Target 9: Encourage the development of content and put in place technical conditions in order to facilitate the presence and use of all world languages on the Internet[1]

Introduction

Internet access is the physical precondition for being able to use the Internet and enjoy its benefits. With the rapid growth in the number of Internet users, reaching an estimated 26 per cent of the world's population by the end of 2009 [ITU, 2010], the amount of content[2] on the Internet has also been increasing, boosted by broadband Internet applications and the emergence of Web 2.0[3] and social networks. There are no agreed figures for the number of webpages,[4] but the best testimony to the incredible size of the web is the fact that search engines have now ceased indexing a significant portion of it.

Internet access is only part of the story, however. The true essence of the Internet is that it fosters communication between humans (and with networked objects), and allows them to obtain and exchange meaningful information. This raises the question of the relationship between content and users. It is imperative that all users worldwide should be able to find Internet content which is meaningful to them. Hence, this content has to be available in their mother tongue, which in turn means that content in local languages is required.

This is the issue of linguistic diversity in cyberspace addressed by WSIS Target 9. The aim is to achieve the greatest possible diversity in order to serve the online needs of people everywhere. This is a relevant and indeed pressing issue for establishing a global information society, since it is directly linked to the basic principle of equality of rights in society, and indirectly to the ecology of cyberspace.

UNESCO, which conducted the process that culminated in the *Convention on the Protection and Promotion of the Diversity of Cultural Expressions* [UNESCO, 2005], is coordinating WSIS Action Line C8, which identifies many priorities to be considered for achieving the target, including the following:[5]

- Develop and implement policies that preserve, affirm, respect and promote diversity of cultural expression and indigenous knowledge and traditions through the creation of varied information content.

- Support local content development, translation and adaptation, digital archives, and diverse forms of digital and traditional media by local authorities.

- Provide content that is relevant to the cultures and languages of individuals in the information society.

- Nurture the local capacity for the creation and distribution of software in local languages, as well as content that is relevant to different segments of population.

- Cooperate with indigenous peoples and traditional communities to enable them to more effectively use and benefit from the use of their traditional knowledge in the information society.

- Promote technologies and research in such areas as translation, iconographies, voice-assisted services and the development of necessary hardware and a variety of software models, including proprietary, open-source software and free software, such as standard character sets, language codes, electronic dictionaries, terminology and thesauri, multilingual search engines, machine-translation tools, internationalized domain names, content referencing as well as general and application software.

Some of the recommendations in Action Line C8 are directly aimed at indigenous populations, who are generally not well represented on the Internet and deserve special attention. Most of the action line refers directly to ways of fostering or encouraging the development of content (or the translation of existing content), and some provisions also address the crucial issue of software. Other WSIS targets, such as Target 4 (Connect public libraries, cultural centres, museums, post offices and archives with ICTs) also have a direct link to Target 9 in promoting linguistic diversity and local content online.

Target 9 and the corresponding action lines are set in an extremely fast-changing context, with rapid growth of Internet users and content. As a result of the concept of users evolving to include objects,[6] the current Internet Protocol addressing scheme has been stretched to its limits, and IPv6[7] has been implemented (besides other protocol improvements) in order to change the order of magnitude of the possible number of different networked entities,[8] which will allow this exponential growth to continue for many years to come.

The persistence of the so-called “digital divide” [ITU, 2010] is also, at least partly, related to the issue of languages and content on the Internet, in particular with regard to online equality [Pimienta, 2009). Equality is often interpreted in terms of physical access, focusing on the gap between those who have access to the technology and those who have not. Policies have been set accordingly to increase access, for example by putting in place shared resources (e.g. telecentre projects and community access points — see also Targets 1 and 4), more affordable technologies (e.g. low-cost PCs) or more pervasive technologies (e.g. cell phones).

However, in order to reduce the digital divide, barriers beyond mere access also need to be addressed to give all users the potential to benefit from information and communication. Users need the skills to use the technology and exploit the benefits it offers.[9] The issue of content availability in the user’s language is fundamental: without content, access itself becomes useless. There is no simple technological or practical answer for content production in cyberspace, an activity which mirrors the complexity inherent in the economic behaviour of individuals.[10] While ICT4D[11] specialists tend to agree that local content production[12] is desirable, the question of how to provide incentives to achieve it in practice remains unresolved, for now.

One reason for the relatively low policy priority given to this issue is the lack of indicators on languages on the Internet. To date, the digital divide has often been measured in terms of access rather than use and content, even though the divides for the latter two are likely to be greater than the access divide.

For example, while Africa accounted for two per cent of total Internet users in 2007 (and 11 per cent of the world’s population), its share of content production, in both national and local languages, was even lower. Furthermore, only 0.6 per cent of total worldwide content in English or French was in African servers (as estimated by *FUNREDES/Union Latine*).[13] The *Language Observatory Project* (LOP)[14] estimated that the percentage of webpages accounted for by local African languages (or at least the small subset of them which are localized)[15] ranged from 0.06 per cent to 0.006

per cent of total Internet content. Even though in some cases these rates may have increased somewhat since then, they are likely to have remained very low. In the case of local languages, the proportion may even be smaller, given the speed of growth of Internet content in the main languages present on the Internet.

Some people consider English as a *lingua franca*[16] in the digital world, thus minimizing the importance of the issue of linguistic diversity and the availability of local content. However, while English is a major language for business or science and the official language of many countries in Asia and Africa, it has been estimated that less than 15 per cent of the world's population understand it.[17] Furthermore, in many countries where English is the official language, a large proportion of the population does not actually use it (the same holds true for the use of French in many francophone African countries or in Haiti, for example). If meaningful access to the digital world is considered a prerequisite for becoming an active citizen of the information society, then the issues of access and content in local languages or mother tongue are central to democracy in the information society.

Evidence from UNESCO studies[18] suggests that not being educated in one's mother tongue is a serious handicap. If the Internet, which has become an important indirect source of information and knowledge and a practical component in education systems, is meant to be for everyone, then it must speak everyone's language, again highlighting the importance of this target.

The top barriers to ICT uptake, or reasons cited for not having the Internet at home, are a lack of perceived need; cost (of equipment and/or service); and a lack of skills [European Commission, 2009]. The results from a recent survey by the US Federal Communications Commission (FCC) also highlight that, in order to further increase connectivity, it is necessary to teach people who are not yet connected how to navigate the web and find online information that is valuable to them, so that they learn to appreciate the benefits offered by connectivity [Horrigan, 2010] (see Box 9.1).

Given that a lot of content on the Internet, and "Internet-related vocabulary," is in English (or at least this might be the perception),[19] countries with higher levels of English language scores could be expected to have higher Internet adoption. There is indeed a positive correlation between the percentage of households with the Internet and the number of Internet users, on the one hand, and proficiency in English, on the other, as proxied by TOEFL[20] scores.[21] This shows that being able to understand and use English terms could be a factor in Internet uptake and highlights the importance of increasing the number of languages on the Internet.

