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

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

Executive summary

The emergence of an information society requires more than access to infrastructure, equipment and services. For users, infrastructure, equipment and services are important because they provide access to information – or content – that they can use in order to pursue their objectives (as governments, businesses, organisations or individuals) and can share with others. They need to be able to find content that is affordable, relevant to their needs and circumstances, and available in languages that they can understand, with formats they can use.

Target 9 deals with the development of content and the technical means required to facilitate linguistic diversity. While these are critical aspects of the information society, quantitative measurement of content and language is challenging both because of limited data and rapid developments in content provision and platforms on which content is accessed. Major developments in these areas since WSIS include the rapid growth of user-generated content, including online social networks; the emergence of mobile apps; and the growing significance of cloud computing models of content storage and access. The opportunities arising from automated translation are now having significant impact on availability of content by language, while the period since WSIS has also seen the introduction of top level Internationalised Domain Names.

Five indicators were selected for monitoring Target 9 (*Partnership*, 2011). Four of these are concerned primarily with language, while one seeks to provide a proxy for local content.

The first two indicators – the proportion of Internet users by language, at country and global levels – have proved extremely difficult to measure because of severe data limitations. Few countries have so far included sufficient questions on Internet usage and language in national censuses and household surveys to enable assessment of Internet user numbers by language, while global estimates currently available are out-of-date and of questionable statistical value. There are also statistical challenges in identifying the language characteristics of national populations. However, it is clear from the evidence that is available that the linguistic diversity of Internet users has increased since WSIS. The proportion of Internet users whose primary language is English has fallen significantly as access to the Internet has become more widespread, from an estimated 80 per cent in the mid-1990s to less than 30 per cent in 2011. More than 300 languages are now available on Wikipedia and more than 100 on major social networks. There has been particularly strong growth in the number of Chinese speakers online.

No satisfactory data are currently available to measure the third indicator selected in 2010 – the number of webpages by language. Comprehensive analysis of this indicator requires a combination

of web-crawling and language identification techniques that has not been undertaken by independent research institutes since 2007. Only very limited data are available from commercial sources. What is available suggests that there is growing linguistic diversity in web content, although English remains the most widely used language on websites. Wider linguistic diversity is likely to be found in user-generated content on social networking and other sites.

The fourth indicator – the number of domain name registrations for each country code top level domain (ccTLD), weighted by population – was selected to serve as a proxy for local content, that is, for content created within each country. The value of this indicator can be improved by including geolocated registrations of global top level domains (gTLDs) within countries and by comparing findings with the number of Internet users as well as total population. Data made available for this report show that the number of domains registered per head of population has been falling by almost three-quarters worldwide since WSIS, but is still much lower in developed countries (18 ccTLDs or 6 TLDs p.c.) than in developing countries (241 ccTLDs or 131 TLDs). The number of Internet users per TLD is falling in developed countries but growing in developing countries because of their high rate of growth in Internet usage.

The fifth indicator – the number and share of Wikipedia articles by language – serves as a proxy for user-generated content online. Extensive data published by the Wikimedia Foundation show that there has been a marked decline in the proportion of articles in English, from 46 per cent of all articles in 2003 to 15 per cent in 2013, and a corresponding increase in the proportion of articles in languages that are not among the ten most-used international languages (from 26 per cent to 58 per cent). However, corresponding data on pageviews show that there is a higher level of linguistic concentration in access and use of Wikipedia content.

The chapter also includes some available data on website usage and on user-generated and social media.

The quantitative measurement of content and language is challenging. Should there be a post-WSIS target related to content and language, it is recommended that indicator 9.1 be retained but suspended until there is more widespread collection of data on Internet usage cross-classified by language spoken (for example, collected by national statistical offices in population censuses); that indicator 9.4 should be retained following revision to include gTLDs and IDNs as well as ccTLDs; and that indicator 9.5 should also be retained but extended to include Wikimedia data on content creators and pageviews. It is recommended that indicators 9.2 and 9.3 should be withdrawn because no satisfactory data are likely to become available for these, but that consideration should be given to including indicators related to online social networks and mobile apps. The possibility is also suggested of incorporating qualitative data in the monitoring of content and language, and of building a more comprehensive portfolio of quantitative and qualitative data for selected representative countries.

Introduction

Eight of the ten targets that were adopted in the WSIS Geneva *Plan of Action* are concerned primarily with access to infrastructure and to facilities that enable effective use of access. However, access to physical infrastructure and facilities is only one aspect of the enabling environment for the effective use of Internet and other online services. Other factors that are essential in enabling the development potential of ICTs to be fulfilled include: the affordability of access; the presence of relevant skills among potential users (including literacy, computer literacy and research and analytical skills); and the availability of relevant content that is readily accessible to users. Accessibility, in this context, is highly dependent on the language(s) in which content is available.

WSIS Target 9 addresses these issues. It has two distinct but interlinked concerns:

- to encourage the development of content online and
- to put in place technical conditions that facilitate the presence and use of all world languages[1] on the Internet.

The availability of content and linguistic diversity are not new issues for the information society, but have been critical to the dissemination of information and knowledge in earlier communications media, including speech, print and broadcasting. Access to information and, thereby, knowledge is a principal factor in enabling individuals to maximise economic opportunity and social networks; in spreading knowledge of health and other social issues; in facilitating business innovation; and in enabling governments to develop policies and programmes that effectively address the social and economic needs of their societies. It is at the heart, therefore, of sustainable development, a primary focus of the post-2015 development agenda, as well as of progress towards the information society. Two aspects of this are equally important:

- the publication of information or content and
- the ability of people and organisations to access content and interact by sharing information with others through communications platforms.

These can be described as the supply and demand sides of content.

Innovations in information and communication technologies and markets before and since WSIS have greatly extended access to both information and interactive communications, creating the opportunity for individuals, organisations, businesses and governments to make more effective use of information in enhancing development outcomes. In particular:

- The Internet has greatly extended the amount of content that is published or made publicly accessible, particularly through the WWW, and greatly extended the reach of that content to much wider groups of potential users. Subject to restrictions in some countries, Internet users anywhere in the world can access the great majority of content that is published online. Access can be easily shared through e-mail, instant messaging and other online platforms as well as offline networks.
- Since WSIS, interactive services such as online social networks, blogs, microblogs, and audio and video file-sharing services – sometimes referred to as Web 2.0 services – have expanded greatly, facilitating enormous growth in publication of user-generated content and interactive information-sharing amongst Internet users.

- Also since WSIS, mobile telephones have evolved from voice telephony devices into multipurpose digital devices that are widely used to share audio, image and video content and to access the Internet as well as for voice and text communications. A new content market has emerged around smartphone applications (mobile apps), overlapping with and supplementing online content accessed through the Internet. The number of apps available for Apple iPhones was reported to have exceeded 1 million in October 2013,[2] while the number of Android apps was reported to have exceeded 1.1 million by February 2014.[3]

Access to information depends on the affordability, availability and accessibility of content as well as connectivity. These different factors are interlinked. A study published jointly by the OECD, UNESCO and the Internet Society (2011) found that there was both "… a strong correlation between the development of network infrastructure and the growth of local content" and "… a significant relationship between the development of international bandwidth and the price of local Internet access," suggesting a virtuous circle between infrastructure, affordability and content production. The availability of relevant skills is also critical to individuals' and organisations' ability to make use of content. UNESCO has developed a set of largely qualitative indicators that can be used to monitor the extent of media and information literacy within different societies, including the quality of national ICT environments, content access, the availability of analytical capabilities and content generation.[4]

Literacy is obviously important in enabling access to content. Some 775 million adults worldwide are estimated to be illiterate, over 10 per cent of the world's population, the majority of them in developing countries.[5]

The languages in which content is available are equally important in determining its accessibility to potential users. There are a little over 7 000 languages in use worldwide today, whose primary speakers are distributed as set out in Table 9.1.

Table 9.1: Distribution of world languages by area of origin

Region of origin of language	Living languages		Number of speakers	
	Count	Percentage	Millions	Percentage
Africa	2,146	30	789	13
Americas	1,060	15	51	1
Asia	2,304	32	3,743	60
Europe	284	4	1,647	26
Pacific	1,311	19	7	0.1
Totals	**7,105**	**100**	**6,236**	**100**

Source: Ethnologue, http://www.ethnologue.com/statistics, viewed 4 March 2014.

Note: speakers of a language in this table are not necessarily located in their primary language's region of origin.

Of these, at least 24 have more than 50 million and 85 more than 10 million first language speakers, some of these distributed amongst numerous versions or dialects. However, almost 50 per cent of living world languages have fewer than 10 000 first language speakers and many of these do not have written form. Some countries have especially large numbers of languages, most notably Papua New Guinea whose seven million inhabitants share 836 different tongues.[6] Estimated language distribution by speaker numbers is set out in Table 9.2.

Table 9.2: Distribution of world languages by number of first language speakers

Population range	Living languages		Number of speakers	
	Count	Percentage	Millions	Percentage
100 million to 1 billion	8	0.1	2,528	41
10 million to 100 million	77	1	2,382	38
1 million to 10 million	308	4	963	15
100,000 to 1 million	928	13	295	5
10,000 to 100,000	1,798	25	61	1
1,000 to 10,000	1,984	28	8	0.1
Fewer than 1,000	2,002	28	0.5	0
Totals	**7,105**	**100**	**6,236**	**100**

Source: Ethnologue, http://www.ethnologue.com/statistics/size, viewed 4 March 2014.

The majority of published content has always, therefore, only been available in a limited range of languages that are much more widely used, particularly languages that have widespread international reach (such as Arabic, French, Portuguese, Spanish and English) and/or are the principal languages in countries with large populations and diasporas (such as Russian and Chinese). This has continued to be the case with professionally published content on the Internet, such as webpages, though minority languages appear to be more widely used in interactive and user-generated content (such as e-mail, instant messaging and social networks), as in voice telephony.

This imbalance in favour of content in a small number of languages has resulted in concern, reflected in this target, over the need for greater linguistic diversity online in order to ensure that all people are able to access content that is relevant to them in a language that is accessible to them, particularly their mother tongue. In 2003, UNESCO adopted a recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace.[7]

Particular attention has been paid by UNESCO and other agencies to the survival of threatened languages, including those spoken by indigenous peoples, and to the preservation of information and knowledge expressed in those languages, in line with the United Nations *Declaration on the Rights of Indigenous Peoples*[8] and other international instruments.

The principal WSIS Action Line that is concerned with content and language is Action Line C8, whose remit covers "… cultural diversity and identity, linguistic diversity and local content" and which is facilitated by UNESCO. Its priorities include:

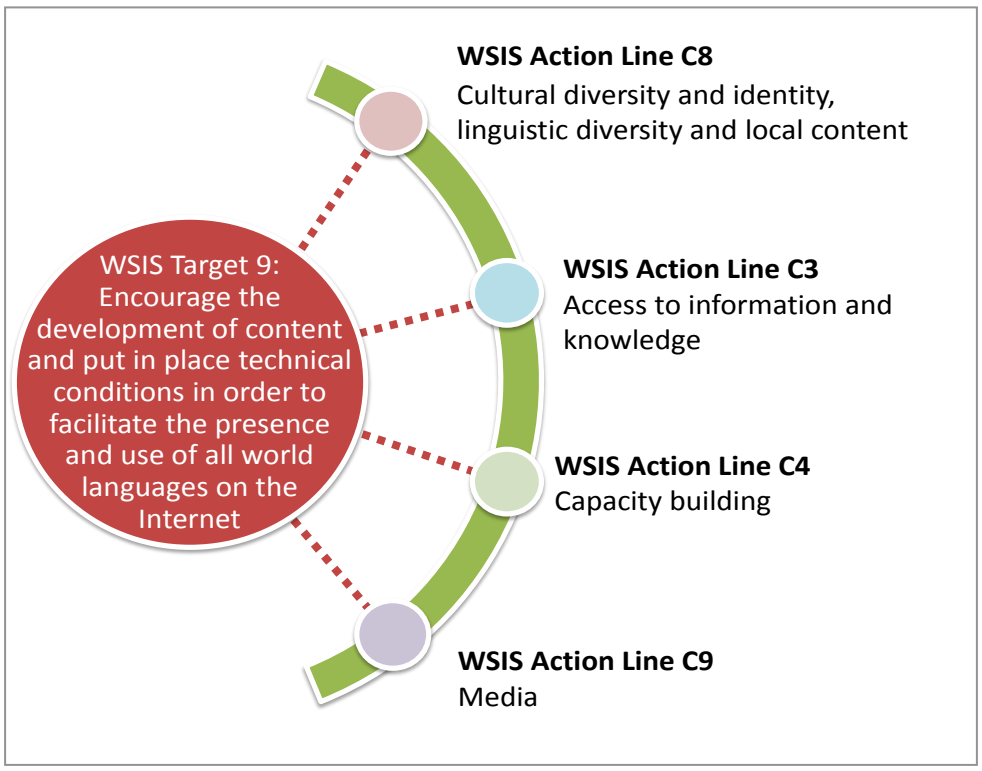
- the development and implementation of policies that "… preserve, affirm and promote diversity of cultural expression and indigenous knowledge and traditions"
- support for "… local content development, translation and adaptation"
- the provision of "… content that is relevant to the cultures and languages of individuals," including marginalised groups such as those who are not literate
- the promotion of software in local languages
- the promotion of technologies and research in areas such as translation, voice-assisted software, multilingual search engines and internationalised domain names, which have the potential to increase the accessibility of content to those who do not have requisite language skills.[9]

Other WSIS action lines are also relevant to content development, access and usage, in particular:

- Action Line C3 encourages governments to promote access to content, including public domain information.
- Action Line C4 is concerned with capacity building, including the eradication of illiteracy, and developing the capabilities of marginalised communities to generate local content.
- Action Line C9 is concerned with media.

This is illustrated in Figure 9.1.

Figure 9.1: Relevance of Target 9 to WSIS action lines



Definitions and challenges of measurement

As noted above, Target 9 is concerned with both:

- online content and
- linguistic diversity online.

The following paragraphs define these terms and discuss the principal difficulties affecting measurement.

Content

The term "content" (or "digital content") is usually used, in the context of ICTs, to include all information and data that are available through digital platforms and services. This includes content on broadcasting platforms, in SMS messages and in mobile apps, as well as content on the Internet. The term "online content" has generally been used more narrowly to refer to content available through the Internet, including webpages; content on social media platforms; and downloadable

material in text, audio, video and other formats. However, these definitions are neither precise nor fixed and need to evolve over time as ICT technology and markets change.

WSIS Target 9 has generally been understood to refer primarily to content on the Internet, and this understanding remains valid for this report. Assessing the target in 2010, the *World Telecommunication/ICT Development Report 2010: Monitoring the WSIS Targets, A mid-term review* (WTDR 2010) defined 'content on the Internet' for this purpose as "… any information (webpages, messages, software …) that is available for retrieval by the user, in any format (for example, text, image, audio, video)."[10]

Although 'local content' is not explicitly addressed by the target, it is substantially addressed in the associated WSIS Action Line. There is no generally agreed definition of local content, which has both geographic and linguistic resonance. Some use the term narrowly to refer to information that is specifically and directly relevant to local communities. In 2011, UNESCO, the OECD and the Internet Society defined it, more widely, to include "… all digital content created for an end user who speaks the same language as the author."[11] Others, including the Partnership on Measuring ICT for Development, have sought to use content published on ccTLDs (country code top level domains) as a proxy for local content (see discussion of Indicator 9.4 below). However, these latter approaches are likely to include much content that is global rather than local in character – for example video material distributed by local online broadcasters or content aggregators – and to exclude much that is local rather than global – for example content posted on global social media platforms such as Facebook, Twitter and YouTube.

There are several points along the value chain at which content can, and arguably should, be measured. UNESCO, the OECD and the Internet Society identified four stages of content production and dissemination in their 2011 analysis of the relationship between infrastructure, affordability and local content:[12]

- **content creation** – the production of content intended for distribution online and the preparation of other content for online distribution
- **content preservation** – including the hosting of content
- **content dissemination** – the publication of content, and enabling of access to content, on websites, social media platforms, mobile apps and through other online media
- **content utilisation** – the extent to which content is accessed by online users and the extent to which content is then used to achieve wider objectives – whether the personal objectives of individual users, the commercial objectives of businesses, or the development objectives of governments and other stakeholders.

Less attention was paid to these content dimensions than to language aspects in the selection of Indicators for Target 9 set out in the *Measuring the WSIS Targets: A statistical framework* (*Partnership*, 2011). Ways of addressing these content dimensions more effectively in future are discussed later in this chapter.

Language

The range and diversity of languages across the world were briefly summarised in the introduction to this chapter. The importance and challenges of measuring linguistic diversity on the Internet arise particularly in two parts of the content value chain:

- content creation and publication (the supply of content)
- content access and usage (demand for content).

A number of challenges constrain the measurement of content creation and publication. The technical structure of the Internet is based on generic and geographic domains rather than on languages. Content can be published by any Internet user, through websites or social media platforms, and is not formally categorised by language when posted. Script recognition algorithms can be used to identify scripts used in posted text, but many of these are shared by several languages – in the case of Latin script, by hundreds. Language recognition algorithms have been developed that can be applied to text and these can be particularly valuable in distinguishing between major languages. However, text analysis does not cover audio, image and video content.

In any event, the size of the World Wide Web (WWW) is now so great that the web in its entirety can no longer be readily analysed through web crawlers (indexation programmes such as those used by search engines). Random samples of web content could be gathered by web crawlers for language analysis, but very large samples would be required to make this statistically viable, which would be expensive. Analyses prepared by Internet companies for business development purposes are unlikely to be made available because of commercial confidentiality.

It is equally difficult to measure content access and usage by language. Almost all countries are to some extent multilingual:

- Some countries have several official languages that are used as primary languages by substantial groups within the population. South Africa, for example, has 11 official languages, including one global language (English) and ten that are spoken predominantly within the country or shared with one of its neighbours, as well as a number of other mother tongues not designated as official languages.[13]
- Many countries designate a global language such as English or French as an official language, even though it is spoken by only a minority of citizens. Some countries, equally, have a *lingua franca* – for example Hindi or Kiswahili – that is widely used alongside both local mother tongues and global languages like French and English. It is estimated that the number of people who understand English may be around 15 per cent of world population: a significant proportion but still relatively low compared to the overall population of Internet users.[14]
- Only two territories are identified by Ethnologue as being monolingual.[15] Most countries have substantial linguistic minorities, with a range of secondary language capabilities, whose Internet access may be better or worse than that of other linguistic groups, on linguistic or other grounds. Examples include speakers of global languages within national minorities, such as Spanish speakers in the United States and Chinese speakers in Malaysia, as well as speakers of local minority languages.

For these reasons, it is misleading to attribute a single language to a country when assessing linguistic diversity.

Secondly, a large proportion of individuals are multilingual, to greater or lesser degree, the extent of multilingualism often (but not necessarily) being associated with higher educational attainment. In Tanzania, for example, while a family may speak a local mother tongue at home, primary education is delivered in Kiswahili and secondary education in English. Many people in countries in West Africa speak French or English, and one or more local languages. Hindi is spoken by many people in India as

a second language, alongside their mother tongue, while English is also spoken or understood by a significant proportion of the population.

For these reasons, it is problematic to associate individuals' ability to access content with the availability of content in their mother tongue. This is particularly so where mother tongues have small speaker communities or where speakers have almost universal familiarity with other languages (for example because these are the languages of primary education). At the same time, it should be recognised that the availability of content in local languages is important for sustaining cultural identity, particularly for minority communities and indigenous peoples, and should not be judged solely as a route to information.

Thirdly, the extent to which a language is present online is not a simple binary question. The WTDR 2010 identified 12 factors that contribute to the online presence of a language, as well as content itself, including:

- a written form for the language
- codification of its script, alphabet and suitable fonts
- the availability of suitable hardware, such as keyboards
- linguistic software (such as word processing and browsing programmes, spell checkers and dictionaries) that enable content to be developed and viewed in the language concerned
- an informed user community driving content production in the educational and creative sectors and the media
- the availability of automated translation enabling access to content published in other languages
- indexation of content by search engines and other intermediaries.[16]

Developments since WSIS

Major developments have taken place in the environment for online content and language since the World Summit. As with other WSIS targets, these affect both what is and should be measured in relation to Target 9, and the potential for successful measurement.

