OECD publishing

Please cite this paper as:

Bruegge, C. (2011-12-08), "Measuring Digital Local Content", *OECD Digital Economy Papers*, No. 188, OECD Publishing, Paris.
http://dx.doi.org/10.1787/5kg0s294n9kf-en

OECD Digital Economy Papers No. 188

Measuring Digital Local Content

Chris Bruegge

OECD DIGITAL ECONOMY PAPERS

The OECD's Directorate for Science, Technology and Industry (STI) undertakes a wide range of activities to better understand how information and communication technologies (ICTs) contribute to sustainable economic growth, social well-being and the overall shift toward knowledge-based societies.

The OECD Digital Economy Papers series covers a broad range of ICT-related issues, both technical and analytical in nature, and makes selected studies available to a wider readership. It includes *policy reports*, which are officially declassified by an OECD committee, and occasionally *working papers*, which are meant to share early knowledge and elicit feedback. This document is a working paper.

Working papers are generally only available in their original language – English or French – with a brief summary in the other. The opinions expressed in these papers are the sole responsibility of the author(s) and do not necessarily reflect those of the OECD or of the governments of its member countries.

STI also publishes the OECD Science, Technology and Industry Working Paper series, which covers a broad range of themes related to OECD's research and policy work on knowledge-based sources of economic and social growth and, more specifically, on the translation of science and technology into innovation.

OECD Digital Economy Papers and
STI Working Papers are available at:
www.oecd.org/sti/working-papers

MEASURING DIGITAL LOCAL CONTENT

Chris Bruegge*

ABSTRACT

This paper discusses the ways to quantify the local content that can be delivered through the internet. Several indicators are proposed; for each indicator the paper discusses available data, presents strengths of a given measure and outlines its potential drawbacks.

RÉSUMÉ

Cet article discute les méthodes appliquées pour quantifier le contenu local qui peut être fourni par le biais d'Internet. Plusieurs indicateurs sont proposés. Pour chaque indicateur le rapport discute les données disponibles, présente les points forts et les désavantages potentiels d'une mesure donnée.

*E-mail: *chris.bruegge@gmail.com*. The views expressed in this paper are those of the author and should not be attributed to the OECD, or its member countries.

TABLE OF CONTENTS

INTRODUCTION

Local content is taking an increasingly prominent place on the Internet. A UNECSO report indicated that close to half of the nearly 400 million Internet users spoke English as a primary language in 2001 (UNESCO, 2001.) Recent studies indicate that the proportion of English speakers relative to other language speakers on the Internet is declining rapidly. In 2010, a market research firm estimated that only 536 million of nearly 2 billion Internet users (~27%) were native English speakers.[1] With this movement away from a dominant, unifying Internet language, local language content has proliferated. Despite the increasing importance of local language digital content, efforts to systematically identify and quantify local content are scarce. This paper makes inroads into the studies of local content by providing analysis of measures which can be used to quantify local content.

There is no accepted uniform definition of local content. For the purpose of this study we will rely on a UNESCO report for the International Telecommunication Union which states that local content must be "understandable and appreciated by local users" (UNESCO, 2001.) In harmony with this description, this paper considers all digital content created for an end user who speaks the same language as the author to be local content. This includes content created for people who do not live in close proximity to the creator, but who thanks to the Internet are part of a world-wide 'local' community of same-language speakers. The language criterion is primarily intended to exclude translated content. No stipulations about the author of local content are made (*i.e.* individuals, governments, and businesses all qualify).

Because local content cannot be measured directly, a set of measures must be used to indirectly infer its size. The remainder of the paper will provide a discussion of the merits of ten measures for local content, as well as challenges associated with the use of these measures. First, we will discuss measures for local content collected by economy. Following the presentation of these measures, we will look at measures related to language. Where a time series or panel data is available, we will also provide a discussion of recent trends or a graphical look at local content creation in 2010 respectively. Finally we give a few concluding remarks.

Quantitative measures of local content

To facilitate classification of local content, it is helpful to divide measures into two groups: Measures which are *i)* associated with a particular economy; and *ii)* measures tied to a particular language. In the first case, local content is attributed to the economy where it was created, regardless of the economy of origin of the author. Measures collected on a language by language basis are assigned to economies according to the proportion of speakers of that particular language residing in the economy. These criteria transform the impossibly complex and subjective problem of identifying and classifying local content into a tractable exercise. Tables 1 and 2 summarise the discussed measures.