Target 9 deals with two separate but related issues: first, the ability to use a given language on the Internet, and second, the provision of appropriate content. The first can be addressed with technical solutions, whereas the second requires a comprehensive set of conditions, many of which are not purely technical. It is one thing to ensure that the technical conditions for a language to be used on the Internet (localization) are in place,[22] but another for this language to gain all the attributes of major languages on the Internet (full representation). See Box 9.2 for more details.

Box 9.1: Understanding digital concepts

In addition to a lack of information about the benefits that can be obtained from using the Internet and a lack of digital literacy skills, the level of English language skills and the availability of local content could also be related factors in Internet usage levels, especially as English is sometimes considered as the *lingua franca* of the Internet. Using the recent US FCC survey results, [Horrigan, 2010] finds that even in the United States where English is the first language, broadband users exhibit varying degrees of understanding of digital concepts, which, in turn, influences what they do online. For example, survey respondents received a series of questions asking them how well they understood various terms related to computers and the Internet. The following shares of broadband users said they understood the listed terms very well: *refresh* or *reload* — 61 per cent, *operating system* — 44 per cent; *Internet browser cookie* — 42 per cent; *JPEG file* — 41 per cent; *spyware* or *malware* — 40 per cent; and widget — 16 per cent. The study found that those with greater understanding of these terms were more intensive users of the Internet. Some 29 per cent of broadband users said they did not understand any of the listed terms very well, while 24 per cent understood five or six of the terms very well. The former group was found to be doing only about half of the online activities that the better informed group did. These findings suggest that skills, including language skills, play a role in how the Internet is being used.

Box 9.2: Requirements for the presence of languages on the Internet

There are many requirements for full representation of a language in cyberspace. A codification scheme for the alphabet and an appropriate keyboard layout are necessary but not sufficient conditions. Indeed, there is a set of conditions which, combined with an appropriate policy framework, will eventually permit the existence of content commensurate in just proportion with a language's population of speakers. These are [Diki Kidiri, 2007]:

- a written form for the language (languages which are only oral do not receive the full range of benefits of digital processing);[23]
- a comprehensive codification for the alphabet and fonts (for reading and writing of documents);
- basic linguistic software for the language (word processor, e-mail management, messaging, browsing);
- a supported keyboard with the alphabet of the language;
- advanced linguistic software (for spellchecking, syntax checking, alternative choices for sequence of letters, online dictionaries, etc.);
- accessibility resources in the language for users with disabilities (such as software for the blind);
- human resources trained to perform the (above) activities of creation and implementation;
- an informed and motivated user community driving content production (this in turn drives indirect requirements for education packages in the language for digital literacy, information literacy and content production);
- comprehensive content responding to the needs of the user community;
- content duly indexed by existing search engines (or alternative search engines specialized in the language);
- articulation with other languages (in particular, translation software from/to a set of other languages);
- sufficient funding from governments or international organizations to support the process leading to the fulfilment of all these conditions.

The task of achieving this target is complex and can only be fully accomplished with the involvement of all stakeholders. Indeed, a sufficient number of users have to be more than just content consumers, and also take responsibility for content production, generating culturally meaningful content and useful services (for example for online commerce and tourism). User involvement would ideally occur as a bottom-up process, but for many languages the critical mass is not present and other key elements may be lacking, such as the existence of qualified leaders who pull the process and/or the existence of a government sensitive to the importance of the issue and willing to put in place the appropriate policies and incentives, in consultation with all relevant stakeholders.

Measuring Target 9 — Proposed indicators

Internationally comparable and recent data on languages and content on the Internet are not widely available. Indicators on access to ICTs are more readily available than indicators on the use of ICTs, ICT skills or digital and information literacy.[24] This dearth of indicators may also have contributed to a relative lack of policy attention given to supporting linguistic diversity on the Internet. Measurement of languages on the Internet is also complicated by technical difficulties, which are likely to worsen over time as the Internet continues to expand in size.

To measure this target, initial focus could be placed on collecting two basic indicators:

1. Proportion of Internet users by language

2. Proportion of webpages by language (Box 9.3)

These indicators could also be collected for different Internet "subspaces" such as the blogosphere, social networks and newsgroups. Ideally, a breakdown by user characteristics (country, region, gender, age, family of languages) would also be available, and compiled at regular intervals so as to follow the rapidly evolving trends in cyberspace.

Box 9.3: Measuring the number of webpages by language

The basic method for measuring the percentage of webpages in a given language involves crawling the space to be measured and applying identification and counting techniques. Basically, if the language has a specific script for encoding, the pages are counted from that parameter. If the same script is shared by many languages (like for instance the Latin script), a language-recognition algorithm is required to identify the different languages sharing the same script.

The major inconvenience of this method is the time required to crawl. Today, this factor even makes the method prohibitive as the size of the space to be crawled expands indefinitely. Alternative techniques are required to apply the crawling method in a smaller space while remaining statistically representative. Some methods use a small randomly obtained sample of the web, but the result is statistically questionable.

LOP focuses on minority languages in top-level domains only (even removing countries with too large a webspace like China). Results, even if they are not fully representative, have shown a very low penetration of local languages on the web for Asian and African languages. It has been demonstrated [Nandasara *et al*, 2008] [Suzuki *et al*, 2002], that the local language percentages follow "Zipf's Law" — the n-th ranked language speaker is one n-th of the population of the top-ranked language. This suggests that the number of webpages written in each language follows a progressive power-law curve where the order of magnitude of presence of local languages continues to decrease; the authors refer to this as the "digital language divide."

Finally, other methods have applied the power of search engines to a set of words which are carefully selected to hold the best semantic and syntactic equivalence across languages. This is the approach that has been followed by FUNREDES/Union Latine, which was able to offer consistent measurements, with more complex indicators, between 1998 and 2007, albeit only for a subset of languages. The continuing expansion of the size of the web is showing the limits of all of these methods, however, and further discussion is required to identify ways to produce reliable figures.

For more details see [Paolillo, *et al*. 2005] and [Pimienta, 2008].

However, a number of additional indicators would be required as well in order to measure linguistic diversity in cyberspace fully. These indicators should, ideally, reflect the behaviour of end users with regard to languages in cyberspace. This includes considerations such as: users' mother tongue and second languages spoken, language set for software interfaces, percentage of e-mails read or written by language, percentage of visited webpages by language. Indirect indicators could also be complied, for example, by looking at the online availability of e-government information and applications in different languages, and to what extent related applications are available for minorities or immigrants in their mother tongue.

Status of Target 9

Unfortunately, the field of linguistic diversity indicators is still in its infancy and data are scant. Before looking at the languages that can be found on the Internet, it is useful to get an overall picture of languages in today's world. The most recent indicative figures on living languages in the world are given in Table 9.1, and the regional distribution in Chart 9.1. A country-level comparison of the number of Internet users and the number of living languages suggests that there is a correlation between Internet penetration rates and language diversity, many countries (for example, in Africa) with low Internet penetration having high language diversity, and vice versa.