Content

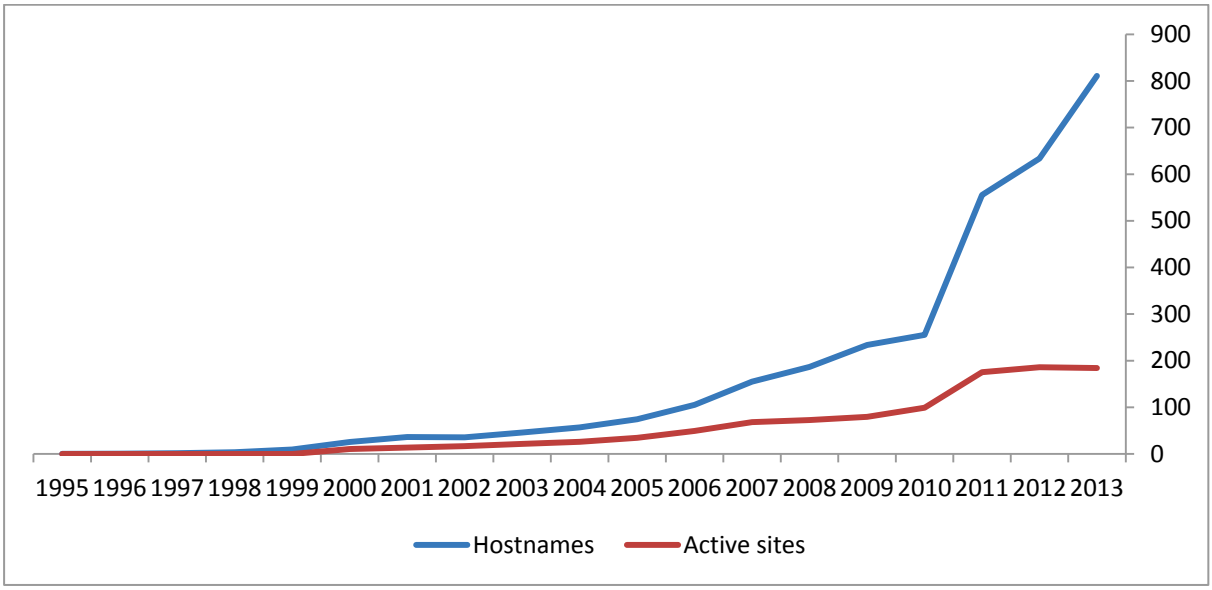
The total volume of recorded data, including both published and unpublished data, is estimated to be doubling every two years.[17] At this rate of growth, the volume of data recorded in 2025 will be more than a thousand times greater than that in 2005. This growth is driven by the rapidly expanding capacity of both computing and communications networks and devices, and by technical developments such as cloud computing, which allows very large data volumes to be hosted and accessed through clusters of data centres rather than requiring storage on users' own devices. Greater network capacity, including the growth of broadband networks, has enabled tremendous growth in the volume of video data downloaded or streamed across the Internet. By 2012, Cisco estimated, video already accounted for 57 per cent of Internet traffic by volume, excluding peer-to-peer file-sharing; with that included, Cisco expected the figure to rise above 80 per cent by 2017 (Cisco, 2013).

Data volumes will increase further as a result of the *Internet of things* (IoT), which will make many objects, as well as people and organisations, active participants in data generation. As with earlier phases of ICT development, the *Internet of things* is likely to be adopted earlier in developed

countries than developing countries, because their communications networks generally have higher specifications and because users in developed countries have more financial resources to buy IoT-enabled devices.

The number of websites and webpages has become increasingly difficult to measure since 2005, especially as search engines no longer crawl the entire web when compiling search results. Netcraft's January 2014 website survey identified over 850 million hostnames and approximately 185 million active sites.[18] Given the difficulties of identifying the number of websites overall, or in the indexable web, trends in website growth may be more useful. The growth of websites since 1995, as assessed by Netcraft, is illustrated in Chart 9.1.

Chart 9.1: Netcraft estimates, total websites, millions of hostnames and active sites, 1995–2013



Source: adapted from data in Netcraft, January 2014 Web Server Survey, data from December of each year, http://news.netcraft.com/archives/2014/01/03/january-2014-web-server-survey.html, accessed 6 March 2014.

Some information concerning the popularity and growth of specific web services is included later in this chapter.

The nature of published online information has diversified considerably since WSIS as a result of the growth in network and device capacity and the emergence of new services, particularly those associated with user-generated content, transactions, and audio and video content enabled by much greater bandwidth.

Social media websites have largely emerged since WSIS and now form an important part of the Internet experience for most users, in developing as well as developed countries. A number of specific social media sites, including those identified below, are now among the most accessed websites in a majority of countries where Internet activity is regularly measured. However, local alternatives are more significant in some markets, notably China.

- The number of monthly active users of Facebook, the leading international social network, which was founded in 2004, has grown from almost zero at the time of WSIS to 1.2 billion in 2013. Although the two are not directly comparable, as some Facebook accounts are held by organisations and individuals may have more than one Facebook account, this figure is equivalent to 45 per cent of that for individuals using the Internet worldwide, as estimated by ITU.[19]

- The number of monthly active subscriptions to Twitter, the leading international microblogging service, which was founded in 2007, has grown to 241 million in 2013. About 500 million tweets were posted daily by the end of 2013.[20] A further 507 million people subscribed to the Chinese microblog service Tencent Weibo, out of an estimated 582 million Chinese Internet subscribers.[21]
- By the end of 2013, as many as 100 hours of video content were reported as being uploaded every minute to YouTube, the leading international video-file sharing site, with more than six billion hours of video being watched each month.[22]

Content on user-generated sites such as these now amplifies the total volume of content available. Much of this can be regarded as local content, in that it is generated by individual users and most likely to be accessed by others within their geographical, occupational or personal communities. User-generated content also offers more scope for publication in local languages. However, data on social media usage are limited because of commercial confidentiality. Data availability is discussed further elsewhere in this chapter.

Another important development in content since WSIS has been the growth in open data, that is, the publication of data gathered and analysed by governments and other public bodies so that this can be used by citizens and third party organisations as well as by official analysts. The growth of open data has been driven by freedom of information legislation, and is more advanced in developed than developing countries, though significant steps have also been taken by a number of developing country governments.[23] Although the total volume of information made available in this way is small compared with the growth of social media and video content, it has particular relevance to the developmental outcomes sought by WSIS.

Language

There have been four significant changes in the relationship between the Internet and language since WSIS.

First, the language in which computers and the Internet are developed has a significant bearing on the development of a multilingual Internet. Computer code and programming were initially dominated by the English language, which influenced the evolution of online systems, for example the use of Latin characters and the ASCII[24] character set in the domain name system. User software was also initially concentrated on a small number of European languages, but has since diversified. The operating system Windows 2000, for example, supported 16 languages, but 45 were available in Windows 2003 and 95 language packs are listed as being available for Windows 7. The number of language packs available for Microsoft Office has likewise grown from 33 in Office 2003 to 65 in Office 2013.[25] Internet browsers are essential portals for users of the WWW, and they too have become more multilingual. Internet Explorer now offers 119 languages, including a number in different dialect forms, while Firefox offers 107 and Google Chrome 117.[26]

Secondly, the widespread use of social media and user-generated platforms on the Internet (and of mobile apps) means that a much higher proportion of content is now generated by individuals, often for small user groups with shared characteristics. Major social media sites support a wide range of languages, enabling individuals to generate content in the language of their choice, subject to the technical limitations of the devices they are using. By December 2013, for example, it was reported that 34 per cent of Twitter tweets were in English, with the next most popular languages being Japanese (16 per cent), Spanish (12 per cent) and Malay (8 per cent).[27] There is anecdotal evidence of

linguistic adaptation in some languages, for example of users adopting Latin characters when communicating in languages that have different character sets such as Russian and Greek.

The third important development has been the introduction of internationalised domain names (IDNs). A major constraint on the domain name system, until 2010, was that only Latin characters within the standard ASCII character set could be used in top level domains (TLDs), preventing the provision of these in languages using non-Latin scripts. ICANN approved procedures that implement top level IDNs by proxying non-Latin scripts against ASCII characters in 2009, and the first IDN TLDs became available during 2010. UNESCO and the European domain registry EURid reported in 2013 that just over 5 million IDNs had been allocated globally, though these represented less than 2 per cent of domains in use worldwide, while more than 90 per cent of the most popular websites did not yet recognise IDNs in URL links (UNESCO and EURid 2013). More information about the development of IDNs can be found in the third part of the chapter.

The fourth development of significance for language on the Internet concerns automated translation. The enormous quantity of content that is available online cannot be translated manually into all languages, or indeed into any specific language – nor is there significant demand for the translation of much online content, for example tweets or Weibo posts. Automated translation programmes offer the only realistic way of enabling translation that responds to demand from online users. The first automated translation programmes were developed in the 1950s; there has been considerable improvement in their performance since the emergence of the Internet, though problems of quality assurance and reliability remain significant, while reliable translation is least available for languages with limited user numbers. The most widely used online translation service, Google Translate, is currently available in 80 languages.[28] Further discussion of automated translation can be found later in the chapter.

Data availability and scope

The Partnership on Measuring ICT for Development adopted five indicators for Target 9 in its 2011 WSIS statistical framework. The indicators are as follows:

> Indicator 9.1: The proportion of Internet users by language, country level
>
> Indicator 9.2: The proportion of Internet users by language, top ten languages, global level
>
> Indicator 9.3: The proportion of webpages, by language
>
> Indicator 9.4: The number of domain name registrations for each country code top level domain (ccTLD), weighted by population
>
> Indicator 9.5: The number and share of Wikipedia articles by language.

This section of the chapter considers the appropriateness and measurability of Target 9; discusses the overall scope and suitability of the five indicators currently selected; summarizes the availability of data envisaged for these indicators in the 2011 WSIS statistical framework; and recommends ways in which, should the target be retained, the portfolio of indicators might be adapted for future monitoring and measurement. Available data for individual indicators are discussed in more detail below.

Access to content and the skills to make use of it, including language, are critically important to the emergence of an inclusive information society. There is, therefore, a powerful case for including measurement of them in the monitoring and measurement of WSIS outcomes. However, the principal challenge of Target 9 concerns the difficulty of defining quantifiable goals against which progress can be measured. There is no obvious upper limit for either content or linguistic diversity. The supply of content – the volume of online information – is growing extremely rapidly, while the range of platforms through which content can be accessed is also broadening. Measures of content that are based around particular forms of content (such as webpages), particular media (such as the Internet) or particular platforms (such as mobile apps) are likely to remain relevant only for short periods of time. One of the principal factors determining linguistic diversity in future is automated translation, the extent and impact of which will also be very difficult to measure.

These factors illustrate the challenges involved in identifying appropriate ways of assessing content and language within the WSIS outcome framework.

The purpose of measuring content and language, like other WSIS targets, is essentially twofold:

- to monitor progress towards the development of an information society, in which there is universal access to the networks, services and content that are required by people, whatever their needs, wherever they live and
- to identify constraints and limitations on this development that can be addressed by governments and other stakeholders in order to accelerate progress and reduce the digital divide.

To have value, indicators for monitoring and measuring progress must enable:

- comparisons to be made at a particular point in time between circumstances in different countries and other relevant categories, such as gender and language groups and
- trends over time to be measured both globally and in individual countries and language groups, also allowing disaggregation where possible by gender, age, disability and other demographic categories.

Data for these indicators must also be relatively easy to gather in the wide range of national contexts concerned, and should be accurate, reliable and timely.

The portfolio of indicators that was selected in 2011 has three main limitations when set against these criteria.

The first concerns the scope of the indicators chosen in relation to the target as a whole. Although the target is concerned with both content and language, the five selected indicators focus predominantly on language. Only one indicator – 9.4 – is concerned primarily with the availability of content, and it is concerned only with the supply side of content (content creation). None of the indicators in the 2011 framework is concerned with the demand side of content (access and use).

The second limitation concerns data availability. Very few data are available for three of the selected indicators (9.1, 9.2 and 9.3), with the result that it is not currently possible to use these as effective measures of either content or language. Reasons for this are discussed below. Data relating to indicator 9.4 are generated by domain name registries and other Internet entities, though these are not widely published or collected. Information derived from historic data sets relating to this indicator has been generously made available and collated for this report through the cooperation

and support of a specialist research consultancy, ZookNIC. Data are readily available for indicator 9.5, thanks to the transparent publication of data-sets by the Wikimedia Foundation and community.

The third limitation is that the selected indicators do not include measurement of the very significant developments in content that have taken place since 2003, especially online social media and mobile apps. Only indicator 9.5 derives data from a social media platform, Wikipedia, but this is an unusual social media platform because content creation in its case involves a much smaller group of people than those that access content. The burgeoning significance of social networks, microblogs and audio and video file-sharing sites in content creation and access is therefore inadequately included in the current portfolio of indicators for this target, distorting the overall picture of content creation and access emerging from assessment over the period since WSIS. The same point can be made concerning mobile apps.

Available data concerning content and language on these new media platforms are very limited. Most popular social media platforms are offered by Online Service Providers (OSPs) free to end-users through a business model that uses data-mining techniques in order to target advertising. While OSPs themselves have extensive data on the geographic, linguistic and other characteristics of content carried on their services, these data have considerable commercial value, not least for targeted advertising and for 'big data' analysis. Little information from them is made publicly available because of their high commercial value and commercial confidentiality. The Wikimedia Foundation, which is a non-commercial enterprise, is an exception to this model and publishes extensive data that are used in the assessment of indicator 9.5 below.

The implications of these challenges for each of the five selected indicators are discussed below. In summary, this section recommends that, if the present target is to be retained:

- Indicator 9.1 should be retained, but suspended until data of sufficient quality become more comprehensively available as a result of national statistical offices incorporating relevant data collection into national censuses and household surveys.
- Indicators 9.2 and 9.3 should be withdrawn as it is not currently possible to obtain reliable data, and unlikely that this situation will change at least in the short or medium term.
- Indicator 9.4 should be retained in revised form, including gTLDs and IDNs as well as ccTLDs in national counts of domain names, and subject to mechanisms being put in place to secure access to comprehensive data sets from either national registries or independent analysts.
- Indicator 9.5 should be retained but developed to include Wikipedia contributors (content creation) and page views (access and use) as well as articles.
- Additional indicators should be developed to replace indicators 9.2 and 9.3. These should be concerned with measuring the volume and linguistic diversity of content on one or more social networks and on mobile apps.

This final recommendation recognises that the pace of change in available platforms since WSIS has been such that an emphasis on webpages is no longer sufficient to measure online content. It is important to recognise that the emergence of further new platforms for content creation, dissemination and access could also render the indicators recommended here outdated during the next decade.

An alternative or supplement to quantitative monitoring and measurement, of the kind envisaged in the 2011 WSIS statistical framework, would be to gather a wider variety of quantitative and

qualitative data on a specified number of countries and territories that are selected to be representative of the world community. While this would not have the same statistical value as monitoring of other WSIS targets, it would enable a more substantive qualitative assessment to be made of trends that are taking place in content and language, alongside those statistical indicators that do prove to be viable. Additional statistical evidence from diverse sources could also be incorporated in monitoring and measurement.

The periodic publication of time series data in tables and charts is only one way of illustrating the spread of online content and language. A number of research institutes and other data analysts have developed considerable expertise in the use of mapping techniques to illustrate trends in online activity, including content and language.[29] Consideration could be given to the potential of techniques such as these for adding insight to those data that are available in this area of WSIS outcomes.

Achievements against Target 9

This section of the chapter considers data availability for the five indicators for Target 9 in more detail, and summarises findings and achievements that can be derived from available data. It supplements these findings with further information, including additional evidence related to these indicators and additional sources concerning social media. The discussion in this section also draws attention to evidence from five selected countries: Brazil, India, Indonesia, Kenya and South Africa. Consideration of each indicator includes a brief assessment of its appropriateness for future monitoring and assessment of Target 9, in the context of the discussion above.

Proportion of Internet users by language, country level

This is measured by Indicator 9.1, which is a measure of the use of the Internet by individuals, classified by language, within each country.[30] Use of the Internet, for this indicator, was intended to include use by an individual within a twelve month period from any location, using any device (including mobile devices). The intention was to calculate this indicator in two forms:

- the proportion of speakers of each language in each country who are using the Internet and
- the proportion of Internet users in each country who are speakers of a particular language.

The attribution of language established for this indicator was to be the ‘usual language’ or ‘mother tongue’ for each individual, as identified in national census or household survey data. The UN Statistical Division defines usual language to mean "… the language currently spoken, or most often spoken, by the individual in his or her present home" and mother tongue to mean "… the language currently spoken in the individual’s home in his or her early childhood".[31] The problems associated with allocating a single language to an individual in either of these ways have been discussed above.

This indicator was developed on the basis of one of the set of ICT core indicators that were agreed by the *Partnership* following WSIS and, which it was hoped, would be included in national population censuses and household surveys conducted by national statistical offices (*Partnership*, 2010). Core indicator HH7 in this set sought to establish the proportion of individuals who had used the Internet in the previous twelve months (since adjusted to three months, see ITU, 2014). The intention was to assess findings from this indicator alongside data concerning individuals’ mother tongues or language preferences collected in the same censuses and surveys.

Data for this indicator are not, therefore, available unless they have been collected in national censuses or household surveys. A model questionnaire and notes for the collection of relevant data in such surveys have been published by the *Partnership*. In 2010, the *Partnership* reported that only 35 developing countries were then collecting data concerning Internet usage (HH7), but no analysis was made of the number of these countries that also collected language data or of analyses of survey outcomes that juxtaposed Internet and language findings.[32] A selective review of national census forms confirms that there is to date limited adoption of the range of questions that could enable assessment as envisaged in the 2011 WSIS statistical framework.[33]

This indicator could have value for assessing access to content by language if this situation changes in the future. It should therefore be retained, but suspended until data of sufficient quality become more comprehensively available as a result of national statistical offices incorporating relevant data collection into national censuses and household surveys.

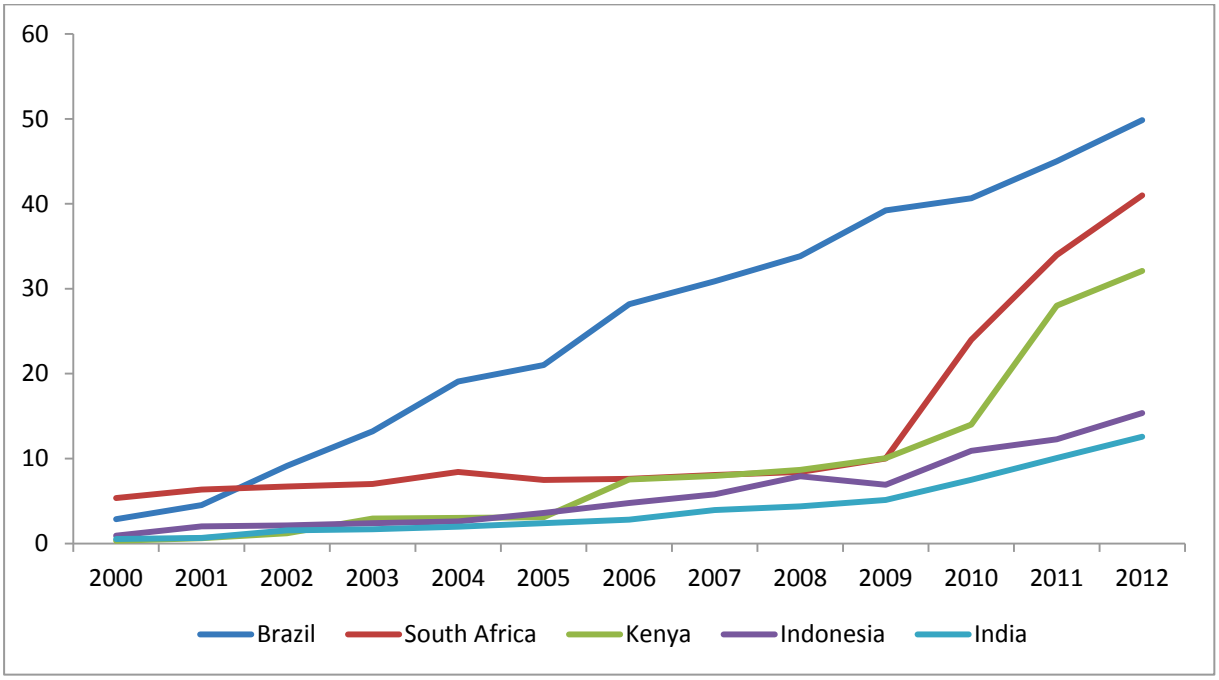
Findings

No reliable data are available for this indicator at present from the sources identified in the 2011 WSIS statistical framework.