Table 1: Summary of measures of local content (by economy)

Indicator	Description	Benefits	Potential Drawbacks
Measures by economy			
ccTLDs ("Country Code Top-Level Domains")	Number of "country code top-level domains" per 1 000 residents per economy	No ambiguity in identification, good fit for local content criteria	Narrow measure of local content due to barriers to entry
Facebook Subscribers	Number of Facebook subscribers per 1 000 residents per economy	Popular (ranked no. 2 site worldwide by Alexa.com), platform for local advertisers	Unavailable in certain areas, substitute products exist, biased towards youth
Online Newspapers	Number of online newspapers per 1 million residents per economy	Measures professional content creation	Varying popularity across economies due to presence of substitute products such as blogs
Streaming Radio Stations	Number of streaming online radio stations per 1 million residents per economy	Good source of local news, language, and cultural media	Regulatory differences between economies, potential variation in the amount of foreign aid used to support local radio across economies
Geotagged Flickr Photos	Number of Flickr photos geotagged per 1 000 residents per economy.	Measures a unique niche of local content not captured by the other measures	Includes photos taken by tourists and other non-locals (*i.e.* photos not intended for a local audience)
YouTube Uploads	Number of YouTube uploads per 1 000 residents per economy	Popular (ranked no. 3 site worldwide by Alexa.com), primarily user-created	Might be biased towards certain cultures

Table 2: Summary of measures of local content (by language)

Indicator	Description	Benefits	Potential Drawbacks
Measures by language*			
Web pages	Number of web pages per language	Broad measure of local content relative to ccTLDs	Classification difficulties, language overlap between economies
Wikipedia Articles	Number of Wikipedia articles per language	Free and easily accessible, reflective of overall shift of the Internet community away from English language	Easy to automate creation of articles
Blogs	Number of blogs per language	Free, accessible to all with Internet access	Measured imprecisely, classification of multilingual blogs is ambiguous
Tweets	Number of Tweets per language	Low overhead, anecdotal evidence showing Twitter communities of minority languages	Measured imprecisely, potential bias due to use of lack of text message vocabulary in given language

Notes: * These measures should be weighted by the number of speakers of the particular language per economy.

Measures by economy

Number of "country code top-level domains"

The "country-code top-level domains" (ccTLDs) are two-letter top-level domains especially designated for a particular economy, country or autonomous territory to use to service their community.[2] "Country code top-level domains" are often used by community-oriented organizations, businesses, and even official town websites. Currently there are 324 ccTLDs listed on the IANA website.[3]

Although the criteria for local content permit a broad range of media to qualify, the spirit of local content is community. Whether local content be user-created, business-created, or government-created, it is intended to draw local readership and promote local language and culture. With this in mind, sites which serve and strengthen the community (*i.e.* ccTLD sites) are ideal measures for local content.

Potential drawbacks

There are drawbacks to using "country code top-level domains" as a measure for local content creation. The registration process for ccTLDs is not uniform across economies. For example, the application process is longer and more costly in some economies than others. These barriers to entry may keep many local content providers out of the market for ccTLDs, and drive them instead to alternate channels of dissemination. Hence, using ccTLDs as a metric for local content only captures a narrow band in the spectrum of local content.

Additionally, some economies have decided to allocate the rights to their ccTLD to third parties. Tuvalu (.tv) and the Federated States of Micronesia (.fm) have taken advantage of the commercial interest in the abbreviations of their ccTLDs.[4] The OECD provides a specific example of an organisation using a ccTLD from outside its own economy of presence. Currently, the OECD uses the .cd domain from the Democratic Republic of the Congo to shorten some of its URLs to oe.cd. In areas where ccTLD rights have been sold, ccTLDs indicate an absence, rather than a presence of local content.

Another bias is introduced by sites which have a ccTLD but provide translations in many different languages. Sites such as *www.rfi.fr* (an online French radio station) provide multiple language versions including *www.english.rfi.fr*; sites in translation such as this one do not meet the language requirement for local content. Nonetheless, they cannot be easily separated from legitimate local content on a large-scale basis.