Overall, the following facts can be reported about the state of languages in the world as at 2009: [25]

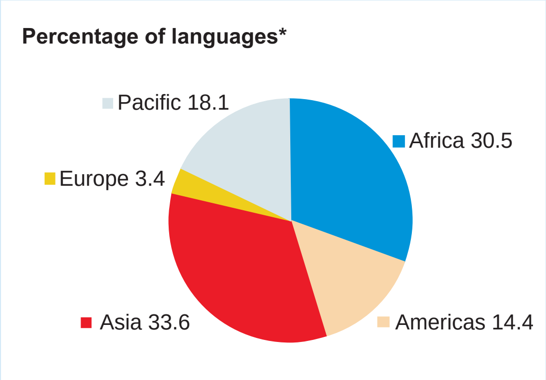
- 1 per cent of languages are spoken by at least 10 million people
- 1 per cent of languages are spoken by 94 per cent of world population
- 96 per cent of languages are spoken by 4 per cent of world population
- More than 50 per cent of languages are spoken by less than 10 000 people

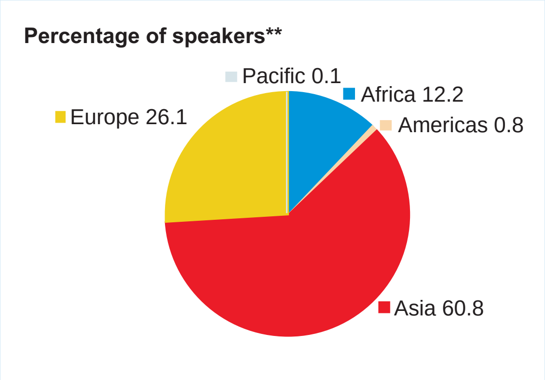
Table 9.1: Distribution of languages by number of first-language speakers, 2009

Population range	Living languages		Speakers	
	Number	%	Number	%
100 million to 1 billion	8	0.1	2 308 548 848	38.7
10 million to 100 million	77	1.1	2 346 900 757	39.4
1 million to 10 million	304	4.4	951 916 458	16.0
100 000 to 1 million	895	13.0	283 116 716	4.8
10 000 to 100 000	1 824	26.4	60 780 797	1.0
1 000 to 10 000	2 014	29.2	7 773 810	0.1
100 to 1 000	1 038	15.0	461 250	0.01
10 to 100	339	4.9	12 560	0.0
1 to 10	133	1.9	521	0.0
Unknown	277	4.0		
TOTAL	**6 909**	**100**	**5 959 511 717**	**100**

Source: [Lewis, 2009].

Chart 9.1: Regional distribution of languages, 2009





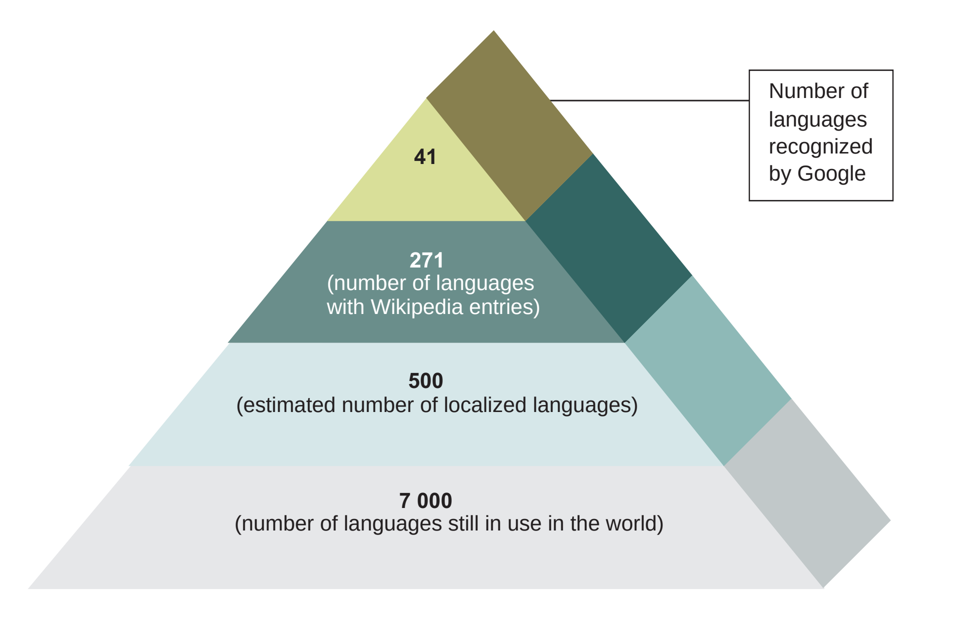
Note: * Living languages that originate in the specified area. ** People who use languages that originate in the specified area as their first language regardless of where in the world they may live.
Source: [Lewis, 2009].

- 70 per cent of languages are spoken in only 20 (developing) countries
- Less than one third of languages have a written form.[26]

There is no straightforward way of mapping these languages onto their online existence as there are very few data and indicators on the state of languages online.

Indeed, indicators of linguistic diversity in cyberspace are very limited. A visual representation of the relative number of languages represented in various ways in the virtual world, compared to the number of living languages in the real world, is given in Figure 9.1.

Figure 9.1: Key figures for languages on the Internet, 2010


Number of languages recognized by Google
41
271
(number of languages with Wikipedia entries)
500
(estimated number of localized languages)
7 000
(number of languages still in use in the world)

Note: There is no agreement among demo-linguists[27] on the size of the language universe, but the figures usually range between 6 000 and 9 000. Differences are explained by the difficulty of defining the distinction between a dialect and a language. The number of localized languages is an estimate.

Source: Ethnologue, SIL International (Summer Institute of Linguistics), Wikipedia and Google.

Technically, measuring the online presence of languages is complicated. Unicode[28] does not encode languages but scripts, a script being a collection of characters, some of which could be shared by different languages. In many cases, a single script may be used to write tens or even hundreds of languages (e.g. the Latin script). In other cases, only one language employs a particular script (e.g. Hangul, which is used only for the Korean language).[29] The approximate figure of 500 localized languages represents, therefore, the set of languages whose character sets are covered by the set of 140 scripts which have been encoded so far, 90 of them related to written languages.

One of the striking developments since the conclusion of WSIS is the growth of user-driven content and the emergence of social networking sites. Although it is not possible to ascertain from available data whether the sites were accessed in local languages, the Asia and the Pacific region registered the largest number of unique visitors between June 2007 and June 2008 (Table 9.2). The Middle East and Africa showed the strongest growth in the numbers of unique visitors (66 per cent), followed by Europe (35 per cent) and Latin America (33 per cent), though the Middle East and Africa and Latin America are growing from much lower numbers. According to Facebook's Timeline [Facebook, 2010a], the number of active users grew from one million people in December 2004 to 400 million in February 2010. Of the active users, 100 million access Facebook through mobile devices [Facebook, 2010b]. Facebook is growing strongly in the Asia and Pacific region and at the start of 2010, for example, there were 17 million users in Indonesia with a monthly growth rate of 13 per cent. Other high growth rates were observed for India and Thailand (both 12 per cent) and Malaysia (10 per cent) [Facebook, 2010c].