Estimates for the overall adoption and use of the Internet in different countries are compiled by the ITU. These are reported in other chapters of this report, which are concerned with access to, and use of, ICT.

Chart 9.2 illustrates the growth in the percentage of Internet users for the five countries that were selected as example countries for this chapter, using ITU estimates. It shows the variable growth rates in Internet use that have been experienced in different developing countries. These have consequential impacts on access to content and the development of content in different languages.

Chart 9.2: Individuals using the Internet, 2000–2012, percentage



Source: ITU World Telecommunication/ICT Indicators Database.

Supplementary data that shed light on Internet adoption and use in different countries are available from some other sources, and these may begin to shed more light on access by language groups. Household surveys were conducted, for example, by the research institute Research ICT Africa (RIA) in some twelve African countries during 2011. These included a question on main household language and extensive questions about language preferences online, as well as more general Internet use.[34] Data from surveys such as this should in time shed further light on variations in Internet access by language group in selected countries at the time that they were taken.

The proportion of Internet users by language, top ten languages, global level

This is measured by Indicator 9.2, which concerns the proportion of Internet users accessing content through the top ten languages in global use (identified in the 2011 WSIS statistical framework as being (in alphabetical order) Arabic, Chinese, English, French, German, Japanese, Korean, Portuguese, Russian and Spanish).[35] It was hoped that this could be reported in two forms:

- the proportion of worldwide speakers of each language who use the Internet and
- the proportion of global Internet users distributed by language.

Some estimates of the proportion of Internet users falling within different language groups have been made at different stages in the development of the Internet. The source identified for this indicator in the 2011 WSIS statistical framework was the Internet data website Internet World Statistics (IWS), which published a calculation for this indicator using data from 2009. These data were derived from a number of sources including ITU and the US Bureau of the Census. IWS has since published an updated tabulation using estimates for Internet use and population for 2011.

As discussed earlier, there are significant statistical challenges involved in estimating the number of speakers using global languages. In making its calculation, IWS states that it allocated a single language to each individual, excluding secondary languages, though it is unclear what methodology was used to select this language in countries where a high proportion of the population is bilingual or multilingual. Other statistical challenges to which IWS drew attention included the need to make adjustments in the data for variable rates of infancy and illiteracy. These challenges, particularly those concerned with language attribution, are sufficient to make it inadvisable to draw strong conclusions from this indicator.

As things stand, there is no realistic prospect of sufficient data becoming available to allow substantive findings to emerge from indicator 9.2. It does not therefore provide a suitable basis for future assessment of Target 9 in the context of the discussion above, and should be discontinued.

Findings

As noted above, some earlier estimates for this indicator were made before IWS published its calculation in 2009. The five-year review of this target, published in 2010, suggested that the proportion of English speakers online in 1996 was as high as 80 per cent.[36] Subsequent estimates were made by Globalstat until around 2005, which found that the proportion of English speakers had fallen to 35 per cent by September 2004.[37]

The 2011 estimates published by Internet World Statistics for this indicator are set out in Table 9.3. Chart 9.3 juxtaposes these 2011 data above with findings from the only Globalstat report, from September 2004, which is still available online. It is unclear how far these two data sets are

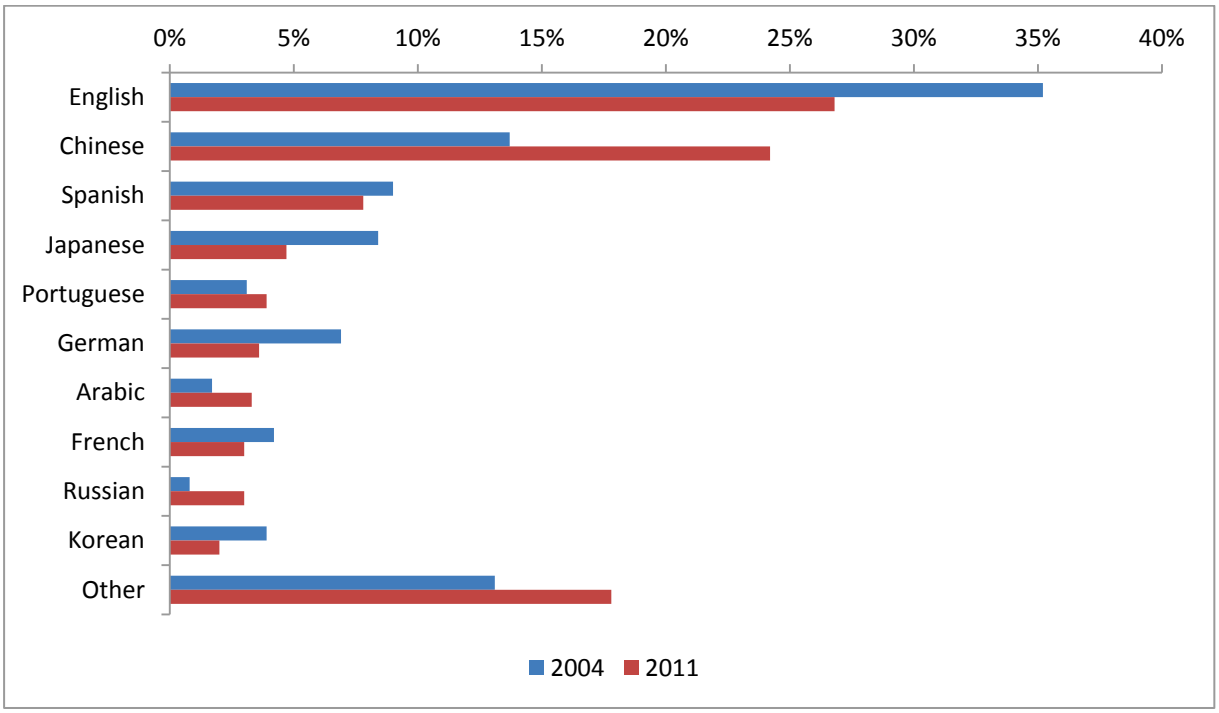
compatible in terms of sourcing and methodology,[38] but they may provide a rough guide to shifts in the proportion of language speakers online. They suggest a decline during the period in the proportion of Internet users who are speakers of English, other European languages and languages from highly developed Asian economies (including Japanese and Korean), and a corresponding growth in the proportion speaking Russian, Chinese and other languages.

Table 9.3: Global Internet users by main global languages, IWS estimates, 2011

Language	Language speakers	Internet users	Internet penetration by language	Language users as percentage of total Internet users
English	1,302,275,670	565,004,126	43%	27%
Chinese	1,372,226,042	509,965,013	37%	24%
Spanish	423,085,806	164,968,742	39%	8%
Japanese	126,475,664	99,182,000	78%	5%
Portuguese	253,947,594	82,586,600	33%	4%
German	94,842,656	75,422,674	80%	4%
Arabic	347,002,991	65,365,400	19%	3%
French	347,932,305	59,779,525	17%	3%
Russian	139,390,205	59,700,000	43%	3%
Korean	71,393,343	39,440,000	55%	2%
Other	2,403,553,891	350,557,483	15%	18%
Total	**6,930,055,154**	**2,099,926,965**	**30%**	**100%**

Source: Internet World Statistics, http://www.internetworldstats.com/stats7.htm.

Chart 9.3: Estimated online language populations, percentage of global online population



Source: Internet World Statistics, http://www.internetworldstats.com/stats7.htm; Globalstat, http://web.archive.org/web/20041019013615/www.global-reach.biz/globstats/index.php3.

The proportion of webpages by language

This is measured by Indicator 9.3, which concerns the proportion of webpages accessible on the WWW that are available in different languages. This was intended to act as a proxy for online content creation by language.[39]

The 2011 WSIS statistical framework recognised that this indicator would be very difficult and expensive to measure because of the size and continued growth of the WWW, and that analysis would require enormous computing resources. Analysts today regard it as highly problematic to estimate the total number of webpages on the Internet, particularly since the volume of the web became so large that search engines ceased to index it as a whole. Netcraft estimated the size of the web in January 2014 at over 850 million hostnames and about 185 million active sites.[40] It was still estimated in early 2014 that 56 per cent of the top ten million websites used English as at least one of their content languages (though this does not mean that English was necessarily the site's predominant language).[41] While broadly representing the overall scale of accessible web content, such numbers can be misleading: for example, they include many pages that do not contain substantive content, while also under-representing the growth in user-generated content such as that on microblogs.

Measurement of this indicator, as defined, would require the systematic and comprehensive use of two complex and expensive sources and methodologies:

- web-crawling programmes, which comprehensively browse the WWW for indexing purposes and
- script and language identification programmes, which can analyse web content to establish the language in which it is written (bearing in mind that an unknown proportion of websites are themselves multilingual).

In the period between WSIS and adoption of this indicator, some relevant data were compiled by an international academic consortium, the Language Observatory Project (LOP).[42] This used a web crawler to search top level domains in a limited range of countries in Africa and Asia. Because of the high resource and financial costs involved, the project was confined to smaller countries, with relatively low content volumes, and to content hosted on ccTLDs, which may be differently distributed by language from that hosted on gTLDs (see Indicator 9.4 below).

However, this project has not reported data since 2007 and does not therefore provide a viable source for this indicator today. While a number of other projects have sought funding for similar work, these have not been successful to date.[43] Therefore, as with indicators 9.1 and 9.2, there are no substantive data available with which to assess progress on this indicator. It would not be feasible for the *Partnership* itself to initiate monitoring of this indicator because of the very high costs involved.

As with indicator 9.2, there is at present no realistic prospect of sufficient data becoming available to allow substantive findings to arise from Indicator 9.3, and it should therefore be discontinued. It may be possible in future to develop an approach that combines random sampling of webpages through a web crawling programme together with language recognition software. For this to be statistically valid, large samples of webpages would be required, and the technique would be expensive. It is therefore only likely to be viable if an independent research institute obtains funding to undertake the work.

A potential alternative source of data concerning content and language across the WWW would be search engines, whose business consists of indexing the web and facilitating access to specific content by end users. These no longer index the entirety of the web, as this is now too large for comprehensive indexing to be effective, but they are likely to have more credible evidence concerning this indicator than other potential sources. However, search engines are commercial businesses. Data concerning content, language and search activity represent an important part of their business model, and are not therefore available for external analysis.

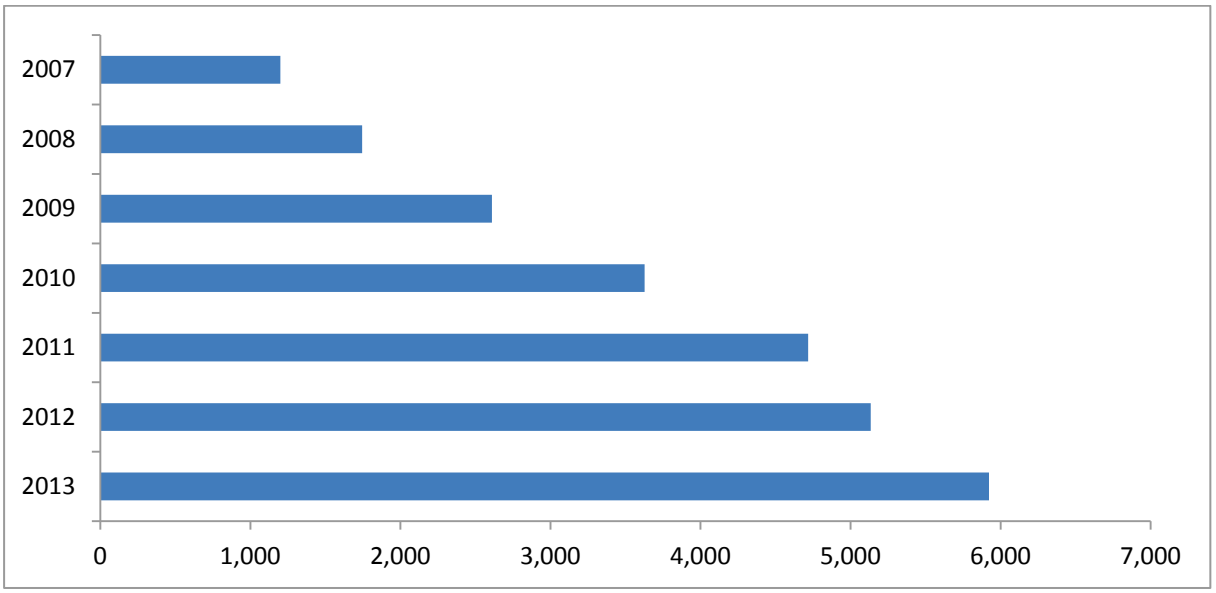
Findings

There are no satisfactory data that can be derived from the source originally anticipated for this indicator, and no substantive alternative data are publicly available.

Some estimates have been made of web content by language in particular regions. In 2013, the UN Regional Commission for Western Asia published data concerning the proportion of websites registered to ccTLD addresses from different Arabic-speaking countries that were in Arabic or English. These data need to be interpreted carefully as only a small proportion of domain registrations in Arabic-speaking countries are ccTLDs (18 per cent in 2013[44]), and because French rather than English is the predominant secondary language in some of them. However, among countries in which English is the prevalent secondary language, the proportion of webpages in Arabic from ccTLD domains varied from less than 17 per cent in Lebanon to more than 50 per cent in Saudi Arabia and Sudan.[45]

Some data have been published on the number of searches that are made using Google search, the leading search engine in most national markets, often with over 90 per cent of the search market, though not (according to 2010 data) predominant in China, Russia, Japan or the Republic of Korea.[46] The number of searches made through Google in 2013 exceeded 2 trillion, amounting to more than 5 billion searches daily. Growth in the number of Google searches daily is illustrated in Chart 9.4.

Chart 9.4: Growth in daily Google searches, 2007–2013, millions of searches



Source: http://www.statisticbrain.com/google-searches/.

With the caveats above concerning particular countries in mind, this can be taken as a useful proxy of the growth of Internet activity and content access in most countries. However, it has not been

possible to obtain more detailed information disaggregating search data by country of origin or language.

The number of domain name registrations for each country code top level domain (ccTLD), weighted by population[47]

This is measured by Indicator 9.4. The 2011 WSIS statistical framework selected the number of ccTLD registrations, weighted by population, as a proxy indicator for the amount of content created within a country. This is the only indicator selected within Target 9 that focuses on content by country rather than by language.[48]

Each online content publisher requires at least one Internet domain, which provides it with a Unique Resource Locator (URL, for example, www.itu.int) on the WWW. The Internet domain name system (DNS), which allocates and manages domains, includes two main types of top level domain:

- Country code top level domains (ccTLDs), such as .uk in the United Kingdom and .za in South Africa, are administered by national ccTLD registries. There is one registry per country code.
- Global top level domains (gTLDs) are administered by a number of international businesses and organisations. The large majority of gTLD registrations are currently for .com (commercial) domains, with a further four gTLDs (.net, .org, .info and .biz) accounting for most of the remainder. The number of different gTLDs available is being greatly expanded following agreement in the Internet Corporation for Assigned Names and Numbers (ICANN), which oversees the domain name system. A substantial number of new gTLDs will therefore enter the market in the near future, some of which are likely to compete significantly with ccTLDs.

Internet users may obtain a relevant domain that is registered by either a gTLD registry (such as Verisign for .com or the Internet Society's Public Interest Registry for .org), or a ccTLD registry (such as Nominet for .uk or ZADNA for .za).[49] Many of these fall within subdomains that identify the type of content publisher involved (for example, .ac.uk for an academic entity in the UK; .co.za for a business in South Africa). These subdomains are administered through the relevant top level registry. There are also a small number of sponsored top level domains (sTLDs) that are used by particular communities, such as .aero (reserved for aviation).

Until 2010, top level domains were only available in Latin characters (though it was possible before then to obtain domains that used non-Latin characters in earlier parts of the domain name). Since 2010, ICANN has authorised a number of top level domains that use non-Latin characters (Internationalised Domain Names or IDNs). Languages that have seen significant use of these top level IDNs include Arabic, Chinese, Korean, Persian, Russian, Thai and a variety of languages in South Asia.

As at March 2014, there were approximately 148 million gTLD registrations worldwide (including a little over 1 million sTLDs), representing 54 per cent of the global market, alongside approximately 125 million ccTLD registrations (including a little over 1 million ccTLD IDNs).[50]

Domain registrations are a relatively good proxy for the number of publishers generating content on the Internet within a country because they have to be unique to a particular content source. No two domain names can be identical. The primary source of information for ccTLD registrations is the national ccTLD registry, of which there is generally[51] only one per country. Some, but not many, registries publish data on the number of registrations within their national domain (including

subdomains). Where these data are not published, it may be possible to obtain them through direct enquiry or through the use of network utility tools. Most national registries belong to one of four regional associations, some of which also publish data across their regions.[52]

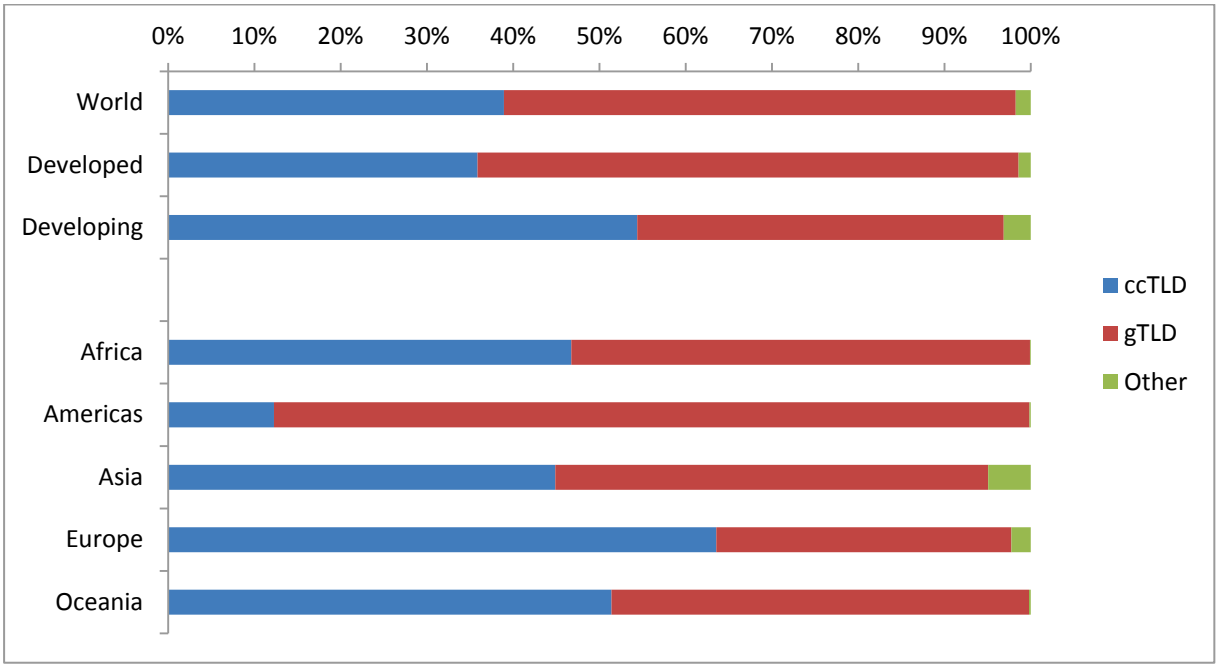
While ccTLD registrations appear at first sight to be a good proxy for local content generation, they have a number of serious limitations and need to be used with caution.

Internet users within a country can choose either a gTLD or a ccTLD registration (or one or more of each). The ratio between gTLD and ccTLD registrations in different countries varies considerably as a result of a number of factors, including the relative cost of registration, the relative complexity and time required for registration processes, and the brand value associated with different national domains. In some countries, for example, the cost of a ccTLD registration is much higher than that for a gTLD. In some, a ccTLD registration may be viewed more positively by a business's customers, because it represents local identity, while in others a gTLD registration may be preferred because it appears to represent global reach and scale. The introduction of many new gTLDs over the next few years could also significantly affect the balance between gTLD and ccTLD registrations. Many of the new gTLDs are designed to appeal to economic, social and cultural sectors or identities, and so offer an alternative option to the geographic branding offered by ccTLDs.

