourish.

References

1.Y.Demchenko, Testing multilingual support in Mail User Agent, TERENA report, 1998
2.Harald Alvestrand et oth. RFC 2130, The Report on IAB Workshop
3.Deborah Anderson, COMMUNICATIONS OF THE ACM January 2005/Vol. 48, No. 1 27
4. Knight, W. Half of all languages face extinction. *New Scientist* (Feb.
16, 2004);
5.UNESCO Draft Recommendation Concerning the promotion and use of multilingualism and universal access to cyberspace
32 C/27, 2003, UNESCO Declaration on Cultural diversity, Paris 2.11.2001
6. CEN TC 304, PT01 Report, User requirements of internationalization and standardization in the field of character set technology, CEN Report, 1995