Given the significance of user-driven content, it is also important to look at the linguistic representation of social networking sites, for example. Thus, as of early 2010, 52 languages were available for translation in Google translate, and the Internet browsers Internet Explorer and Mozilla supported 63 and 70 languages, respectively. Furthermore, many “international” websites, such as Amazon, eBay, Google and Facebook, now propose localized versions, often also in local languages. Thus, Facebook was available in 67 languages, Blogger in 50, Youtube in 19, Flickr in 8, Twitter in 6, and Linkedin in 4 languages.

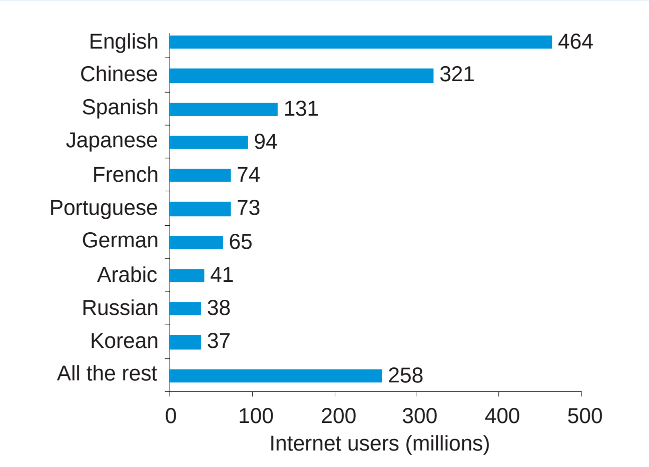
Table 9.2: Growth of social networking sites, by region, June 2007-June 2008,
Total worldwide audience, age 15+, home and work locations

	Unique visitors (thousands)		
	June 2007	June 2008	Percentage change
Worldwide	464 437	580 510	25%
Asia and the Pacific	162 738	200 555	23%
Europe	122 527	165 256	35%
North America	120 848	131 255	9%
Latin America	40 098	53 248	33%
Middle East and Africa	18 226	30 197	66%

Source: comScore World Metrix [ComScore, 2008).

Microsoft also plans to translate Windows 7 and Office 2010 into a further 59 local languages by June 2011, in addition to the 101 languages (including Azeri, Georgian, Macedonian, Uzbek, Bosnian, Punjabi and Kyrgyz) into which its most popular software packages have already been translated.[30]

Chart 9.2: Top ten languages on the Internet, 2009



Source: Internet World Stats, www.internetworldstats.com/stats7.html.

As the field of linguistic diversity indicators is still in its infancy, the only indicator which has been available thus far (and without any reliable degree of precision) is the distribution of Internet users by language. This was offered, with a transparent methodology, by Globalstat, until 2005,[31] and subsequently by Internetworldstats, although limited only to the figures for the top ten languages in terms of users (Chart 9.2), and without much information on the methodology and sources.[32]

Some of these figures are available over time. Between 1996 and 2007, the percentage of English-speaking Internet users[33] has dropped from 80 per cent to around 30 per cent,[34] reflecting the fact that non-English speakers are increasingly getting online.

As for the number of webpages per language, the published data have shown huge discrepancies since the first studies published back in 1996 (see [Pimienta *et al*, 2010] for more details).

The most promising effort has been made by LOP, which has been able to produce unique figures for minority languages in top-level domains of Africa and Asia. However, the rapidly increasing size of the web makes it more and more difficult to crawl entire top-level domains. The most consistent publisher of data, and the only one with the capacity to produce indicators by subcategories, has been FUNREDES/Union Latine. However, its scope is limited to Latin languages, German and English, and, unfortunately, as its methodology is based on search engines, since 2007 it is no longer in a position to produce reliable figures. Indeed, in recent years the indexes of search engines have stopped reflecting a substantial proportion of the online universe and the results are beginning to show obvious biases and errors in their search occurrence counting figures.

Therefore, just as Internet stakeholders are starting to acknowledge the importance of the production of reliable indicators for linguistic diversity in order to set policies and measure their impact, the task of producing such indicators is becoming even more difficult. New creative approaches are urgently required in order to overcome the technical barriers to measurement created by the exponential growth of cyberspace.

It is also possible to look at what can be referred to as “linguistic productivity,” given by the number of webpages in a given language divided by the number of Internet users in that language. This could offer an indicator of content production.

The FUNREDES/Union Latine studies have produced indications of the production of webpages per language[35] (Tables 9.3 and 9.4). A comparison of the figures over time shows that productivity initially rises in the phase of Internet user growth, and subsequently declines once a certain penetration threshold has been reached. This supports the hypothesis that increasing user access could be the first step towards content creation, followed by digital literacy programmes focusing on why and how to create content in cyberspace. The studies have shown, for instance, that French content was first driven by Quebec, then by Belgium and Switzerland and finally by France, following the growth of Internet penetration in the respective countries. With respect to Spanish content, Portuguese-speaking Brazil produced more Spanish content than most of the smallest countries of Latin America. In 2001, the US produced one tenth of the number of webpages in Spanish produced by Spain, although at that time it had more Spanish-speaking Internet users than Spain. At the same time, Mexico, a country with the highest immigrant population in the US, has a very low content productivity. The findings could be used to design language policies in cyberspace. For example, Spanish content would have been boosted by programmes targeting content production by Spanish-speaking immigrants in the US. Finally, Germany, where English is not the first language, shows an extremely high productivity of English webpages, and, in fact, all OECD countries are producing a significant proportion of their pages in English.[36]

Table 9.3: Production of webpages in French, by region and by country, 2007 (percentage of pages — productivity*)

By country		By region	
France	60% — 1.1	Europe	75%
Canada	20% — 1.1	America	22%
Belgium	7% — 0.6	Africa / Arab States	0.3%
Switzerland	5% — 0.9	Asia / Pacific	0.2%
Other	8% — 0.8	Other	2%

Note: * Computed as the ratio of % of production per % of Internet users in the given language.
Source: FUNREDES/Union Latine.

Table 9.4: Production of webpages in other languages, by country, 2007 (percentage of pages — productivity*)

Spanish		English		Portuguese	
Spain	56 % — 3.4	United States	66 % — 1.0	Brazil	71 % — 0.9
United States	10 % — 0.4	United Kingdom	7 % — 0.6	Portugal	15 % — 1.0
Argentina	9 % — 0.9	Canada	4 % — 0.7	United States	4 % — 5.0
Mexico	8 % — 0.5	Australia	2 % — 0.3	Spain	4 % — 3.7
		Germany	1 % — 39.0		

Note: * Computed as the ratio of % of production per % of Internet users in the given language.
Source: FUNREDES/Union Latine.