Fortunately, it is possible to use geolocation techniques to identify the country of origin of gTLD registrations, and current data concerning this are published by the Internet domain search database WHOIS.[53]

The balance between gTLD and ccTLD registrations in world regions in December 2013 is illustrated in Chart 9.5. (Data for this Chart include the six leading gTLDs,[54] which account for over 99 per cent of gTLD domains. They exclude 15 ccTLDs that function as virtual gTLDs (see below). IDNs are included as 'Other domains' rather than as ccTLDs or gTLDs.)[55]

Chart 9.5: The balance between gTLD and ccTLD registrations in world regions, 2013



Source: Data supplied by ZookNIC, compiled from ccTLD, Whois and other sources.

Chart 9.5 shows that ccTLD registrations are a minority of total registrations worldwide. In addition, there are considerable variations in the ratio between gTLD and ccTLD registrations in different world regions. Country code registrations are least common in the Americas, because gTLDs have always been the norm in the United States. They are also particularly uncommon in the Arabic-speaking region. Country code top level domain (ccTLD) registrations make up a significant majority of registrations in Europe and are also particularly common in the CIS region. This degree of variation makes ccTLDs alone an unreliable proxy for local content.

The overall balance between gTLD and ccTLD registrations has been relatively stable over the period since WSIS. The proportion of ccTLDs among total registrations worldwide was between 38 per cent and 39 per cent in each of the three years assessed for this report (2003, 2008 and 2013), though there has been a small increase in the proportion of ccTLD registrations in developed countries since 2008 and a more significant decrease (from 60 per cent to 54 per cent) in the proportion in developing countries.[56]

A further complication arises from the fact that not all ccTLD registrations represent local content or local registrants. Some, but not all, ccTLD registries accept registrations from non-domestic users. Some 15 ccTLDs with suitable domain extensions have been marketed or used as, in effect, virtual gTLDs. Examples of these include .me (Montenegro; used for personal websites), .co (Colombia, used as an alternative to .com for business registrations), .nu (Niue; used for various meanings in different languages), .tv and .fm (Tuvalu and the Federated States of Micronesia, both used by broadcasters).[57] The most extreme example of a ccTLD outreaching its domestic market is .tk, the ccTLD for the New Zealand dependency of Tokelau that has some 1 200 inhabitants but makes domain registration available free of charge and was responsible for more than 20 million ccTLD registrations by December 2013 – more than any gTLD other than .com, and 5 million more than the next highest ccTLD, Germany.[58] The 15 ccTLDs that act as virtual gTLDs have been excluded from the analysis in this section of the chapter, leaving a total of approximately 96 million ccTLDs under discussion.

Large businesses and other organisations often have multiple registrations. Global businesses such as Google and Amazon, for example, make use of ccTLDs in many countries as well as their global .com domains (and are in the process of setting up new gTLDs with their own identities). Google can be accessed through some 200 national domains as well as through google.com.[59] Some traditional businesses and organisations also procure a number of domains from both gTLDs and ccTLDs, so that they can visibly provide tailored services in particular countries and also in order to protect trademarks, brand identities etc. These multiple registrations may be more common in larger than in smaller countries because of the greater number of large businesses and organisations in those countries.

The factors discussed above suggest that, rather than using ccTLD registrations alone as a proxy for local content creation, it would be preferable to include both ccTLD registrations and gTLD domains that are registered from within the same national territory, as identified through geolocation. If indicator 9.4 is to be retained, therefore, it should be revised to include both ccTLD and gTLD data, and to incorporate IDNs.

Data required for this indicator are difficult to collate, particularly longitudinal data. The analysis below would not have been possible without the assistance of the consultancy ZookNIC, which has maintained historic data on registrations throughout the period since WSIS. The future viability of this indicator is dependent on access to comparable data being available.

Findings

Domain name data for both gTLDs and ccTLDs are compiled regularly by the consultancy ZookNIC.[60] These data are derived from a number of sources, including published ccTLD registry reports, analysis of TLD root zone files, network utility tools and direct correspondence with registry operators. Historic data from these sources illustrate the growth of registrations within each country or territory over time. The following paragraphs present findings concerning both ccTLD registrations (the existing indicator) and total registrations (ccTLD plus gTLD and other registrations) in different countries and territories over the period since WSIS.

The data and analysis in this section have been prepared in collaboration with ZookNIC, using data from its comprehensive database of domain name registrations, which have been generously made available for this purpose. As indicated above, the data raise a number of interpretation challenges. In addition to the balance between gTLD and ccTLD registrations and the incidence of virtual gTLDs, there are definitional differences between registries concerning what to include in domain counts – for example, some registries may not include inactive domains. A number of registry policies – such as pricing, limits on the number of domains a single entity can register, and identification or residency requirements – also complicate direct comparisons between registries.

It is clear, too, that content made available through local domains is not necessarily local in nature, while a good deal of content that is local in nature is generated on social media sites and therefore not reflected in domain counts. However, the data summarised below, particularly those for ccTLD and gTLD registrations together, do represent a worthwhile proxy for web content that is generated by country and add significantly to our knowledge of relevant trends.

Table 9.4 shows the overall numbers and proportions of registrations by region – and for developed and developing countries as defined by ITU – for each of the three years 2003, 2008 and 2013.

Table 9.4: Total registrations by world region, 2003–2013

	2003		2008		2013	
	Millions	Percentage	Millions	Percentage	Millions	Percentage
World	59.7	100%	173.4	100%	245.2	100%
Developed	49.6	83%	135.9	78%	197.4	81%
Developing	7.1	12%	34.7	20%	45.0	18%
Other/Unknown	3.1	5%	2.8	2%	2.7	1%
Africa	0.3	0.5%	1.0	0.6%	2.3	0.9%
Americas	23.9	40%	71.8	41%	98.9	40%
Asia	5.3	9%	29.8	17%	36.9	15%
Europe	25.8	43%	63.7	37%	98.0	40%
Oceania	1.2	2%	4.2	2%	6.4	3%

Source: Data supplied by ZookNIC, compiled from ccTLD and other sources (see above).

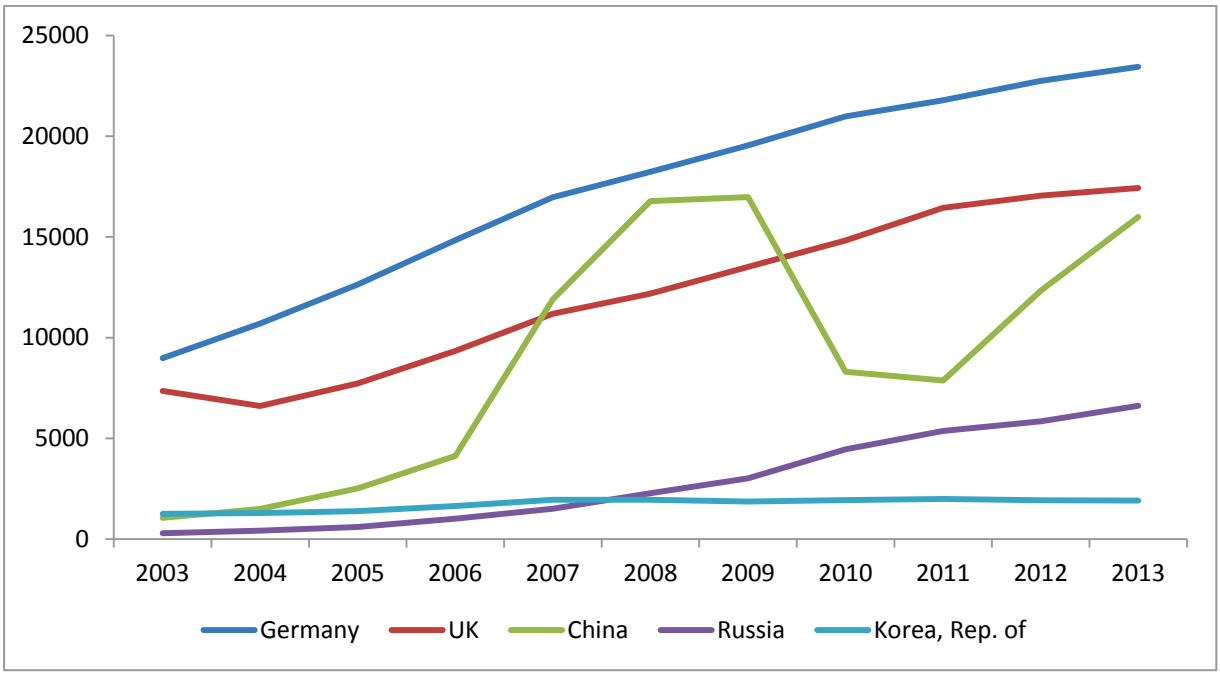
Note: Figures exclude 15 ccTLDs that act as virtual gTLDs.

This table shows that the Internet has continued to be dominated by content providers in Europe and the Americas throughout this period. The proportions of domain registrations from these two continents are very substantially greater than that from Asia, which has a substantially higher

population. The proportion of registrations from Africa remains below 1 per cent, while the continent has a little less than 15 per cent of world population.

Charts 9.6 and 9.7 show the trend in the development of ccTLD and total registrations for selected countries – including five leading Internet user countries (Chart 9.6) and the five developing countries selected for review in this chapter (Chart 9.7) – year-on-year since 2003.

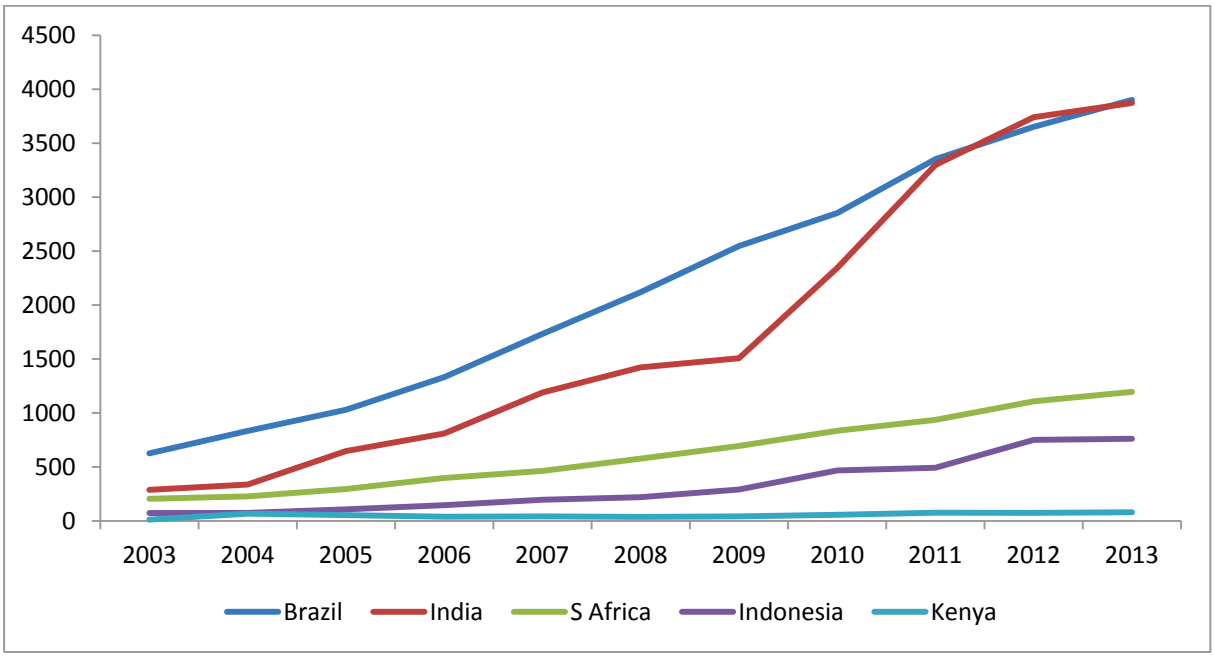
Chart 9.6: Total domain registrations (thousands), 2003–2013, leading Internet countries



Source: Data supplied by ZookNIC, compiled from ccTLD, Whois and other sources (see below).

Note: Figures exclude 15 ccTLDs that act as virtual gTLDs.

Chart 9.7: Total domain registrations (thousands), 2003–2013, developing countries



Source: Data supplied by ZookNIC, compiled from ccTLD, Whois and other sources (see below).

Note: Figures exclude 15 ccTLDs that act as virtual gTLDs.

These illustrate steady rates of growth in registrations, with more rapid growth in Brazil and India than in other developing countries illustrated. The number of registrations in Kenya is still low, but has also grown steadily, from 12 000 in 2003 to 80 000 in 2013. The rapid growth, decline and return to growth in China, illustrated in Chart 9.6, resulted from a period of aggressive price competition for registrations in the middle years of the decade, followed by a return to more normal registration pricing.[61]

Indicator 9.4, which was adopted in the 2011 WSIS statistical framework, proposed to measure the number of ccTLD registrations per head of population. This indicator can also be calculated as a proportion of the number of Internet users in a country as estimated by ITU. The same calculations can be made for total registrations, including both ccTLDs and gTLDs. Table 9.5 sets out the numbers of people per ccTLD and per TLD registration in world regions for the three years 2003, 2008 and 2013, while Table 9.6 sets out the numbers of Internet users per ccTLD and per TLD for the same three years.

Table 9.5: Persons per ccTLD and TLD registration, world regions, 2003–2013

	2003		2008		2013	
	per ccTLD	per TLD	per ccTLD	per TLD	per ccTLD	per TLD
World	278	106	101	39	75	29
Developed	62	24	27	9	18	6
Developing	1518	727	264	159	241	131
Africa	5103	2859	1839	933	1053	492
Americas	329	36	137	13	80	10
Asia	1804	723	238	137	260	117
Europe	42	28	18	12	12	8
Oceania	51	27	21	9	12	6

Source: Data supplied by ZookNIC, compiled from ccTLD, Whois and other sources (see below).
Note: Figures exclude 15 ccTLDs that act as virtual gTLDs.

Table 9.6: Internet users per ccTLD and TLD registration, world regions, 2003–2013

	2003		2008		2013	
	per ccTLD	per TLD	per ccTLD	per TLD	per ccTLD	per TLD
World	34	13	23	9	29	11
Developed	26	10	16	6	14	5
Developing	83	40	39	23	74	40
Africa	82	46	151	76	221	103
Americas	99	11	61	6	49	6
Asia	124	50	39	23	84	38
Europe	14	9	9	6	9	5
Oceania	22	12	12	5	8	4

Source: Data supplied by ZookNIC, compiled from ccTLD, Whois and other sources (see below).
Note: Figures exclude 15 ccTLDs that act as virtual gTLDs.

Table 9.7 presents the gross and Internet-user populations per ccTLD and TLD registration for a selection of countries in 2012/2013, including the leading Internet and developing countries included in charts 9.6 and 9.7, and other countries representing different economic groupings.

Table 9.7: Gross and Internet-user population per domain registration, 2012/2013

Country	Persons		Internet users	
	per ccTLD	per registration (total registrations)	per ccTLD	per registration (total registrations)
Germany	5.1	3.4	4.3	2.8
UK	5.8	3.6	5.0	3.1
China	144.2	91.8	61.0	38.8
Russia	29.1	21.6	15.5	11.5
Korea, Rep.	49.8	25.4	41.9	21.3
Brazil	60.7	51.4	30.2	25.6
India	730.1	323.2	91.85	40.7
Indonesia	2434.3	330.3	373.9	50.7
Kenya	1449.8	555.8	465.4	178.4
South Africa	59.4	44.2	24.4	18.1
Australia	8.5	4.2	7.0	3.5
Spain	26.4	11.9	19.0	8.6
Chile	38.7	32.4	23.8	19.9
Venezuela	123.8	89.1	54.5	39.3
Iran, Islamic Rep.	166.1	104.9	43.2	27.3
Thailand	1053.7	84.8	279.2	22.5
Viet Nam	204.3	123.8	80.7	48.9
Mozambique	6457.5	3697.4	313.2	179.3
Niger	119664.4	4766.1	1687.3	67.2
Burkina Faso	201547.6	15197.5	7517.7	566.9

Source: Data supplied by ZookNIC, compiled from ccTLD, Whois and other sources.

Note: Population data are 2013 estimates; Internet user data are 2012 estimates.

While there are limits to the extent to which ccTLD and TLD registrations can be seen as proxies for local content creation, these data clearly show that there is still very considerable diversity in the extent to which the Internet has become pervasive in different countries, and the extent to which content is being published on the web by content providers in different types of country. In particular:

- Developed countries, which have very high rates of Internet access and use, typically also have high numbers of TLD registrations, with the result that, in many cases, they record fewer than ten people and fewer than five Internet users per TLD.
- Middle income developing countries, most of which have rapidly rising Internet user rates, have higher numbers of citizens and Internet users per registration, often with between 30 and 100 citizens (or between 20 and 50 Internet users) per registration.
- Least developed countries are likely to have much lower levels of registration density, as indicated by the figures for Mozambique, Niger and Burkina Faso in Table 9.7.

As expected, there appears from this evidence to be a broad association between the density of TLD registrations and levels of economic development. However, these are clearly not the only factors involved. Within Europe, for example, significantly more citizens and Internet users are recorded per registration in Spain than in Germany or the United Kingdom. Within West Africa, Burkina Faso has a much lower density of both ccTLD and TLD registrations than its neighbour Niger, which has a comparable level of GDP and a comparable Human Development Index ranking.

One additional finding from these data that is worth noting is the relationship between growth rates of Internet use and registration density. The data in Table 9.6 show that the number of Internet users per registration has fallen over the past five years in the Americas, Europe and Oceania, where Internet usage levels are generally high (and therefore no longer growing at a significant rate relative to population). The number of Internet users per registration rose over the period 2008–2013 in Africa and Asia, because they have higher growth rates in Internet usage than in registrations. Where Internet usage levels are relatively low, it is these rather than registration levels that are likely to be the primary determinants of registration density. Measuring registrations against population is therefore a more reliable proxy for local content generation than measuring them against Internet users.

There is considerable scope for further analysis of data concerning registrations, which could shed further light on patterns in the national and international development of the Internet.

Internationalised Domain Names (IDNs)

Internationalised Domain Names can be registered in three different ways:

- through a non-ASCII TLD (such as .中国 for China or .СРБ for Serbia (which can accommodate either ASCII or non-ASCII characters at lower levels (that is, 'before the dot') (IDN TLD)
- by using non-ASCII characters 'before the dot', combined with an ASCII character ccTLD (IDN.ccTLD)
- by using non-ASCII characters 'before the dot', combined with an ASCII character gTLD (IDN.gTLD).

A full range of IDNs became available in 2010, when the first IDN TLDs were authorised. However, IDN.ccTLDs have been available since 2004, while IDN.gTLDs first became available within .com and .net before then.

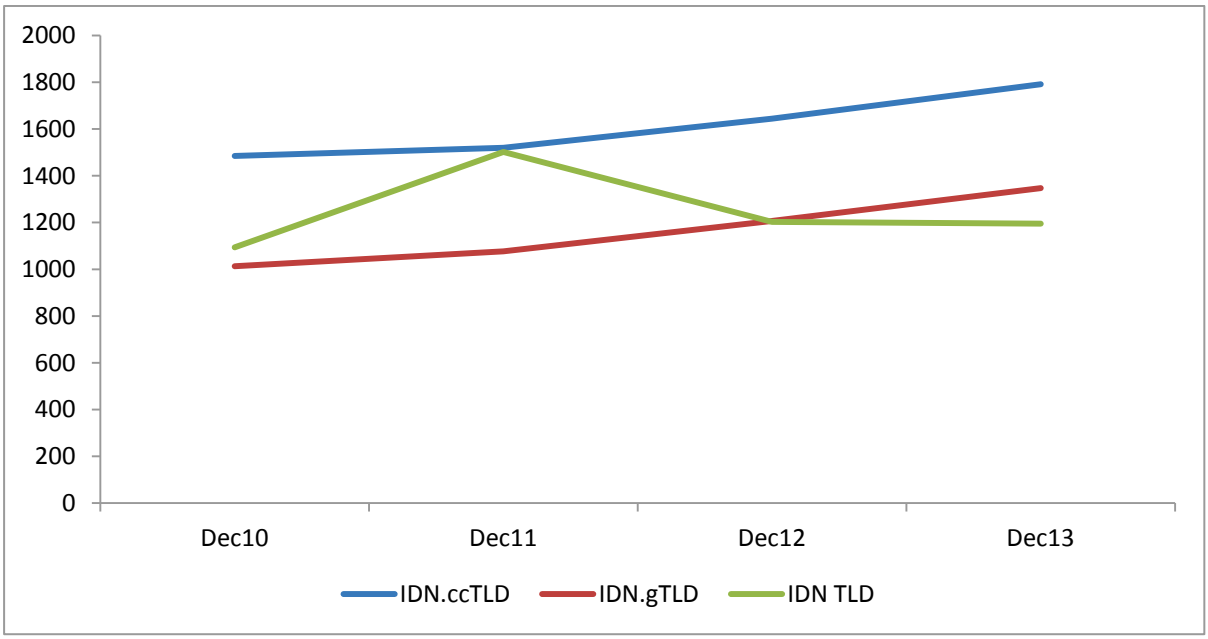
Findings

By late 2013, there were 44 IDN TLDs available, including 41 IDN ccTLDs, representing 31 countries (there were seven representing different language scripts in India), and three IDN gTLDs (one in Japanese and two in Chinese).[62]

Data for IDNs have been kindly provided by ZookNIC for the period since the introduction of IDN TLDs in 2010. Chart 9.8 illustrates the total number of IDN domains extant in December of each year since then.

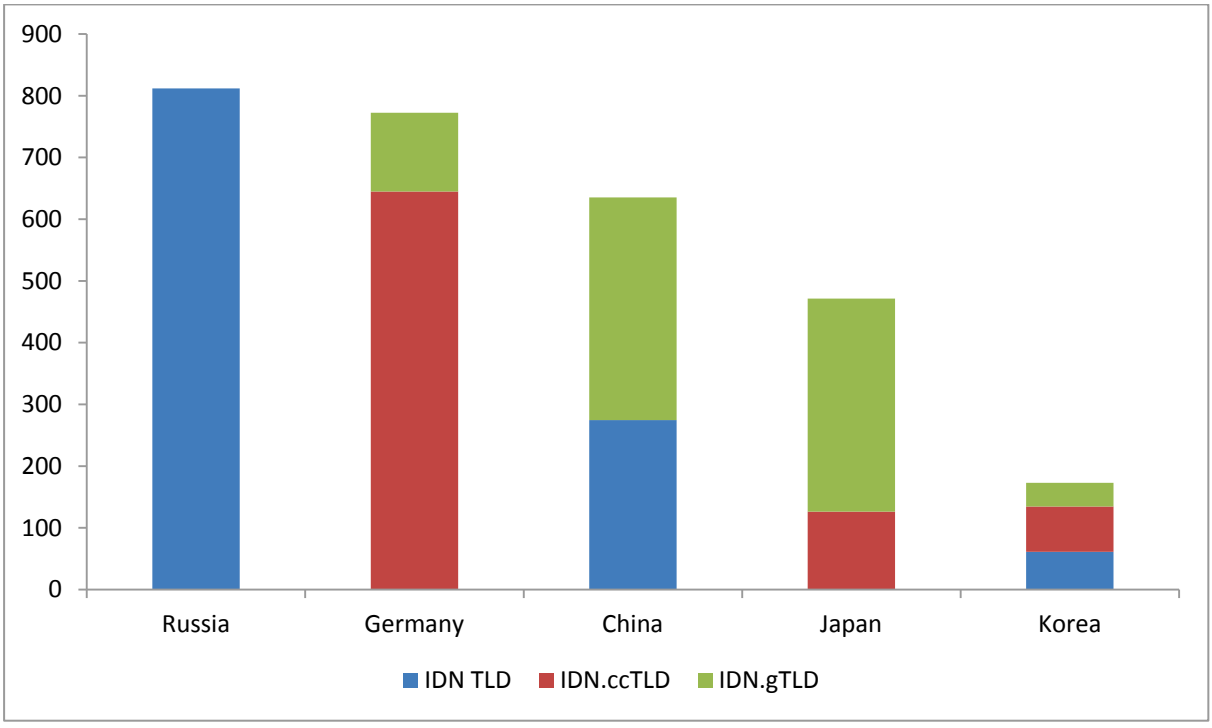
Chart 9.9 illustrates the preponderance of leading countries within each type of domain.

Chart 9.8: Types of IDN, 2010–2013, registrations, thousands



Source: Data from ZookNIC, compiled from ccTLD, Whois and other sources.

Chart 9.9: Number and type of IDN, leading countries, 2013, thousands



Source: Data from ZookNIC, compiled from ccTLD, Whois and other sources.

These data show that there has been only modest uptake of IDNs in recent years. Following an initial surge, since 2011 there has been a decline in the number of registrations with IDN TLDs and only modest growth in the total number of IDN registrations, to stand at just over 4 million in December 2013, according to ZookNIC's data. (This compares with a figure of 5 million cited by the .eu registry EURid and UNESCO in their annual report on world deployment of IDNs for 2013 (UNESCO and EURid, 2013). The difference between these figures probably results from different counting norms

concerning unused registrations.[63]) IDNs remain a small proportion of global registrations and this proportion is not growing rapidly at present.

The number and share of Wikipedia articles by language[64]

This is measured by Indicator 9.5, which is concerned with the number of Wikipedia articles by language. It seeks to observe this over time, as a proxy for user-generated online content creation, using data published by the Wikimedia Foundation.[65]

Wikipedia is the largest and most widely used online encyclopaedia. Founded in 2001, its content is created by an online community of independent contributors and editors. Although the most substantial volume of Wikipedia content is in English, by the end of 2013 content was available online in 287 languages. However, 162 of these are listed as having fewer than 1 000, and 63 as having fewer than 100, articles.[66] According to Alexa's rankings, which are derived from selective toolbar-based monitoring, Wikipedia is one of the ten most visited WWW sites, both worldwide and in the majority of the 126 countries on which it publishes data, though it is likely to rank lower than this on pageviews. By February 2014, it received more than 20 billion page views per month, accessing more than 30 million pages of content. In only a few countries – including China – was it not the predominant reference site.[67]

Wikipedia and its related sites (such as Wiktionary and Wikinews) are coordinated by the Wikimedia Foundation, a non-profit organisation that publishes wide-ranging statistical information about its content and other aspects of performance.[68] Although indicator 9.5 is specifically concerned with the language distribution of Wikipedia content (articles), publication of these data also allows analysis of content creation and content access/usage by language. These related aspects of Wikipedia content are also discussed below.

Wikipedia data provide an illustration of trends in online content and language from a website that has a high level of popularity across the globe – though one that is less widely used in some substantial Internet markets, such as China, Russia and the Republic of Korea, than it is in others.

Wikipedia is also to some extent a proxy for the development of user-generated content. However, while its content is user-generated, the ratio between content creation and content access on Wikipedia – between contributors/editors and users/readers – is very different from that on social networking and microblogging platforms, where the majority of users both create and access content. Some information on other social media platforms is therefore included later in this chapter.

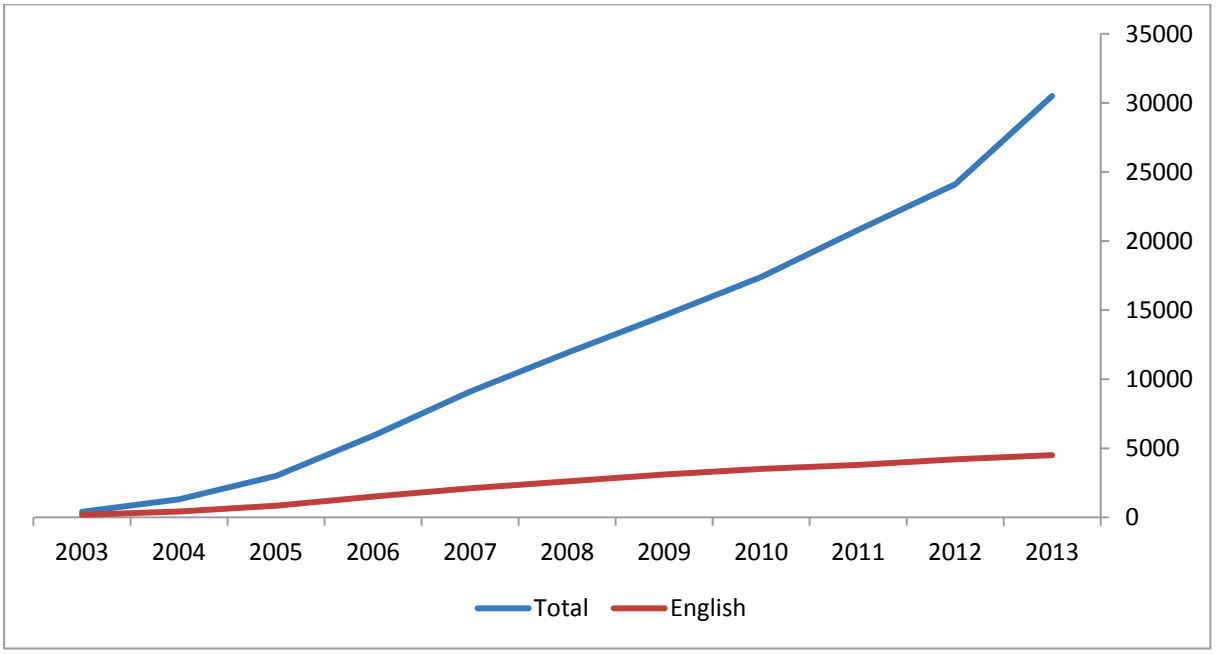
Indicator 9.5 has value as a proxy indicator for changes in Internet content by language, though it has significant limitations and its representativeness should not be assumed. Wikipedia is, at present, a major Internet presence in most countries and territories, but reliance on it as a proxy for content by language will lead to under-representation of languages where it is not the primary reference site (particularly Chinese). Furthermore, Wikipedia content is not necessarily representative of other Internet content (or of social media content), nor can Wikipedia's continued pre-eminence as a reference tool be assumed. Nevertheless, much more extensive data are publicly available on Wikipedia content than on other comparable platforms. If the target is retained, therefore, indicator 9.5 should be retained, and extended to include data concerning the languages of contributors/editors (content creation) and page views (content access/usage) as well as articles

(content created). The treatment of bot-generated and automatically translated articles within analysis (see below) should be reviewed, in conjunction with the Wikimedia Foundation.

Findings

The total number of Wikipedia articles has risen from 398 000 at the end of December 2003 to 30 500 000 in December 2013. The proportion of articles written in English has declined during this period fell from 46 per cent to 15 per cent. These trends are illustrated in Chart 9.10.

Chart 9.10: Wikipedia articles – total and English language, 2003–2013, thousand articles



Source: Wikipedia statistics at http://stats.wikimedia.org/EN/TablesArticlesTotal.htm.

Data for the number of articles in each language available on Wikipedia are published in time series dating back to 2003. The number of articles by language recorded by Wikimedia at December in each of these years, for the ten languages identified in indicator 9.2, and for all other languages, is set out in Table 9.8.

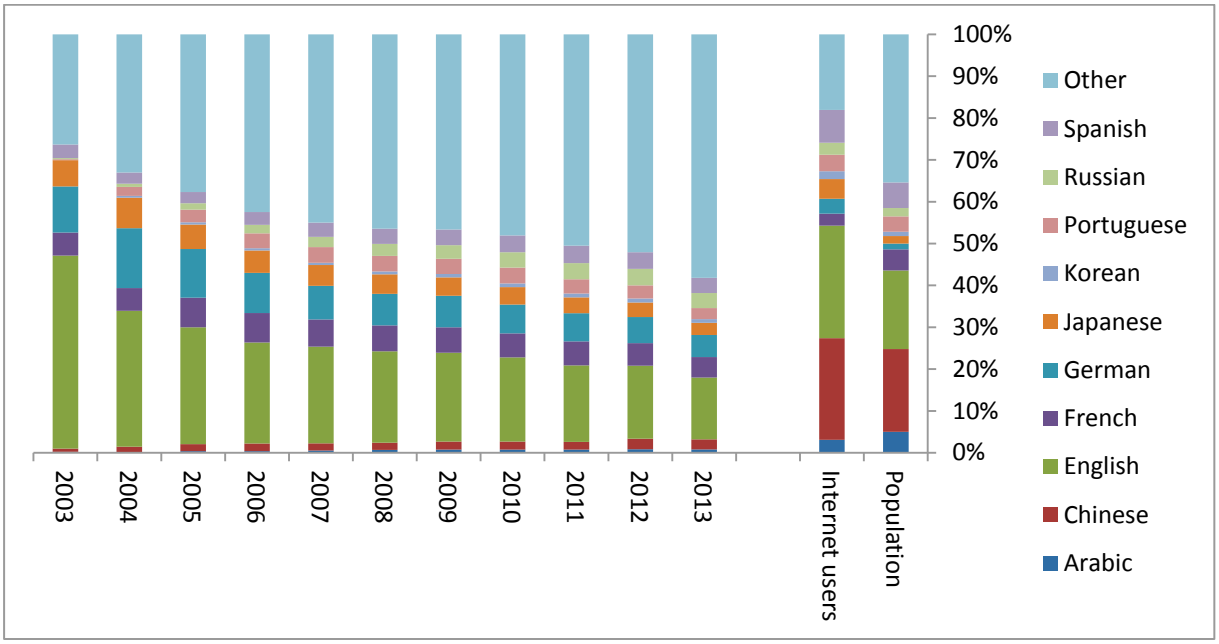
Table 9.8: Wikipedia articles by language, 2003–2013, thousand articles

Year	Arabic	Chinese	English	French	German	Japanese	Korean	Portuguese	Russian	Spanish	Other
2013	250	733	4500	1500	1600	895	262	810	1100	1100	17748
2012	204	610	4200	1300	1500	843	228	760	946	964	12541
2011	160	382	3800	1200	1400	788	186	710	801	870	10501
2010	135	329	3500	1000	1200	726	151	659	638	701	8359
2009	112	279	3100	886	1100	644	119	528	474	550	6807
2008	80	207	2600	738	897	552	83	443	343	431	5526
2007	49	158	2100	592	732	452	49	340	223	310	4095
2006	22	106	1400	410	557	310	30	206	118	178	2463
2005	11	51	837	212	349	175	16	91	46	80	1129
2004	2	17	422	71	186	95	5	28	10	35	429
2003	1	3	184	22	44	25	0	1	1	13	105

Source: Wikipedia statistics at http://stats.wikimedia.org/EN/TablesArticlesTotal.htm, accessed 11 April 2014.

The distribution of articles by language for the ten years since 2003, revealed in Table 9.8, is illustrated in Chart 9.11. This also sets this distribution against the distribution of languages spoken by the population as a whole and by Internet users resulting from the IWS data for 2011 presented under Indicator 9.2.

Chart 9.11: Distribution of Wikipedia articles by language, 2003–2013



Source: Wikipedia statistics at http://stats.wikimedia.org/EN/TablesArticlesTotal.htm, data for December each year.

Chart 9.11 shows that there has been a strong reduction in the proportion of Wikipedia articles that are in English, which has fallen from 46 per cent in 2003 to 15 per cent in 2013; and a corresponding increase in the proportion of articles that are in languages other than the ten most-used international languages, up from 26 per cent in 2003 to 58 per cent in 2013. In fact, several languages that are not in the top ten most popular languages, as identified in indicator 9.2, have substantially higher numbers of articles on Wikipedia than some of those included in this chart. These include several European languages. Dutch, Italian, Polish and Swedish all accounted for more than 1 million articles in 2013, some four times or more than the figure for Arabic or Korean; while Ukrainian and Catalan accounted for just under half a million each. There were also high numbers of articles in some non-European languages, including just under a million in Vietnamese.[69]

These article counts need to be interpreted with care.

- Different language groups in the Wikipedia community take different views of the appropriateness of bot-generated content and automated translation. These differences account for the rapid growth of content in some languages, for example the seventyfold and 20-fold growth in content in the Filipino languages Waray-Waray and Cebuano during 2013. There are even 119 000 articles in the artificial language Volapuk, almost all created in or before 2008. These differences in Wikipedia practice by language community exaggerate the growth in content in 'Other languages' overall.[70]
- Articles also vary in length and depth. Wikipedia did not at the time of writing publish comprehensive data on the number of words by language in Wikipedia content, or on the number of longer articles by language against which the variation in length and depth of

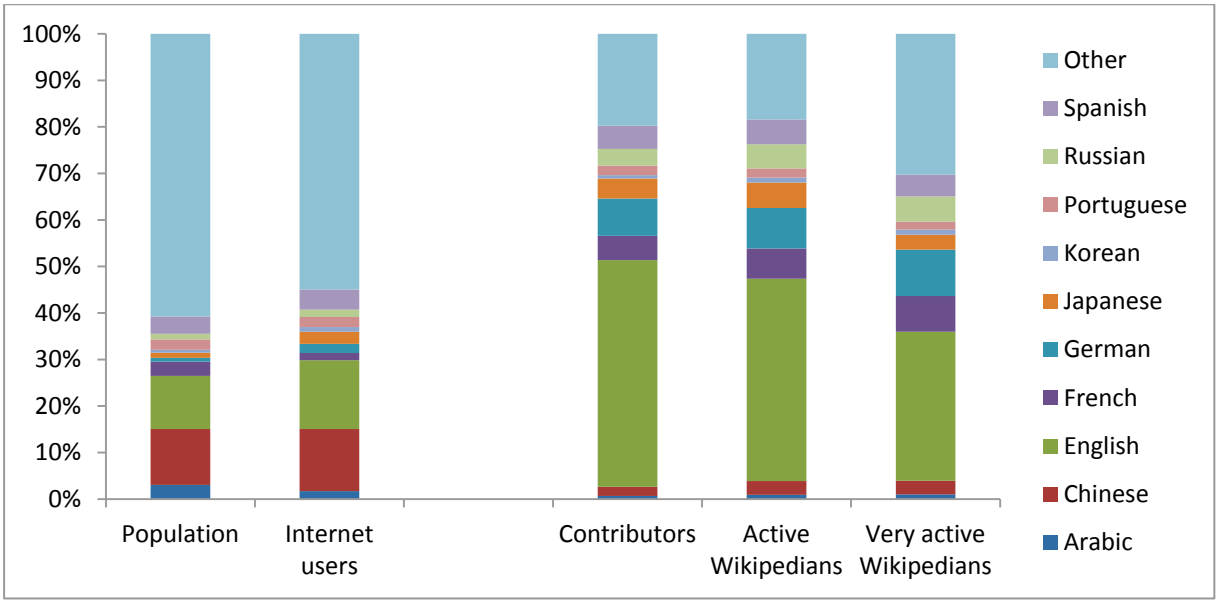
language can be assessed. However, comprehensive datasets for these indicators, covering all languages, should be available from the latter half of 2014.[71]

- As well as considering the language in which content is written, it is also useful to consider the cultural diversity of article content. To do so goes beyond the remit of this report, but some statistical research has been undertaken using articles concerned with geographic locations.[72] Suitable measures of the range of content could be included in future monitoring and measurement.

Regardless of these caveats, the data presented above suggest a strong current of diversification in the languages in which Wikipedia content is available. These can be juxtaposed to some extent with data representing content creation and content access/use.

Chart 9.12 illustrates the proportion of Wikipedia contributors (those who have contributed ten edits or more throughout the life of Wikipedia), active contributors (those who contribute five or more edits per month) and very active contributors (those contributing more than 100 edits per month), in the different language groups. (Figures for contributors therefore include historic contributors who no longer participate, while the other columns include only those who are currently active.) It shows a significantly higher predominance of content creation by contributors writing in English than is suggested by the proportion of articles in Chart 9.11. The higher proportion of minority language users among very active contributors suggests that contributions in those languages may tend to come from a small number of enthusiasts rather than a wider circle of occasional contributors.