Table 9.5: Twenty ccTLDs with the strongest average annual growth of registrations, 2005-2009

	ccTLD		**CAGR 05-09***	**% ccTLD/population 2009****
1	Former Soviet Union	su	102.4	na
2	China	cn	95.9	1.0
3	Tajikistan	tj	88.2	0.3
4	Viet Nam	vn	77.5	0.1
5	Russia	ru	55.8	1.7
6	India	in	54.4	0.0
7	Réunion	re	52.7	0.4
8	Iran (I.R.)	ir	49.4	0.2
9	Venezuela	ve	48.8	0.5
10	Poland	pl	44.5	4.1
11	Latvia	lv	41.0	3.5
12	France	fr	40.0	2.5
13	Portugal	pt	39.4	2.7
14	Guadeloupe	gp	39.1	0.3
15	Lithuania	lt	39.0	3.3
16	Palestinian Authority	ps	36.1	0.1
17	Bosnia and Herzegovina	ba	35.3	0.3
18	Kenya	ke	34.1	0.0
19	Albania	al	32.8	0.1
20	Peru	pe	31.1	0.1

Note: * India 05-08; ** India 2008. CAGR: Compound Annual Growth Rate.
Source: ITU, based on data provided by ZookNIC: www.zooknic.com.

Another way of monitoring the development of local content is by using country code top-level domains (ccTLDss)[37] registered in each country as a proxy.[38] This is based on the assumption that domain-name registration reflects to some extent the availability of local web presence and hence local content even though it provides no information about the languages used. Based on the 98 countries for which data were available in both 2005 and 2009 (or 2008 in some cases), the 20 ccTLDs displaying the strongest average annual growth rates between 2005 and 2009 (except India: 2008) are given in Table 9.5. Most of the countries concerned are growing from very low starting levels, even though in Poland, Latvia, Lithuania, Portugal and France there are already more than two ccTLD registrations per 100 inhabitants.

The 20 ccTLDs with the highest growth rates for the number of ccTLD registrations per 100 inhabitants between 2005 and 2009 are given in Table 9.6. Some of the countries where the number of registrations per 100 inhabitants was already high in 2005 continued to experience strong growth. Many of the relatively smaller European countries may be found in this list.

If ccTLDs are accepted as a proxy for the availability of local content, a greater number of ccTLD registrations per 100 inhabitants could be expected to be statistically associated with higher levels of household Internet usage. Plotting Internet users per 100 inhabitants against the number of ccTLD registrations per 100 inhabitants for 2009 (Chart 9.3) shows a positive correlation (Spearman rank correlation = 0.8, and very highly statistically significant). This shows that there is indeed a rather strong statistical association between these two variables, pointing to the

Table 9.6: The twenty countries with the largest increase in ccTLD registrations per 100 inhabitants, 2005-2009

	Country	% ccTLD/population 2005	% ccTLD/population 2009	Difference (percentage points)
1	Netherlands	10.0	21.2	11.2
2	Switzerland	9.9	17.8	7.9
3	St. Vincent and the Grenadines	0.9	8.4	7.5
4	Denmark	11.6	18.6	7.0
5	Sweden	4.1	9.8	5.7
6	United Kingdom	7.3	12.8	5.5
7	Austria	5.5	10.6	5.1
8	Luxembourg	4.5	9.6	5.0
9	Germany	11.1	15.9	4.9
10	Belgium	4.5	8.9	4.5
11	Australia	2.9	7.1	4.2
12	Norway	5.3	9.2	3.9
13	New Zealand	4.9	8.8	3.9
14	Czech Republic	2.1	5.8	3.7
15	Iceland	4.5	8.2	3.6
16	Poland	0.9	4.1	3.2
17	Greenland	4.4	7.2	2.8
18	Faroe Islands	3.2	6.0	2.8
19	Estonia	2.5	5.3	2.8
20	Hungary	2.0	4.8	2.7

Source: ITU, based on data provided by ZookNIC: www.zooknic.com.

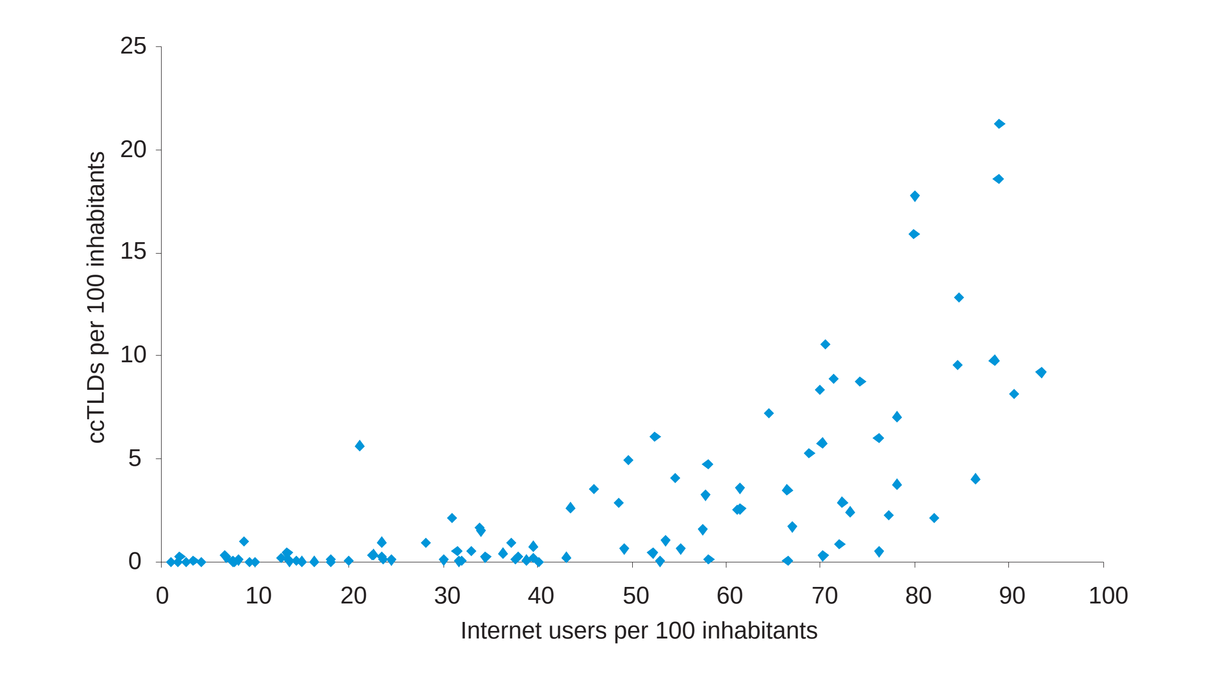
potential importance of providing local content for Internet use. Just as local languages are important, the availability of local content can also be considered key to Internet development and for achieving Internet and ICT-related development goals.

It is not known in which languages these domains exist. However, efforts to measure Internet webpages in certain languages[39] over several years also point to a correlation between the growth in the number of users and the growth of content. Nevertheless, over time, and as the Internet user penetration rate becomes higher, the correlation diminishes and less content per user is produced. This suggests that promoting user-created content is a good candidate for policy consideration, at least in the early stages of Internet and local content development.