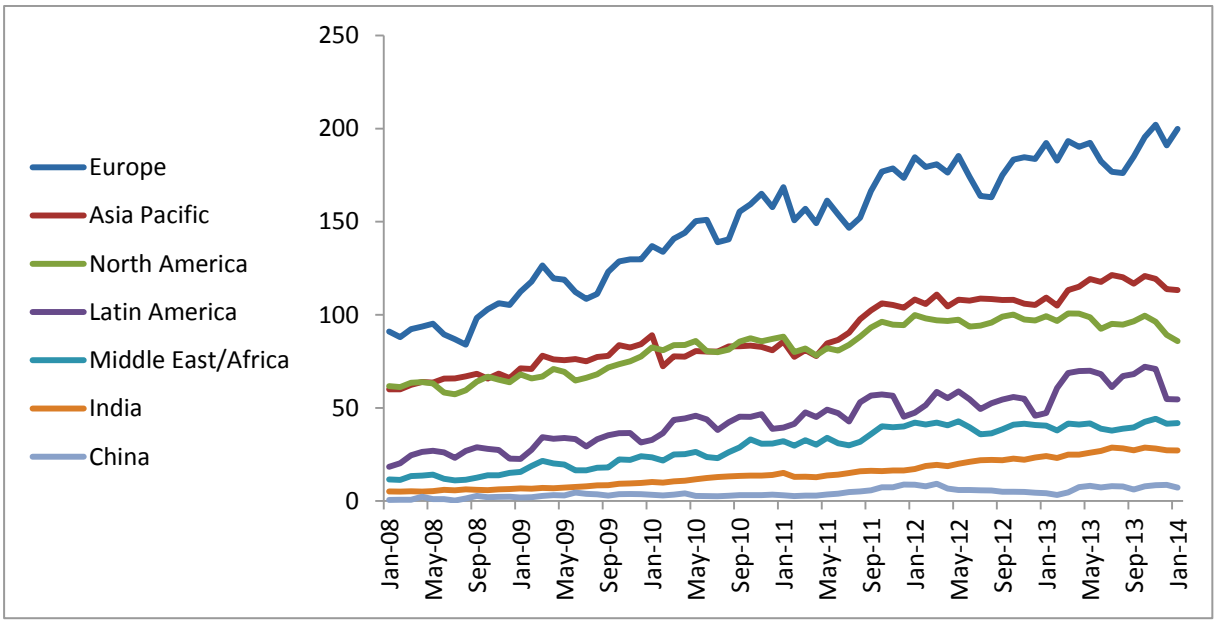
Chart 9.12: Linguistic density of Wikipedia contributors, December 2013



Source: Wikipedia statistics accessed via http://stats.wikimedia.org/EN/Sitemap.htm, accessed 11 April 2014.

The growth of Wikipedia monthly unique visitors by country/region, derived from Comscore's sampling methodology rather than from Wikimedia data, is illustrated in Chart 9.13.

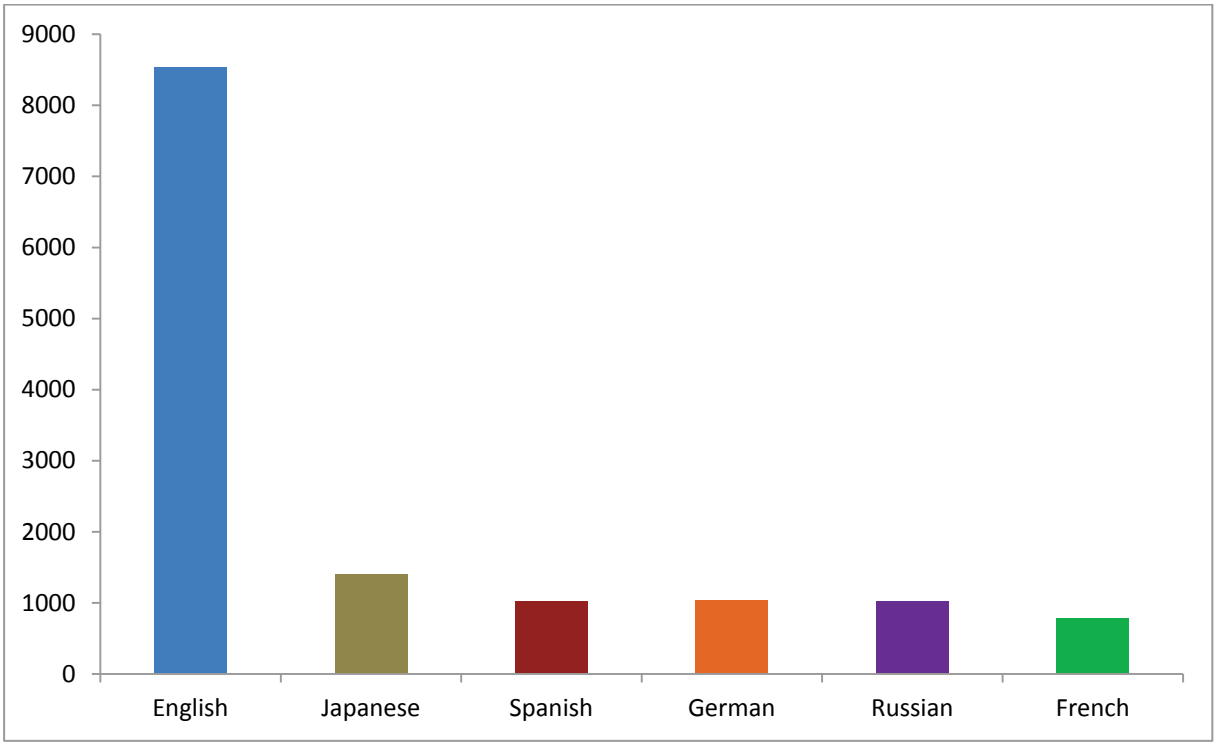
Chart 9.13: Wikipedia unique visitors by country/region, 2008–2013, million pageviews



Source: Wikimedia data at http://reportcard.wmflabs.org/#, derived from Comscore, viewed 10 March 2014.

The distribution of monthly pageviews by language, as at December 2013, is set out in Chart 9.14.

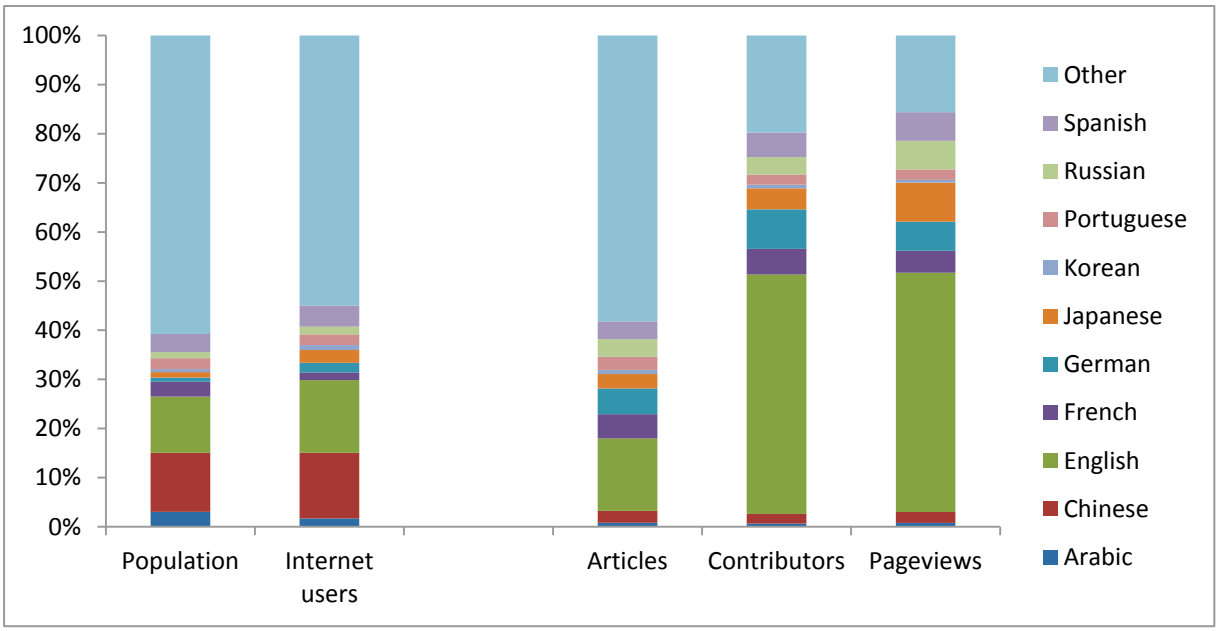
Chart 9.14: Wikipedia monthly pageviews (millions), leading languages, December 2013



Source: Wikimedia data accessed via http://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm.

These measures of content creation, content itself and content access can be drawn together as in Chart 9.15.

Chart 9.15: Wikipedia contributors, articles and pageviews, leading languages, December 2013

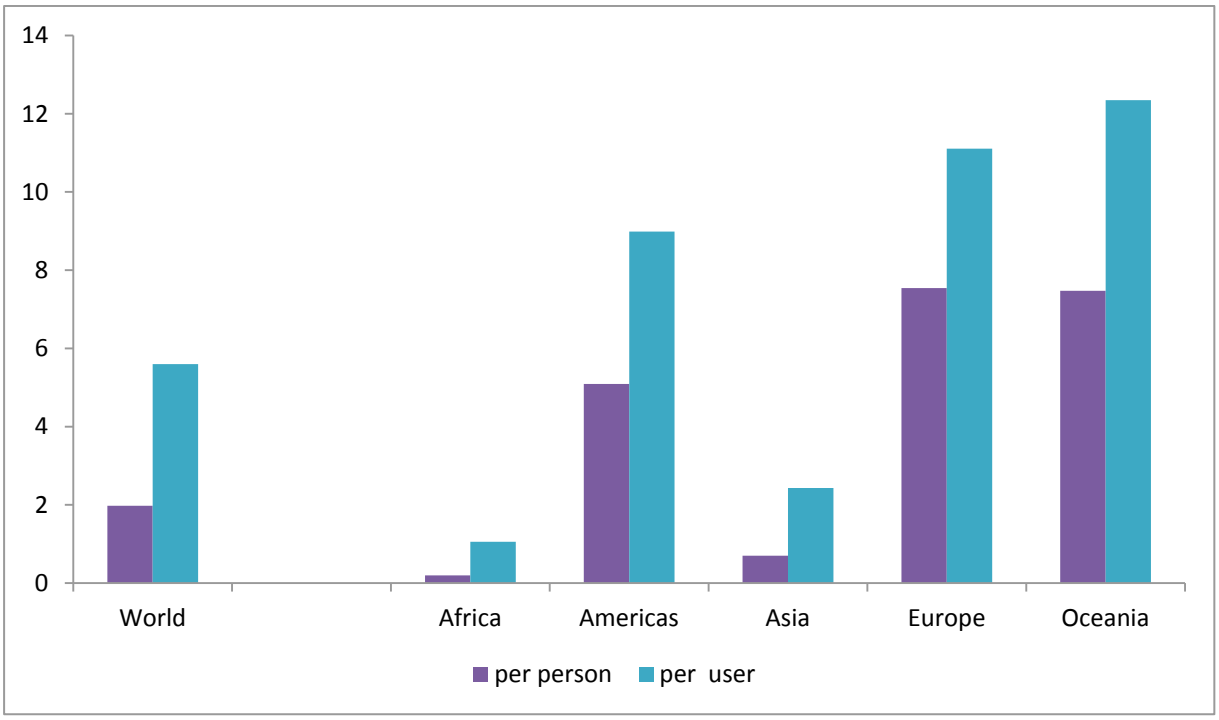


Source: Wikimedia data accessed via http://stats.wikimedia.org/EN/Sitemap.htm, accessed 11 April 2014.

Chart 9.15 illustrates that, while the proportion of English language articles on Wikipedia has declined, English has remained the predominant language for access by Wikipedia users.

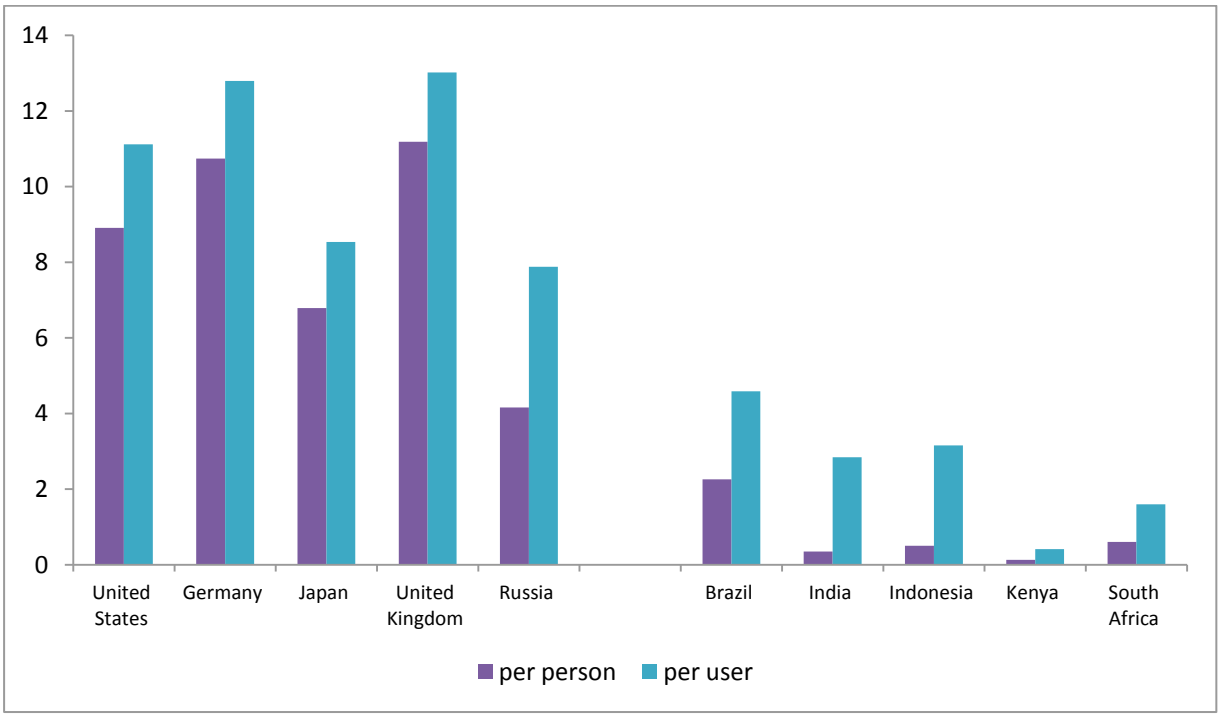
As noted above, Wikimedia data allow comparisons to be drawn between different regions and countries, including the five countries selected for specific attention in this chapter. The distribution of Wikipedia pageviews between different world regions, averaging monthly page views in the first three months of 2014, is set out in Chart 9.16. (Regions used by Wikipedia may differ slightly here from those used elsewhere in this report. Data for global population and estimates of Internet users are also higher than those used elsewhere in this report, but differences in the data sets do not allow these to be adjusted more precisely.) Comparable data for selected leading Internet using countries and for the five countries selected for this chapter are set out in Chart 9.17.

Chart 9.16: Wikipedia monthly pageviews per person/per Internet user, by region, Jan–Mar 2014



Source: Wikimedia data at Wikipedia data at http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryOverview.htm.

Chart 9.17: Wikipedia monthly pageviews per person/per Internet user, countries, Jan–Mar 2014



Source: Wikimedia data at http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryOverview.htm.

Other data published by Wikimedia show that significant proportions of users in most countries use English or, to a lesser extent, other international languages when accessing Wikipedia content. In many countries, the proportion of page views in English is around 10 per cent of the total.[73] This

could be due to an expectation among users that English language content on most subjects will be more extensive than that in other languages.

In some countries, the proportion of page views in English is very much higher than that in local languages: in Pakistan, for example, in the first three months of 2014, 95 per cent of pageviews were in English, with only 1 per cent in Urdu; in India 74 per cent were in English; in Malaysia 73 per cent were in English, with 13 per cent in Chinese and 6 per cent in Malay; in Ethiopia 91 per cent were in English and only 2 per cent in Amharic; in Tanzania 88 per cent were in English and only 3 per cent in Swahili. French is similarly dominant in Francophone Africa, accounting for 77 per cent of pageviews in Senegal and 77 per cent in Mali, while a further 16 per cent and 14 per cent respectively were in English.[74]

Additional evidence concerning content and language

As discussed above, the five Indicators that were selected for Target 9 in 2011 can provide only a partial account of the development of content and language online since WSIS.

- There are severe limitations to the availability of data, particularly for Indicators 9.1, 9.2 and 9.3.
- Indicator 9.4, as adjusted above, provides useful data concerning the publication of online web content, but not of access to or usage of web content, nor of the publication, access and use of content on social media sites.
- Indicator 9.5 provides comprehensive data concerning content, content creation and usage for one specific form of user-generated content, but does not provide evidence concerning other, more common, forms of social media where content is generated by wider user groups.

The following paragraphs supplement information derived from the indicators above concerning two important aspects of the overall environment for online content and language as it has evolved since WSIS – access and use of websites, and access and use of social media platforms.

Website usage

A number of sources are available that identify the most accessed websites in different countries. The Internet analysis companies Alexa and Comscore research the use of websites globally and in particular countries in order to provide advisory services to online businesses and organisations. Both make use of user samples that provide data through monitoring software (in Alexa's case, a toolbar) together with weighting adjustments. They are not therefore comprehensive and the reliability of their results cannot be guaranteed.

More data are made publicly available by Alexa, which reports on the use of websites in 126 countries.[75] However, Alexa data do not include access through mobile devices,[76] and this is likely significantly to affect findings in countries, such as most of those in Africa, where mobile devices have become the primary platform for Internet access. Some corrective to this can be found in data from the browser company Opera, which monitors access to websites on its mobile browser Opera Mini. However, this browser accounts for a small proportion of the mobile browser market and is particularly popular in certain countries, and so may also be unrepresentative.[77]

Nevertheless, data from Comscore and Alexa are widely used within the industry. Alexa's published findings illustrate in particular:

- the preponderance of a small number of global sites, particularly those providing search and social media content, in the majority of countries surveyed and
- variations between countries in the significance of local websites and sites in local languages.

Table 9.9 lists the most popular websites globally and in a number of leading world countries in early 2014, using visitor data from Alexa. Table 9.10 adds equivalent data for the five countries selected for special observation in this chapter, together with November 2010 data (the latest available) for Opera users in those countries.

Table 9.9: Website popularity, global and selected countries, 2014

	Global	USA	Germany	UK	Spain	China	Japan	Korea, Rep.	Russia
1	Google	Google	Google.de	Google.uk	Google.es	Baidu	Yahoo.jp	Google	Yandex
2	Facebook	Facebook	Facebook	Google.com	Google.com	QQ	Google.jp	Facebook	VK
3	YouTube	YouTube	Google.com	Facebook	Facebook	Taobao	Amazon.jp	Naver	Google.ru
4	Yahoo	Yahoo	YouTube	YouTube	YouTube	Sina.com	Google.com	YouTube	Google.com
5	Baidu	Amazon	Ebay	BBC	Blogspot.es	Hao123	YouTube	Google.kr	Mail.ru
6	Wikipedia	Wikipedia	Amazon.de	Ebay	Twitter	Weibo	FC2	Baidu	YouTube
7	QQ	LinkedIn	Wikipedia	Yahoo	Live.com	Tmall	Facebook	Daum	Odnoklassniki
8	Twitter	Ebay	Spiegel	Amazon	Wikipedia	Sohu	Rakuten	QQ	Facebook
9	Live.com	Twitter	Bild	Wikipedia	Yahoo	360.cn	Wikipedia	Yahoo	Wikipedia
10	LinkedIn	CraigsList	Yahoo	LinkedIn	LinkedIn	163.com	Ameblo	Taobao	LiveInternet
11	Taobao	Bing	Web.de	Live.com	Marca	Soso	Livedoor	Tistory	LiveJournal
12	Amazon	Pinterest	GMX	Twitter	Wordpress	gmw.cn	Nicovideo	Blogspot.kr	Avito
13	Google.in	Blogspot	T-Online	Daily Mail	El Mundo	ifeng	goo.ne.jp	Wikipedia	Rambler
14	Sina.com	Go.com	Xing	Paypal	El Pais	Xinhuanet	Naver	nate.com	rbc.ru
15	Blogspot	CNN	Uimserv	Guardian	Amazon.es	Google.hk	Twitter	sinacom.cn	Twitter
16	Hao123	Live.com	Blogspot	Wordpress	Milanuncios	Alipay	dmm.co.jp	gmarket	RuTracker
17	Weibo	Paypal	Gutefrage	Amazon	Lacaixa	People.com.cn	msn.com	ask.com	ucoz.ru
18	Wordpress	Instagram	Chip.de	Pinterest	Pinterest	China.com	xvideos	hao.123	sberbank.ru
19	Yahoo.jp	Tumblr	xhamster	Tumblr	Ebay.es	Youku	Kakaku	blog.me	AliExpress
20	vk.com	ESPN	focus.de	Telegraph	as.com	Sogou	Baidu	ecplaza	lenta.ru

Source: http://www.alexa.com/topsites, accessed 9 April 2014.

Table 9.10: Website popularity, computer and mobile platforms

	Brazil		India		Indonesia		Kenya		South Africa	
	Alexa 2013	Opera 2010	Alexa 2013	Opera 2010	Alexa 2013	Opera 2010	Alexa 2013	Opera 2010	Alexa 2013	Opera 2010
1	Google.br	Google	Google.in	Google	Google.com	Facebook	Google.com	Facebook	Google.za	Facebook
2	Facebook	Orkut	Google.com	Facebook	Facebook	Google	Facebook	Google	Google.com	Google
3	Google.com	Live.com	Facebook	Orkut	Blogspot	Detik	YouTube	Wikipedia	Facebook	Mxit
4	YouTube	YouTube	YouTube	YouTube	YouTube	YouTube	Yahoo	Wapdam	YouTube	YouTube
5	UOL	Globo	Yahoo	Getjar	Yahoo	Yahoo	Google.ke	YouTube	Yahoo	Wikipedia
6	Globo	Twitter	Blogspot.in	Zedge	Google.id	Wapdam	StandardMedia	Yahoo	Gumtree	Mygamma
7	Yahoo	MSN	Wikipedia	Yahoo	Kaskus	Twitter	Twitter	BBC	LinkedIn	Getjar
8	Live.com	Facebook	LinkedIn	Songs.pk	Wordpress	Wikipedia	Wikipedia	Getjar	Wikipedia	Thumbtribe
9	Blogspot.br	uol.com.br	IndiaTimes	Wikipedia	Detik	Getjar	Nation	My Opera	News24	Zamob
10	Mercadolivre	4shared.com	Flipkart	Vuclip	Twitter	Vivanews	Blogspot	Reference.com	FNB	Yahoo

Source: http://www.alexa.com/topsites, accessed 9 April 2014; Opera, State of the Mobile Web, November 2010, http://www.operasoftware.com/archive/smw/2010/11/index.html.

These data, while imprecise for reasons described above, illustrate a number of important points concerning the development of content since 2003.

- A small number of international websites account for a high proportion of web access both globally and in the majority of countries. These sites include search engines (particularly Google

and Yahoo, which have become the principal conduits or portals for Internet users seeking content, often used now as a substitute for entering URLs as well as for pure search), online social networks (particularly Facebook), blog sites (particularly Blogspot and Wordpress), microblogs (particularly Twitter and, in China, Weibo), video file-sharing sites (specifically YouTube), and online reference sites (particularly Wikipedia). In a high proportion of countries monitored by Alexa, Google, Facebook and YouTube feature in the top five positions in the rankings, in some cases through ccTLD rather than gTLD domains. Some additional data on social media websites can be found below.

- There are a small number of countries in which these global sites are not predominant or not so predominant, usually because of the presence of local (or local language) alternatives. This is particularly the case in four countries with large populations, whose languages use non-Latin alphabets – China, Japan, Republic of Korea and Russia. The preponderance of Chinese alternatives to international search and social networking sites is so strong that three of these feature among Alexa's top ten global websites.[78]
- In most countries, some local sites also have audiences within the top ten and certainly within the top 20 websites. As well as social networks, these include e-commerce sites (such as Taobao in China and Mercadolive in Brazil) and mainstream national media (such as the BBC and several newspaper websites in the UK, *Der Spiegel* and *Bild* in Germany, and the *East African Standard* and *Daily Nation* in Kenya).
- Some differences are suggested between computer and mobile access, though data here are unreliable because of the different dates involved and the limited market share of the Opera browser. Nevertheless, mobile usage illustrates the popularity of content platforms that are specific to mobile devices, such as the South African instant messaging service Mxit and the mobile app store Getjar. More analysis is needed of the differential use of content between computer and mobile platforms.

Social media usage

As noted above, social media and other sites offering user-generated content have become very prominent in Internet usage since 2003, and must be included in any current or future assessment of online content and language. These sites include social networks such as Facebook, LinkedIn and RenRen; blog sites such as Blogspot and Wordpress; microblogs such as Twitter and Tencent Weibo; messaging and VoIP sites such as Yahoo Messenger and Skype; and audio, image and video filesharing sites such as Flickr, Instagram and YouTube. They provide new spaces for content creation, sharing and usage, including content intended for both general and specific readerships. As social media content is user-generated, it may be more likely than other online content to be written in users' primary languages, though this is difficult to assess with the limited data that are available at present.

While this chapter is concerned with content on the Internet, it should also be recognised that there has been a correspondingly rapid growth in the volume of content that is specific to mobile phones, originally including SMS messages but more recently including content accessed and shared through mobile apps. Much of this content is also user-generated. While these mobile content platforms have not been discussed in this chapter, they should be included in future assessments of trends in content and language that measure outcomes relating to WSIS Target 9.

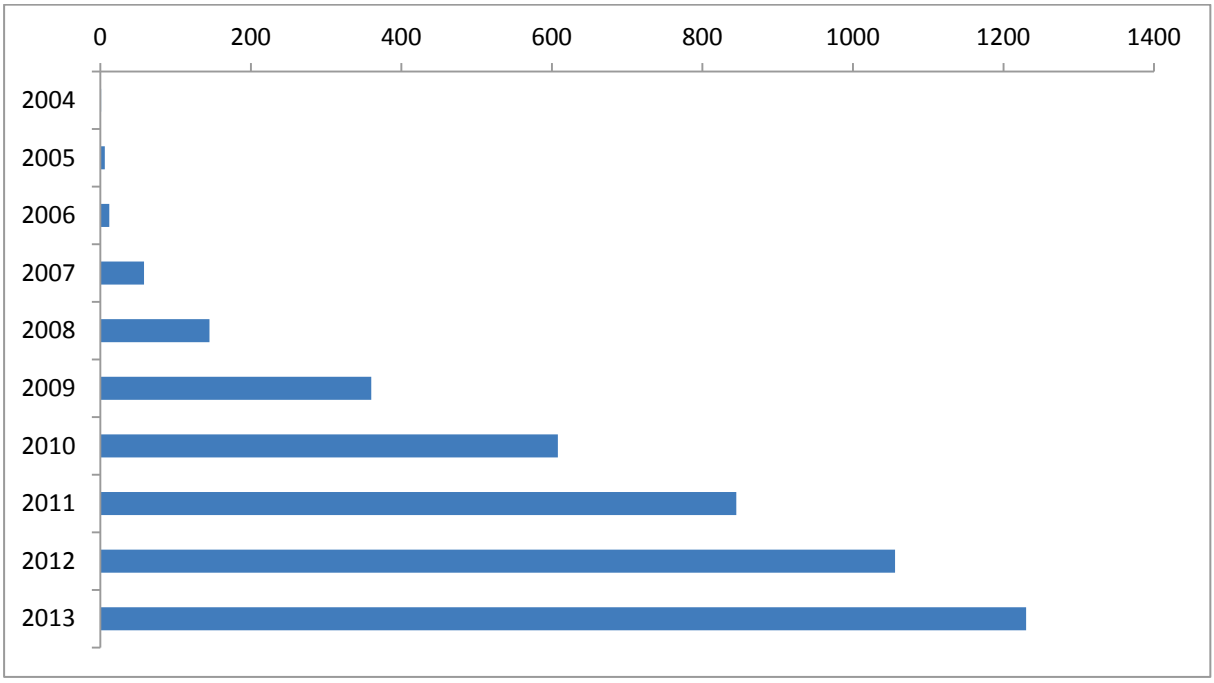
Most prominent social media sites are commercial businesses, whose business models rely on data mining to target advertising at site users. As a result, while they collect extensive data about their users (content creators and readers) and the content they create and access, these data are commercially confidential and not available for public analysis. The following paragraphs provide some information, derived from published sources, which address the impact of these sites on content and language.

It is important when analysing data concerning global social media platforms to remember that they are not universally predominant. Alternative services are popular in a number of countries, including China, Russia, Republic of Korea and Japan. In particular, global social media platforms such as Facebook, Twitter and YouTube are not generally available in China. A report published in 2013 estimated that over 90 per cent of Chinese Internet users then had at least one social media account, with the most popular platforms including Qzone (blogs and photo-sharing), Tencent Weibo (microblog), Sina Weibo (microblog and social network), Wechat (messaging), PengYou and RenRen (social networking) and 51.com (gaming).[79]

The most prominent social network in most countries, but not in China, is Facebook, which was established in 2004. By the end of 2013, Facebook was clearly established as the predominant social network worldwide, dominating the market for social media in most countries and identified as one of the two most popular websites in a substantial majority of countries in Alexa counts of web usage. By the end of 2013, it registered more than 1.3 billion monthly and 757 million daily active users, was available in 70 languages and was accessed by as many as 40 per cent of active Internet users daily. Some ten billion Facebook messages were said to be posted daily.[80]

Detailed (for example, country-level) information on the growth of Facebook is not readily available, but some general data have been published. The growth in the number of those using Facebook at least monthly is illustrated in Chart 9.18.

Chart 9.18: Growth in Facebook monthly active users, 2004–2013, millions of users



Source: The Guardian newspaper website, http://www.theguardian.com/news/datablog/2014/feb/04/facebook-in-numbers-statistics, accessed 6 March 2014. Data sourced from Facebook.

Data published in 2012 showed that the countries with most Facebook users, after the United States, were Brazil and India (with over 50 million users each), followed by Indonesia and Mexico. The most popular languages after English were Spanish (with around 80 million users), followed by Portuguese (principally because of users in Brazil), French, Indonesian and Turkish. The fastest growing languages between May 2010 and November 2012 were Portuguese and Arabic.[81]

However, as with Wikipedia data (above), there were substantial differences between language behaviour in different countries. Data have also been published showing that, while more than 96 per cent of Brazilian Facebook users chose Portuguese as their default language in 2012, almost all of those in India chose English with less than 1 per cent selecting Hindi.[82]

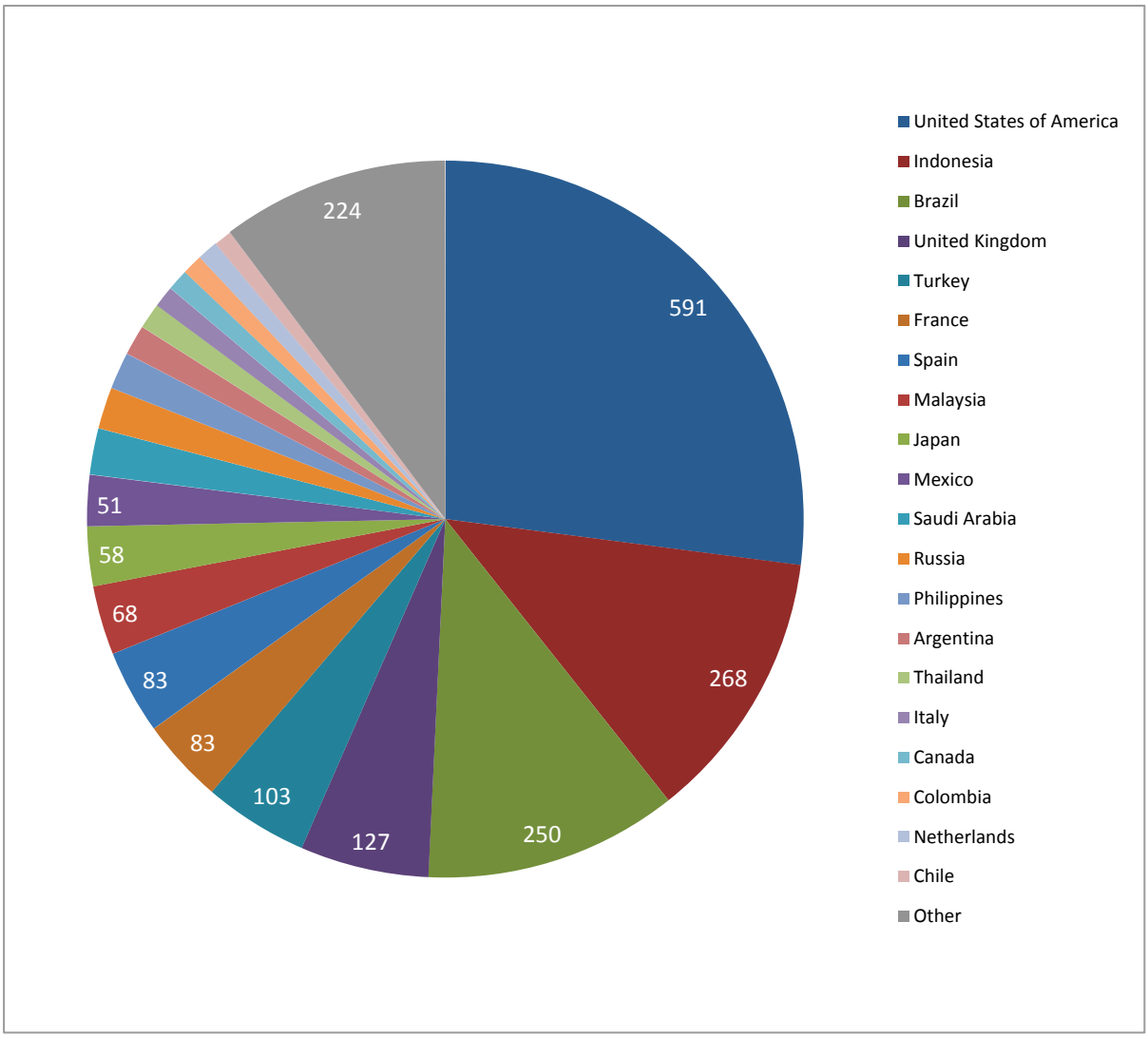
Differences in user behaviour on Facebook are well-illustrated by data on the language distribution of Facebook use in Arabic-speaking countries published by the UN Economic and Social Commission for Western Asia in 2013 (UNESCWA, 2013).[83] These showed the preponderance of Arabic use on Facebook varying from 81 per cent in Yemen to just 4 per cent in Tunisia. English was the predominant user language in six, and French in three, Arabic-speaking countries. The incidence of Arabic usage on Facebook had increased since 2011 in a number of countries in the region, including Iraq, Egypt and Jordan, while significant use in languages other than Arabic, English and French was evident in those countries with large expatriate populations.

Further country and language data are available on commercial terms from Facebook but have not been reviewed for this report.[84]

The most prominent microblog in most countries, though not in China, is Twitter, which had built a user community of some 646 million between its establishment in 2006 and January 2014, 115 million of whom were active at least once a month.[85] The number of 'tweets' – messages of up to 140 characters – posted daily by Twitter users was approximately 58 million. Around 60 per cent of users published tweets, according to published data, the remaining 40 per cent being passive readers.

The DOLLY project[86] at the University of Kentucky measures those tweets that can be geolocated because of settings that have been enabled by terminal users – a total of between 1 per cent and 2 per cent of tweets, mostly created on mobile devices. While not random and so not necessarily representative, this provides a sample of over 2 billion tweets posted during 2013. Chart 9.19 illustrates the geographic distribution of tweets originating in the 20 most popular countries within this sample, using data kindly provided by the project.

Chart 9.19: Geocoded origin of a sample of 2 billion tweets, 2013, million tweets



Source: data supplied by the DOLLY project, University of Kentucky.[87]

Some analysis has also been undertaken of tweets by language. Researchers who analysed 380 million geolocated tweets posted from 191 countries between October 2010 and May 2012 identified at least 78 languages within their dataset, the leading languages being English (by a very substantial margin), Spanish, Malay and Indonesian. As with the data reported above, it is unclear if geolocated tweets are representative of tweets in general. English was used in 10 per cent or more of tweets within this dataset that were posted from other leading European countries (France, Italy and the Netherlands), and in 5 to 10 per cent of those from a number of other countries (including Turkey, Chile and Venezuela).[88] A separate study of over 6 million Twitter users from 246 countries and territories, undertaken in 2010, also found that English was predominant, accounting for 53 per cent of tweets in total, for more than 10 per cent of those from the Netherlands, Indonesia and Mexico, and 9 per cent of those from Brazil (Poblete *et al.*, 2011). As Twitter is a form of publication, this may represent users seeking to maximise their global readership. This linguistic pattern may also have changed significantly since 2010, because of the high rate of growth in the number of Twitter accounts worldwide.

The most prominent video filesharing site in most countries is YouTube, which is owned by Google. YouTube reported in February 2014 that its content receives more than 1 billion unique visitors

monthly, those visitors watching approximately six billion hours of content. Its service is localised in 61 countries and available in 61 languages.[89]

There has been similarly strong growth in the posting of image content. It is difficult to confirm the reliability of data, but in March 2014, it was estimated that 200 million users of the photo-sharing website Instagram were adding 60 million items daily to a total already exceeding 20 billion.[90] Flickr was estimated in 2013 to have 87 million users, posting more than 3.5 million images daily to a total exceeding 8 billion.[91]

E-commerce sites and Internet banking represent other forms of content that are local or user-specific in character, access to which should be considered when reviewing content availability and access. Available evidence suggests that participation in e-commerce and Internet banking varies considerably between countries, as a result of economic conditions as well as online behaviour. The Internet research company Comscore found that 29 per cent of Internet users worldwide made use of Internet banking in April 2012, for example, including 45 per cent of Internet users in North America but less than 9 per cent of those in the Middle East and Africa.[92]

Conclusions and recommendations

The measurement of online content and language is far from easy. Nevertheless, it is clear from the evidence presented in this chapter that there has been tremendous growth in the creation, sharing and access of online content in the decade since WSIS, and that there has been growing diversity in the range of languages used for both content creation and content access. While there is still a long way to go before content and language are equally available to all, the trends described in this chapter are broadly positive.

- On the supply side, the number of websites (calculated as the number of allocated URLs) has grown enormously between 2003 and 2013, and the number of webpages even more substantially. Traditional websites have been supplemented by new forms of user-generated content, which are extensively used by individuals, businesses and organisations. Social media such as Facebook, Twitter and Weibo have expanded the range of content available, including local content, and have provided new platforms for both content creation and content access. The volume of video content uploaded to YouTube has also grown enormously since 2003, exemplifying growth in non-text content that is facilitated by the increasingly widespread availability of broadband networks.
- Alongside the Internet, mobile apps have added new opportunities for content creation and access since they first became available in 2008. The number of apps available for Apple iPhones was reported to have exceeded 1 million in October 2013,[93] while the number of Android apps was reported to have exceeded 1 million by February 2014.[94]
- The growth in demand for Internet content has also increased enormously. The number of Internet users has risen from an estimated 1.02 billion in 2003 to an estimated 2.75 billion in 2013, from 16 per cent to 39 per cent of world population.[95] The pace of growth during this period has been particularly marked in developing countries, which accounted for an estimated 31 per cent of total Internet users by 2013.[96]

There has therefore been exceptional growth since 2003 in the volume of content generated, in the numbers of people, businesses and organisations engaged in content creation, in the number of

people accessing content, and in the volume of content accessed. There is also now much greater diversity in the range of online content, thanks to the emergence of social media and mobile applications and to the spread of electronic commerce. The availability of development-related content has been facilitated by the emergence of open data and the spread of electronic government and transactions.

In spite of this, the evidence presented in this chapter shows that there remains a powerful digital divide in both content creation and content access between developed and developing countries. Developed countries in Europe, the Americas and parts of Asia continue to generate the majority of web content. Evidence from the registration of TLDs suggests that there is a broad, but by no means precise, association between measures of development (such as GDP per capita and HDI) and Internet content generation. Low-income developing countries tend to have particularly low levels of TLD registration as well as Internet usage.

It is difficult to assess the developmental impact of digital divides in content creation and access, not least because overall data volumes are distorted by the high demands on bandwidth generated by video content, the majority of which is likely to be for entertainment use. However, the evidence in this chapter tends to confirm the finding of UNESCO, the OECD and the Internet Society that there is a virtuous circle between infrastructure supply, affordability of access and the development and use of local content. Societies that enjoy high quality broadband access at low prices are likely to see greater Internet use, increasing demand for local content that is then supplied by governments, businesses, independent organisations and individual users of the Internet exploiting the potential of social media platforms.

An important policy implication of this is that one of the ways in which governments can most effectively stimulate the market for local content is through the enabling environment for investment in communications networks and services. However, infrastructure is insufficient to address all of the disadvantages that affect content creation and access in developing countries. A number of international reports have addressed aspects of the social and economic context in developing countries that also inhibit content production and use.

On the supply side, these include the small size of many developing country content markets for information and cultural goods, the existence of global services (such as search engines and social media platforms) that facilitate free access to information and information sharing that might otherwise provide a basis for local service development, and complex arrangements for the registration of new businesses, which inhibit service innovation. The growth of cloud computing may reduce some of the financial and administrative costs involved in innovation, encouraging more diverse content generation at local level, but evidence on this is not yet clear.[97]

On the demand side, as well as limited infrastructure capacity and cost, access and use of content are constrained by illiteracy and the lack of media and information literacy skills.