Comparison of the production of webpages in Spanish and French shows a higher production of webpages in French, which could be interpreted as the result of policies to boost content creation such as those developed by the *Organisation Internationale de la Francophonie*.[40] See also Box 9.4.

Local content creation also requires more basic digital and information literacy as key components of digital inclusion policies. Appropriate education and technical training could encourage new Internet users to be both producers and consumers of content, sometimes referred to as "prosumers." Beyond formal education systems, libraries can also have an important role to play in promoting digital and information literacy and in creating local content, thus establishing a strong linkage between Targets 9 and 4.

Chart 9.3: Internet users and ccTLD registrations per 100 inhabitants, 2009*



Note: *This includes the 98 countries for which data were available both in 2005 and 2009 (or in some cases 2008). The 2009 Internet users penetration rates are ITU estimates.

Source: ITU, based on Eurostat and national sources and data provided by ZookNIC: www.zooknic.com.

Box 9.4: Active policies to boost content creation: The case of Catalan

In 2006, an original generic name was created "*to serve the needs of the Catalan linguistic and cultural community on the Internet.*"[41] It is an interesting case, since Catalan is a fascinating success story of how a language which had been threatened during the years when Spain was ruled by dictatorship has been able, through appropriate linguistic policies, to make a complete revival. The result of measurements of Catalan webpages in 2006 and 2007 by FUNREDES/Union Latine (see http://funredes.org/lc) has shown a productivity greater than that of Spanish worldwide, though not greater than that of Spanish in Spain, and it was too soon to measure a strong impact in terms of Catalan content progression. Indeed .cat is the first TLD intended to promote a given language and its results in terms of boosting the presence of Catalan on the web deserve to be followed carefully as lessons can be learnt for other languages.

It is not known whether the existence of a specific top-level domain would trigger content creation in a given language, but the promotion effect alone could be an important factor, as witnessed by proposals which have been inspired by .cat, for Welsh (.cym) or Quebec (.qc).

Conclusions and recommendations

Although there is still a long way to go before achieving Target 9, a huge step has been accomplished with the WSIS process insofar as the issue of linguistic diversity has been acknowledged and has been given much higher priority on the global Internet-related policy agenda.

This is the result of a combination of several factors. Many players in the field have worked to draw attention to the issue of linguistic diversity on the Internet, from governments (like the Canadian Government *International Develop-*

ment Research Centre, IDRC[42] in Asia[43] and Africa[44] with its Pan programmes) to civil society (like, for instance, the World Network for Linguistic Diversity, MAAYA).[45] The drastic changes in Internet demographics in recent years have also contributed to raising awareness of this issue, especially the rapid increases in the number of Internet users from the Asia and Pacific and the Arab States regions.

Until recently, the (technical) complexity of the issue and measurement costs, coupled with the idea that English can act as a *lingua franca* for the Internet, have held back progress on this target. However, attitudes are changing and the subject has gained higher visibility at the last three annual meetings of the Internet Governance Forum, which is driving WSIS multistakeholder participation, and at the WSIS Forum meetings held annually in Geneva.[46]

Some changes have already taken place. The issue of internationalized domain names (IDN) has received considerable attention. More than half of the 1.7 billion people who use the Internet speak languages with non-Latin scripts,[47] highlighting the importance of making provision to support all languages on the Internet. In 2009, the Internet Corporation for Assigned Names and Numbers (ICANN) approved plans to open up Internet domain names to non-Latin script characters and allow domain names in Arabic, Chinese and other scripts.

The implementation of domain names in local languages, which is now under way, is likely to increase demand for linguistic diversity on the Internet, adding a bottom-up driver to the top-down efforts made at the political level and in the context of the WSIS process. The number of initiatives to promote linguistic diversity is rapidly increasing, and this process can be expected to show some tangible outcomes by 2015, with likely increases in the number of languages that can be used on the Internet, the availability of local content, and the number of language versions of the main software and applications used on the Internet.

Nonetheless, the lack of available indicators to measure Target 9 makes it difficult to track progress on its achievement, at both the international and national levels. While there may be some anecdotal evidence and success stories, which may serve as good-practice ideas for other initiatives, there are not enough data to allow an objective assessment of where the target currently stands. The main global stakeholders need to work together, highlight the importance of the issues at stake, guide countries in what policies they could implement, and potentially come up with and implement a plan for concrete actions. Indeed, the objective of linguistic diversity in cyberspace warrants support on the global ICT and Internet-related policy agenda, not least to be able to exploit synergies among the various stakeholders.

Collecting indicators in this field calls for a recommendation to the main global stakeholders. Following the Barcelona symposium held in September 2009,[48] an urgent effort is required to set up an action plan which should include:

- incentives to the scientific community for initiating research projects geared to creating indicators for measuring languages and content on the Internet;
- direct actions coordinated by international organizations to support projects for the development of indicators;
- mainstreaming linguistic diversity in all aspects of ICT and information society measurement efforts.

International organizations can further contribute by making governments more aware of the importance of linguistic diversity in cyberspace. This could include suggestions for programmes supporting the issues, and appropriate guidelines.

Strategies also need to be designed to support the IDN implementation process and use it as a starting point for building a more comprehensive perspective on the subject. In addition to the implementation of IDNs, many efforts are emerging in different countries and sectors. It would be important to take stock of these initiatives and improve coordination between them.

ICT4D donors should consider mainstreaming linguistic and cultural diversity in their project frameworks, as was done for gender issues. One possibility would be to consider giving higher funding priority to projects that specifically take linguistic diversity and local content into account. Developing countries' diasporas could also play a strategic role in all ICT4D efforts, and should be involved in content production for local languages.

Technological innovation aimed at bringing communities not yet represented on the Internet online, and at improving digital literacy, including through the development of mobile web technologies, should be promoted.

Finally, the role of translators should be valued in the context of linguistic diversity, as well as cybervolunteer networks[49] that can help with the promotion of languages online. The Bamako International Forum on Multilingualism[50] was established in January 2009. An action plan was drawn up containing a chapter entitled "For a multilingual cyberspace," including a broad set of recommendations which could serve as valid inputs.[51] As a follow-up, and as a way to push the transformation of recommendations into an action plan, the main stakeholders should set up and commit to a coordinated roadmap, including continuous monitoring of progress in linguistic diversity and local content in cyberspace.

Notes

[1] Substantial contributions to this chapter were made by Daniel Pimienta from FUNREDES, with inputs from Daniel Prado (Union Latine), Jean-François C. Morfin (Intlnet), Viola Krebs (ICVolunteers), and Deirdre Williams (St Lucia). ITU is also very grateful for the ccTDL data provided by Matthew Zook from ZookNIC.

[2] On the Internet, content is any information (webpages, messages, software...) that is available for retrieval by the user, in any format (e.g. text, image, audio, video).

[3] The term *Web 2.0* refers to what is perceived as a second-generation of the web, characterized by dynamic, shareable and user-generated content, such as in social networks.