Governments can stimulate content generation and access to content by addressing these constraints. Government websites and open data policies provide an example to other potential content providers as well as offering content that is of direct relevance to local users. Governments can also use social media platforms to disseminate public information, though this should not diminish the use of traditional media. In the longer term, efforts by governments to address media and information literacy, through education and lifelong learning, should raise the proportion of

citizens with the skills and confidence needed to access and exploit the online content resources available to them.

Language is a critical dimension of this. As this chapter has emphasised, it is very difficult reliably to measure online content and access to that content by language, though it is possible to establish trends in language use of online services such as Wikipedia, where data are published. Other online service providers gather comparable data for use in their commercial development but these are not available for independent analysis. The following conclusions are suggested by the evidence:

- There is increasing diversity in the range of languages available and used online. The predominance of English, which was very pronounced in the early period of the Internet, has now reduced, though it is still estimated that more than half of the top ten million websites use English as at least one of their content languages.[98] There has been a marked increase in the web presence of some languages using non-Latin scripts, especially Chinese, though South Asian languages and Arabic have shown less dynamic growth.
- Language is less of a constraint on social media sites, where content is user-generated, than it is on conventional websites. The number of languages available on social media sites has grown significantly, with almost 300 now available on Wikipedia and around 100 each on Google and Facebook. Users are able to post information in the language of their choice, which may or may not be their primary language, though this will be partly determined by the audience they seek to reach as well as by personal language preferences and capabilities. Unfortunately, very little statistical information is publicly available about the languages used in social media and how these are changing over time.
- It is too early at present to assess how much impact the introduction of IDNs will have on linguistic diversity on the Internet, though early evidence suggests that this has not been as significant as had been anticipated. The role of IDNs should continue to be monitored.
- It is clear that there is still a long way to go before content is as readily available in national and local languages as it is in global languages, particularly English. The clearest exception to the continued leading presence of English online is the Chinese Internet market, which is dominated by Chinese language sites that have benefited from constraints on access to global social media platforms in their primary market. In some developing countries, Wikipedia evidence suggests that existing Internet users are more likely to access content in English than in local languages, though this is partly because Internet access has not yet penetrated deeply into social groups that do not have English as a secondary language.

The most significant emerging trend in this field concerns automated translation. Although there will always be quality and reliability challenges, this has the potential to allow end-users to access content written in languages with which they are unfamiliar, when that content would otherwise be inaccessible to them. While the challenge of automated or machine translation has been addressed by computer scientists and linguists since the 1950s, the search for effective and reliable translation mechanisms has become more substantial since the Internet became widespread, focusing on statistical and example-based methodologies. However, dependence on analysis of existing manual translations in developing translation algorithms means that automated translation is likely to be more successful between major languages and offers less of a solution for minority languages that are rarely translated or written. Translation between languages with very different structures and characteristics, such as Latin languages and Chinese, is also problematic.

By 2013, the leading online translation service Google Translate was available in 80 languages, with a free plug-in in 60 languages available to content developers.[99] The capabilities of Google Translate, Bing Translator and similar services will be enhanced by the growth of capacity in cloud data centres that have the computing power to explore very large sets of manually-translated originals. Continued efforts to improve translation capabilities will be the most effective way in which computing and Internet professionals can advance linguistic diversity on the Internet, particularly in facilitating the reach of content into minority language communities.[100]

The growth in content, including local content, over the past decade, which is described above, and the related spread of language diversity online have resulted primarily from developments in the communications market rather than from interventions by governments and international agencies. Increased access to the Internet, the increased capacity of networks to carry high content volumes, and the low cost of publication online have accelerated the growth in web content, while new platforms such as social media and microblogs have enabled all Internet users to contribute their own content at minimal cost and inconvenience.

Internet businesses have responded to this growth in content by providing new platforms for content distribution and extending the range of languages in which content can readily be published.

Governments have supported content growth by facilitating the enabling environment for Internet investment and services, while, in most countries, imposing few restrictions on content access.

The Internet professional community has contributed to greater linguistic diversity by enabling IDNs and fostering the development of automated translation software and services.

The spread of online content and linguistic diversity are critical aspects of the "… people-centred, inclusive and development-oriented Information Society" envisaged in the WSIS outcome documents. It is therefore important to understand trends in both of these aspects of WSIS implementation. However, it is difficult to establish quantitative targets for them, both because there is no stable or finite limit to their potential achievement, and because of severe limitations in the data sets that are currently available. If the target is to be retained for measurement post-2015, revisions will need to be made in the current indicators, and these will need to be supplemented by a wider range of evidence in order to achieve a representative understanding of relevant trends and developments in different countries and regions. The recommended changes are as follows:

- Indicator 9.1 should be retained, but suspended until data of sufficient quality become more comprehensively available as a result of national statistical offices incorporating relevant data collection into national censuses and household surveys.
- Indicators 9.2 and 9.3 should be withdrawn as it is not currently possible to obtain reliable data, and unlikely that this situation will change at least in the short or medium term.
- Indicator 9.4 should be retained in revised form, including gTLDs and IDNs as well as ccTLDs in national counts of domain names, and subject to mechanisms being put in place to secure access to comprehensive data sets from either national registries or independent analysts.
- Indicator 9.5 should be retained but developed to include Wikipedia contributors (content creation) and pageviews (access and use) as well as articles.
- Additional indicators should be developed to replace indicators 9.2 and 9.3. These should be concerned with measuring the volume and linguistic diversity of content on one or more social

networks and on mobile apps. The emergence of further new platforms for content creation, dissemination and access may require further adjustments to indicators in due course.

An alternative or supplementary approach would involve gathering a wider variety of quantitative and qualitative data on a number of specific countries and territories that are selected to be representative of the world community. While this would not have the same statistical value as monitoring of other WSIS targets, it would enable a more substantive qualitative assessment to be made of trends that are taking place in content and language, alongside those statistical indicators that do prove to be viable. Additional statistical evidence from diverse sources could be incorporated in this monitoring and measurement, along the lines suggested in this chapter.

The periodic publication of time series data in tables and figures is only one way of illustrating the spread of online content and language. Consideration could be given to the potential of mapping and other techniques to add insight to those data that are available in this area of WSIS outcomes.

List of references

CISCO (2013), *Cisco Visual Networking Index: Forecast and Methodology, 2012-2017*, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf.

Dubai School of Government (2011), *Civil Movements: the Impact of Facebook and Twitter*, Arab Social Media Report 1-2, 2011, http://www.dsg.ae/portals/0/ASMR2.pdf.

Gantz, J. and Reisel, D. (2011), *Extracting Value from Chaos*, International Data Corporation, http://uk.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

ITU (International Telecommunication Union) (2005), *World Summit on the Information Society Outcome Documents: Geneva 2003 - Tunis 2005*, http://www.itu.int/wsis/outcome/booklet.pdf.

ITU (2010), *World Telecommunication/ICT Development Report 2010: Monitoring the WSIS Targets, A mid-term review*, http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtdr2010.aspx.

ITU (2013), *World Telecommunication/ICT Indicators database 2013, 17th edition*, http://www.itu.int/ITU-D/ict/publications/world/world.html.

ITU (2014), *Manual for Measuring ICT Access and Use by Households and Individuals, 2014 edition*, http://www.itu.int/en/ITU-D/Statistics/Pages/publications/manual2014.aspx.

Mocanu, D. *et al*. (2013), *The Twitter of Babel: Mapping World Languages through Microblogging Platforms*, PLOS ONE, http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0061981.

OECD (Organisation for Economic Co-Operation and Development), UNESCO and the Internet Society (2011), *The Relationship between Local Content, Internet Development and Access Prices*, http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/local_content_study.pdf.

Partnership on Measuring ICT for Development (2010), *Core ICT Indicators*, http://www.itu.int/en/ITU-D/Statistics/Pages/coreindicators/default.aspx.

Partnership on Measuring ICT for Development (2011), *Measuring the WSIS Targets: A statistical framework*, http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wsistargets2011.aspx.

Poblete, B. *et al*. (2011), *Do All Birds Tweet the Same?: Characterizing Twitter Around the World*, paper delivered at the ACM Conference on Information and Knowledge Management, Glasgow, 2011, http://www.ruthygarcia.com/papers/cikm2011.pdf.

UNCTAD (United Nations Conference on Trade and Development) (2013), *Information Economy Report 2013: The Cloud Economy and Developing Countries*, http://unctad.org/en/PublicationsLibrary/ier2013_en.pdf.

UNESCWA (UN Economic and Social Commission for Western Asia) (2013), *Regional Profile of the Information Society in the Arab Region, 2013*, http://www.escwa.un.org/information/pubaction.asp?PubID=1492.

UNESCO (United Nations Educational, Scientific and Cultural Organization) (2003), *Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace, 2003*, http://portal.unesco.org/en/ev.php-URL_ID=17717&URL_DO=DO_TOPIC&URL_SECTION=201.html.

UNESCO (2013), *Global media and information literacy assessment framework: country readiness and competencies, 2013*, http://unesdoc.unesco.org/images/0022/002246/224655e.pdf.

UNESCO (2014), *Education for All Global Monitoring Report 2013/4*, http://unesdoc.unesco.org/images/0022/002256/225660e.pdf.

UNESCO and EURid (2013), *World Report on IDN Deployment 2013*, http://www.eurid.eu/files/publ/insights_2013_idnreport.pdf.

Endnotes

[1] The term 'world languages' is understood here to mean all languages used in the world today, rather than the small number of languages which are extensively used worldwide.

[2] Sam Costello, 'How Many Apps Are in the iPhone App Store', http://ipod.about.com/od/iphonesoftwareterms/qt/apps-in-app-store.htm, accessed 5 March 2014.

[3] See http://www.appbrain.com/stats/number-of-android-apps, accessed 5 March 2014.

[4] UNESCO, 2013,. "Digital literacy" is defined in WTDR 2010, p.190 as "… equipping people with ICT concepts, methods and skills to enable them to use and exploit ICTs"; "information literacy" as "… providing people with concepts and training in order to process data and transform them into information, knowledge and decisions" including "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."

[5] UNESCO, 2014, p. 70.

[6] Data from Ethnologue, http://www.ethnologue.com/statistics/size; http://www.ethnologue.com/country/PG.

[7] See http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/official_documents/Eng%20-%20Recommendation%20concerning%20the%20Promotion%20and%20Use%20of%20Multilingualism%20and%20Universal%20Access%20to%20Cyberspace.pdf.

[8] The text of the *Declaration* can be found at http://www.itu.int/dms_pub/itu-s/md/03/wsis/doc/S03-WSIS-DOC-0005!!PDF-E.pdf.

[9] The full remit for Action Line C8 can be found in World Summit on the Information Society, Geneva Plan of Action, 2003, para. 23, http://www.itu.int/dms_pub/itu-s/md/03/wsis/doc/S03-WSIS-DOC-0005!!PDF-E.pdf.

[10] WTDR, 2010, p. 189, http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtdr2010.aspx.

[11] UNESCO, OECD and ISOC, *op. cit.*, p. 36.

[12] *ibid*, pp. 12–13.

[13] See, for example, http://www.southafrica.info/about/people/language.htm#.UxXAtf3iufk.

[14] WTDR 2010, p. xxxi.

[15] These are the Democratic People's Republic of Korea and the British Indian Ocean Territory: see http://www.ethnologue.com/statistics/country.

[16] WTDR 2010, p. 178.

[17] For example, by the market research firm the International Data Corporation – see John Gantz and David Reinsel, 2011.

[18] A note on Netcraft's methodology can be found at http://www.netcraft.com/active-sites/.

[19] Data sourced from Facebook, published at http://www.theguardian.com/news/datablog/2014/feb/04/facebook-in-numbers-statistics. Data sourced from Facebook.

[20] Twitter data at https://about.twitter.com/company, accessed 6 March 2014; Tencent Weibo data from Data from http://www.go-globe.com/blog/social-media-china/.

[21] See http://www.go-globe.com/blog/social-media-china/; http://www.techinasia.com/social-media-and-social-marketing-china-stats-2013/.

[22] YouTube data at http://www.youtube.com/yt/press/statistics.html.

[23] Kenya's Open Data Initiative, supported by the World Bank, is described at https://opendata.go.ke/.

[24] American Standard Code for Information Interchange.

[25] Microsoft information at http://support.microsoft.com/kb/292246 and http://windows.microsoft.com/en-GB/windows/language-packs#lptabs=win7.

[26] Information from browser websites.

[27] Twitter data reported at http://mashable.com/2013/12/17/twitter-popular-languages, accessed 6 March 2014.

[28] Information on Google Translate from http://translate.google.co.uk/about/intl/en_ALL/ and http://translate.google.com/manager/website/?hl=en.

[29] Examples of this kind of work at the Oxford Internet Institute can be found at http://geography.oii.ox.ac.uk.

[30] For details of the indicator as planned, see Partnership (2011), p. 82.

[31] *ibid*.

[32] Report (to ECOSOC) of the Partnership on Measuring Information and Communication Technologies for Development, March 2012, para. 28, http://unstats.un.org/unsd/statcom/doc12/2012-12-ICT-E.pdf.

[33] National census forms are collected at http://unstats.un.org/unsd/demographic/sources/census/censusquest.htm.

[34] The questionnaire can be found at http://www.researchictafrica.net/docs/HH_Master_Questionnaire.pdf.

[35] This indicator is summarized in the 2011 WSIS statistical framework, p. 83.

[36] WTDR 2010, p. xxxi; source unidentified.

[37] The Globalstat data set for this date is still available online, at http://web.archive.org/web/20041019013615/www.global-reach.biz/globstats/index.php3.

[38] Brief discussions of the sources and methodologies used can be found with the data at online locations cited above.

[39] This indicator is summarized in the 2011 WSIS statistical framework, p. 83.

[40] Netcraft, January 2014 Web Server Survey, http://news.netcraft.com/archives/2014/01/03/january-2014-web-server-survey.html, accessed 6 March 2014.

[41] Data from the web technology analyst Web3Tech, covering the top ten million websites, reported at http://w3techs.com/technologies/overview/content_language/all, viewed 7 April 2014. The methodology behind this and other language figures reported is unclear.

[42] This was initiated and coordinated by the University of Technology in Nagaoka, Japan, see http://gii2.nagaokaut.ac.jp/gii/blog/lopdiary.php/lopdiary.php?catid=109&blogid=8.

[43] Information from Daniel Pimienta.

[44] Data from ZookNIC, see below.

[45] UNESCWA, 2013, p. 106.

[46] As reported at http://ptgmedia.pearsoncmg.com/images/9780789747884/supplements/9780789747884_appC.pdf. Data from June 2010.

[47] This section of the chapter has been written in conjunction with Matthew Zook of ZookNIC.

[48] Details of this indicator can be found in the 2011 WSIS statistical framework, p. 84.

[49] The registration process is often conducted though intermediary organisations or businesses known as registrars, which are accredited by the relevant registry.

[50] Data from CENTR, Domain Wire, edition 6, December 2013, available at https://centr.org/system/files/agenda/attachment/domainwire_stat_report_2013_3_0.pdf.

[51] Not always as some countries have both Latin and IDN ccTLDs.

[52] These are AfTLD (Africa), APTLD (the Asia-Pacific region), CENTR (Europe), LACTLD (Latin America and the Caribbean).

[53] Current data are published at http://www.whois.sc/internet-statistics/country-ip-counts/.

[54] These are .com, .net, .org, .biz, .info and .mobi.

[55] Data for this chart have been supplied by ZookNIC, compiled from ccTLD, Whois and other sources.

[56] *ibid*.

[57] A list of other ccTLDs marketed in this way can be found in http://geography.oii.ox.ac.uk/#geography-of-top-level-domain-names.

[58] CENTR, *op. cit*.

[59] These are listed at http://en.wikipedia.org/wiki/List_of_Google_domains.

[60] ZookNIC's published data can be found at http://www.zooknic.com/.

[61] Information from ZookNIC.

[62] See http://en.wikipedia.org/wiki/List_of_Internet_top-level_domains.

[63] Information from ZookNIC.

[64] The assistance of Erik Zachte and Tilman Bayer, Data Analyst and Senior Operations Analyst, respectively, for the Wikimedia Foundation, is acknowledged in the preparation of this subsection.

[65] The indicator is described in the 2011 WSIS statistical framework, p. 85.

[66] These are listed, with approximate numbers of articles at March 2014, at http://meta.wikimedia.org/wiki/List_of_Wikipedias.

[67] See http://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm and http://stats.wikimedia.org/EN/TablesArticlesTotal.htm. Data on the popularity of websites globally and by country are published by the web information company Alexa at http://www.alexa.com/topsites/global and http://www.alexa.com/topsites/countries. Alexa uses a selective toolbar-based methodology for data-gathering which has significant limitations.

[68] This information is available on a variety of sites, a useful portal being http://stats.wikimedia.org/EN/Sitemap.htm.

[69] The most recent data can be found at http://meta.wikimedia.org/wiki/List_of_Wikipedias. The figures in this paragraph were viewed on 19 March 2014.

[70] Other data sets published by the Wikimedia Foundation may allow some adjustments to be made, though it will be difficult entirely to remove bot-generated content.

[71] Data up to 2010 for all languages and up to date for the majority of languages can be found at http://stats.wikimedia.org/EN/TablesDatabaseWords.htm.

[72] This research is reported at http://www.tracemedia.co.uk/terra/.

[73] Data by country are published at http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdown.htm.

[74] Wikipedia data from http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryOverview.htm.

[75] Alexa publishes these data at www.alexa.com/topsites. A list of countries can be found at http://www.alexa.com/topsites/countries. Historic data are available on commercial terms but have not been accessed for this report.

[76] Information confirmed in correspondence with Alexa.

[77] Mobile browser market shares are reported at http://www.netmarketshare.com/browser-market-share.aspx?qprid=0&qpcustomd=1, viewed 19 March 2014. Opera Mini is particularly popular in India, Indonesia, Russia, China and Brazil: see http://www.buzzom.com/2012/06/opera-mini-7-browser-now-launched-on-feature-phones-and-blackberry-devices/.

[78] These are the search engine Baidu, the messaging and multipurpose site QQ and the online marketplace Taobao.

[79] See http://www.go-globe.com/blog/social-media-china/, viewed 7 April 2014.

[80] See Alexa data at http://www.alexa.com/topsites/countries; http://www.statisticbrain.com/facebook-statistics/; http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/#.Uw-G0Pl_vNk.

[81] See http://www.oneskyapp.com/blog/top-10-languages-with-most-users-on-facebook/ and http://www.socialbakers.com/blog/1064-top-10-fastest-growing-facebook-languages. Data sources unspecified.

[82] See http://www.oneskyapp.com/blog/language-breakdown-for-the-top-5-facebook-countries-outside-us/. Data source unspecified.

[83] The report derives reported findings from Dubai School of Government, 2011, p.15.

[84] See http://www.insidefacebook.com/2010/05/26/facebooks-latest-language-data-country-by-country/.

[85] Data in this paragraph are from http://www.statisticbrain.com/twitter-statistics/.

[86] Digital OnLine Life and You.

[87] Some data from the DOLLY project are published at http://www.floatingsheep.org/p/dolly.html.

[88] Delia Mocanu *et al*., 'The Twitter of Babel: Mapping World Languages through Microblogging Platforms', PLOS ONE, available at http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0061981.

[89] See https://www.youtube.com/yt/press/en-GB/statistics.html.

[90] See http://expandedramblings.com/index.php/important-instagram-stats/#.Uw-IiPl_vNk, viewed 7 April 2014.

[91] See http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer.

[92] Data from Comscore at http://www.comscoredatamine.com/2012/06/1-in-4-internet-users-access-banking-sites-globally/. Usage rates for the total population are, therefore, even lower in the Middle East and Africa than in North America and Europe because of the lower proportions of the total population in the former regions that are currently online.

[93] Sam Costello, 'How Many Apps Are in the iPhone App Store', http://ipod.about.com/od/iphonesoftwareterms/qt/apps-in-app-store.htm, accessed 5 March 2014.

[94] See http://www.appbrain.com/stats/number-of-android-apps, accessed 5 March 2014.

[95] ITU statistics at http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2013/ITU_Key_2005-2013_ICT_data.xls.

[96] *ibid*.

[97] For a discussion of this, see UNCTAD, 2013.

[98] Data from the web technology analyst Web3Tech, covering the top ten million websites, reported at http://w3techs.com/technologies/overview/content_language/all. The methodology behind this and other language figures reported is unclear.

[99] Information on Google Translate from http://translate.google.co.uk/about/intl/en_ALL/ and http://translate.google.com/manager/website/?hl=en.

[100] As automated translation becomes more widespread, increased care will be needed to ensure that algorithms do not treat existing automated translations as equivalent source material to manual translations.