[4] It is worth noting that the size of the "invisible web" — the part of the web containing dynamic pages which are not indexable (such as data bases) — is estimated as being 500 times larger than the visible web [Bergman, 2001].

[5] See WSIS Geneva Plan of Action, 2003, at: http://www.itu.int/wsis/docs/geneva/official/poa.html#c8, §23.

[6] The term "Internet of objects," or "Internet of things," alludes to the idea that many devices (such as household coffee machines or refrigerators) will eventually be accessible via the Internet through innovative applications. See http://www.itu.int/wsis/tunis/newsroom/stats/The-Internet-of-Things-2005.pdf for more details.

[7] Internet Protocol version 6 (IPv6) is the next-generation Internet layer protocol for packet-switched networks and the Internet.

[8] IPv4 addresses use 32 bit/s, while IPv6 uses 128 bit/s, allowing 2128 (about 3.4×1038) addresses.

[9] See the eleven barriers to use of ICT for human development, presented in a spiral form in [Pimienta, 2009].

[10] It is not straightforward to describe individual decisions to produce and browse or not in a particular language. Nonetheless, incentives could be put in place to stimulate the production of local content, and/or content in local languages.

[11] Information and Communication Technologies for Development (ICT4D) refers to the application of ICTs for development in general and poverty reduction in particular.

[12] Local content production does not imply that the locally produced content is only consulted locally. As a matter of fact, empirical data lead to the observation that locally produced content in a globalized world tends to be consulted globally. The existence of diasporas and of people interested in research and study of local cultures from abroad are only two of the reasons to explain that phenomenon.

[13] All references from *FUNREDES/Union Latine* can be consulted in the FUNREDES/Union Latine Observatory of Linguistic and Cultural Diversity at: http://funredes.org/lc.

[14] The Language Observatory Project (http://www.language-observatory.org/) is a consortium of universities initiated and coordinated by the University of Technology of Nagaoka, Japan. For more information, see http://gii2.nagaokaut.ac.jp/gii/.

[15] "Localizing content" refers to the process of translation into the local language and adaptation of the content to the local culture. "Localizing language" is meant in this chapter as the process of allowing a language to be usable in the digital environment, and it refers then basically to a technical process of codification of the corresponding character set.

[16] A *lingua franca* is a language systematically used to communicate between persons not sharing a mother tongue.

[17] While figures for speakers of a given language (especially a second language) vary considerably and up-to-date figures are hard to find, the low rate of change allows the use of less recent figures. In 1996, David Graddol, writing for the British Council, estimated "speakers of English" to be 750 million (source "The Future of English" — http://www.britishcouncil.org/learning-elt-future.pdf), which, compared with the 5.7 billon world population for 1996, gives a percentage of 10 per cent. Ethnologue, citing sources from 1999, gives a figure of a little more than 508 million English speakers (341 million first language) which, compared with the 6 billon of world population for 1999, gives a percentage of less than 8.5 per cent. [Malherbe, 1999] offered the figure of 600 million, which set against the 1999 world population gives a percentage of a little more than 10 per cent. Union Latine (cited in [Pimienta, 2009]) gave a figure of 670 million in 2008, which also means a proportion of just over 10 per cent. Other sources may give higher figures (for example [Prado, 2010] provides estimates of 824 million or 729 million for 2009 depending on the assumptions made), which is why it is necessary to leave a margin by stating less than 15 per cent. In a similar context, [Graddol, 2006], using data from the World Tourism Organization, states that three-quarters of tourism situations involve non-English speakers in non-English-speaking countries.

[18] Children who are not educated in their mother tongue have lower grades as well as a higher rate of failure to go on to the next level (see http://www.unesco.org/en/languages-in-education/).

[19] Practical, or technical, computer language also does not exist in all languages. For example, words such as *copy, paste, file, mouse, drive, operating system, network, burning a CD, download, broadband, Instant Messenger,* etc. are often taken from English. This might constitute a barrier to people who do not speak English or do not feel at ease using English terminology.

[20] *Test Of English as a Foreign Language*. See http://www.ets.org. This is an imperfect proxy, because TOEFL exams tend to be taken by those who plan to live, work or study abroad. Therefore, it tends to be a relatively small proportion of individuals in a country who take TOEFL exams. Furthermore, this proportion may vary widely across countries, and is likely to reflect other factors such as the resources allocated to English teaching. Nonetheless, in spite of these limitations, TOEFL test scores currently constitute the only widely available measure of English proficiency where English is not the mother tongue.

[21] [OECD, 2006] and [van Welsum and Xu, 2007] use TOEFL scores as a proxy for English language skills, and [Lee, 2009] uses TOEFL scores to proxy the effect of English proficiency on economic growth.

[22] In Asia and Africa, IDRC supports the localization of local languages through initiatives which gather many partners from various countries, in a collaborative manner. The Asian group is also strongly involved in the process of the development of IDNs. Both groups organize regional meetings which contribute to the necessary awareness-raising among stakeholders.

[23] However, broadband evolution and the combined management of audio and video may eventually open the door for some type of digital processing for oral languages, while the Internet is becoming less and less text centred.

[24] Digital literacy consists of equipping people with ICT concepts, methods and skills to enable them to use and exploit ICTs. The related concept of information literacy consists of providing people with concepts and training in order to process data and transform them into information, knowledge and decisions. It includes methods to search and evaluate information, elements of information culture and its ethical aspects, as well as methodological and ethical aspects for communication in the digital world.

[25] Based on data from [Crystal, 2006], Ethnologue and UNESCO.

[26] Some sources state than only one tenth of languages have a written form.

[27] People with linguistic skills researching on the demographic elements of languages.

[28] UNICODE (http://unicode.org), one of the pillars of the progress made in multilingualism, is a non-profit association formed as a consortium which sets standards for encoding character sets and scripts.

[29] See http://unicode.org/standard/supported.html and http://unicode.org/iso15924/iso15924-num.html for more complete explanations and for the set of supported scripts.

[30] Balancing Act Africa, Issue No. 481, 20 November 2009 (http://www.balancingact-africa.com/).

[31] Thanks to the "Wayback engine" of archive.org, it is possible to see a snapshot of one of the latest products: http://web.archive.org/web/20041019013615/www.global-reach.biz/globstats/index.php3.

[32] http://www.internetworldstats.com/stats7.htm.

[33] These figures refer to both first and second language English speakers.

[34] The last figure for 2009, given by InternetWorldStats, is 27.6 per cent of English-speaking users on the Internet (source http://www.internetworldstats.com/stats7.htm).

[35] Limited to English, French, Portuguese and Spanish.

[36] See http://funredes.org/lc for more details.

[37] Top-level domains (TLDs) are divided into two classes. Generic top-level domains (gTLDs) include for example ".com" or ".org," while country code top-level domains (ccTLDs) are used and reserved for countries or dependent territories expressed as two letter country codes [OECD, 2009].

[38] This is an imperfect proxy, as it does not take generic top-level domains (gTLDs) registered in each country into account, nor the fact that these registrations can be influenced by factors such as registration pricing policies. The relative shares of ccTLDs and gTLDs vary across countries. Nonetheless, domain-name registrations are an indicator of interest in adopting a web presence and ultimately an indicator of the development of the Internet [OECD, 2009].

[39] See http://funredes.org/lc and the observation of Latin languages, German and English.

[40] The *Organisation Internationale de la Francophonie* (OIF), a governmental organization bringing together 56 States, contributes to promoting diversity. Its four main action lines include one for "digital technologies." OIF was present during the WSIS process as an advocate of linguistic and cultural diversity in cyberspace. Supporting content creation in French and in the local languages of the Member States is an indirect part of the objective "to support and value digital expression of francophone communities." This translates, for example, into a programme to fund initiatives in the field (Information Highway Francophone Fund). Progress in content development in Spanish and French, and the productivity of French content relative to other languages, suggests that policies to support the production of local content can have measurable effects.

[41] See http://www.puntcat.cat/en_index.html.

[42] http://idrc.ca.

[43] http://www.panl10n.net/.

[44] http://www.panafril10n.org/.

[45] The MAAYA network brings together more than 25 partners with a dynamic of fostering collaboration in concrete projects. UNESCO, which is a member of the MAAYA network, is the UN body which promotes the theme of linguistic diversity, both in the real world (protecting endangered languages) and in the virtual world (for instance in a signed agreement with ICANN) and, as mentioned in the introduction, coordinates with ITU for WSIS Action Line C8. See http://maaya.org.

[46] See for instance the plenary session chaired in Rio de Janeiro by the Brazilian Minister of Culture in http://www.intgovforum.org/Rio_Meeting/IGF2-Diversity-13NOV07.txt.

[47] See http://news.bbc.co.uk/2/hi/8333194.stm.

[48] International Symposium on Multilingualism and Cyberspace, see http://maaya.org/spip.php?article105.

[49] In Latin America, the project "En mi Idioma" (in my language) enables the Misak community of Colombia to design a website for language education. Enlace Quiche, in Guatemala, trains bilingual teachers and supports the production of educational packages in Maya

languages. ICVolunteers.org coordinates a network of volunteers, involved in cybervolunteerism and translation (over 10 000 volunteers registered). Its language and migration project assists migrants with their language needs and makes use of the web to do so.

[50] http://www.acalan.org/fr/confeven/forum/forum.php.

[51] See http://www.acalan.org/fr/confeven/forum/plan_action.pdf.

References

Bergman, M.K. (2001), White Paper: The Deep Web: Surfacing Hidden Value. in Ann Arbor, MI: Scholarly Publishing Office, University of Michigan, University Library vol. 7, no. 1, August, 2001 http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104.

ComScore (2008), “Social Networking Explodes Worldwide as Sites Increase their Focus on Cultural Relevance,” http://www.comscore.com/Press_Events/Press_Releases/2008/08/Social_Networking_World_Wide.

Crystal, D. (2006), “Language and the Internet,” Second Edition, Cambridge University Press, 9/2006.

Diki-Kidiri, M. (2007), “Comment assurer la présence d’une langue dans le cyberespace?,” Unesco CI.2007/WS/1., 2007, http://unesdoc.unesco.org/images/0014/001497/149786F.pdf.

European Commission (2009), *Europe’s Digital Competitiveness Report*, European Commission, Brussels.

Facebook (2010a), Company Timeline page, http://www.facebook.com/press/info.php?timeline (accessed 2 March, 2010).

Facebook (2010b), Statistics page, http://www.facebook.com/press/info.php?statistics (accessed 2 March, 2010).

Facebook (2010c), “Facebook’s February 2010 US Traffic by Age and Sex: All Groups Growing, Men More Quickly,” http://www.insidefacebook.com/ (accessed 2 March, 2010).

Graddol, D. (2006), “English Next,” British Council, http://www.britishcouncil.org/learning-research-english-next.pdf.

Horrigan, J. (2010), “Broadband Adoption and Use in America,” Results of FCC Survey, OBI Working Paper Series No. 1.

ITU (2010), “Measuring the Information Society 2010,” Geneva.

Lee, C. G. (2009), “English language and economic growth: Cross-country empirical evidence,” paper prepared for the Conference on Globalisation, Development and Growth in Asia,” held at the University of Nottingham, Malaysia Campus, January 14-15, 2009.

Lewis, M. (2009), “Ethnologue: Languages of the World, Sixteenth edition,” Dallas, Tex.: SIL International. http://www.ethnologue.com/.

Malherbe, M. (1999), “Les langues de l’humanité,” Editor: R. Laffont, Paris.

Nandasara, S.T., Kodama, S., Choong, C,Y., Caminero, R., Tarcan, A., Riza, H., Nagano, R.L. and Y. Mikami. (2008), “An Analysis of Asian Language Web Pages,”.in The International Journal on Advances in ICT for Emerging Regions (ICTer), Volume 1, No 1, 2008.

OECD (2006), *2006 OECD Information Technology Outlook*, OECD, Paris.

OECD (2009), *OECD Telecommunications Outlook 2009*, OECD, Paris.

Paolillo, J., D. Pimienta, D. Prado, Y. Mikami and X. Fantognanet (2005), “Measuring linguistic diversity on the Internet,” UNESCO, Paris, http://portal.unesco.org/ci/en/ev.php-URL_ID=20882&URL_DO=DO_TOPIC&URL_SECTION=201.html.

Pimienta, D. (2008), “Accessing Contents,” in: Global Information Society Watch, APC, Hivos and ITeM. http://www.giswatch.org/gisw2008/thematic/AccessingContent.html.

Pimienta, D. (2009), “Digital divide, social divide, paradigmatic divide,” in: International Journal of Information Communication Technologies and Human Development, V1, N1, January-March 2009. Older version accessible at: http://funredes.org/mistica/english/cyberlibrary/thematic/Paradigmatic_Divide.pdf.

Pimienta, D., Prado D. and A. Blanco (2010), “Twelve years of measuring linguistic diversity in the Internet: balance and perspectives,” UNESCO, Paris. http://unesdoc.unesco.org/images/0018/001870/187016e.pdf.

Prado, D. (2010), “Présence des langues romanes dans la connaissance,” presentation made at “Présence, poids et valeur des langues romanes dans la société de la connaissance,” OIF, Paris, 30 April 2010.

Suzuki, I., Y. Mikami, A. Ohsato and Y. Chubachi (2002), “A Language and Character Set Determination Method Based on N-gram Statistics,” in: ACM Transactions on Asian Language Information Processing, Vol.1, No.3, September 2002, pp.270-279.

UNESCO (2005), Convention on the Protection and Promotion of the Diversity of Cultural Expressions, Document CLT-2005/CONVENTION DIVERSITE-CULT REV., Paris.

van Welsum, D. and Xu T. (2007), “Is China the new centre for the offshoring of ICT-enabled services?,” DSTI Information Economy Report, March 2007, OECD, Paris.