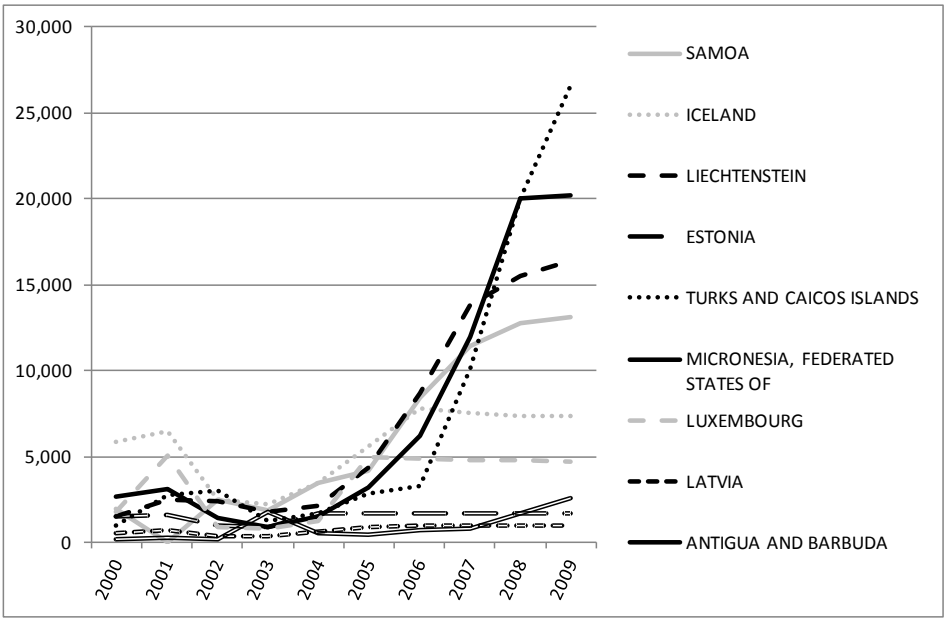
Available data

Using a Google wildcard search, it is possible to identify the number of indexed web pages with a particular "country-code top-level domain" in 246 economies from 2000-2010 (Please refer to Figure A1 in the Annex for search parameters). The ccTLD for a particular economy is unambiguous to define. This eliminates a large source of measurement error in quantifying local content. The drawback of using the Google platform is that the search algorithm is proprietary, and hence the methodology non-transparent. The algorithm seems to find more results for searches which are conducted on a more frequent basis. Additionally, searches performed months apart return very different results. For this reason, all data should be gathered at the same time.

During the past ten years, the median growth in indexed sites with a particular ccTLD was 40 % per year in the 246 economies contained in our data. At the median, the number of indexed pages per ccTLD doubles every 25 months. Average growth over the same period was an astounding 3202 % per year. The average is distorted by small economies which increase from only a handful of sites to hundreds or even

thousands in just a year. Figure 1 provides a graphical representation of the growth in the number of ccTLDs (per 1000 inhabitants) in the top-7 economies.

Figure 1. Top-7 economies (ccTLD per 1000 inhabitants

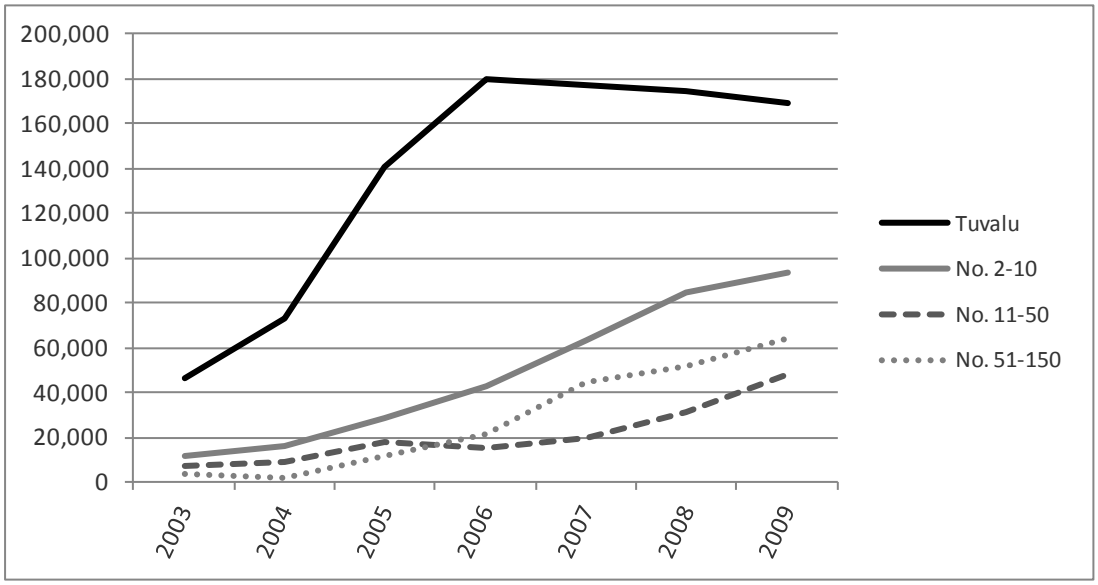


Notes: Tuvalu has more ccTLDs per capita than the next 50 economies combined. For scaling purposes Tuvalu was not included on the graph. Please refer to Figure 2 below for a graphical depiction of Tuvalu compared with other economies.

Source: Google.com

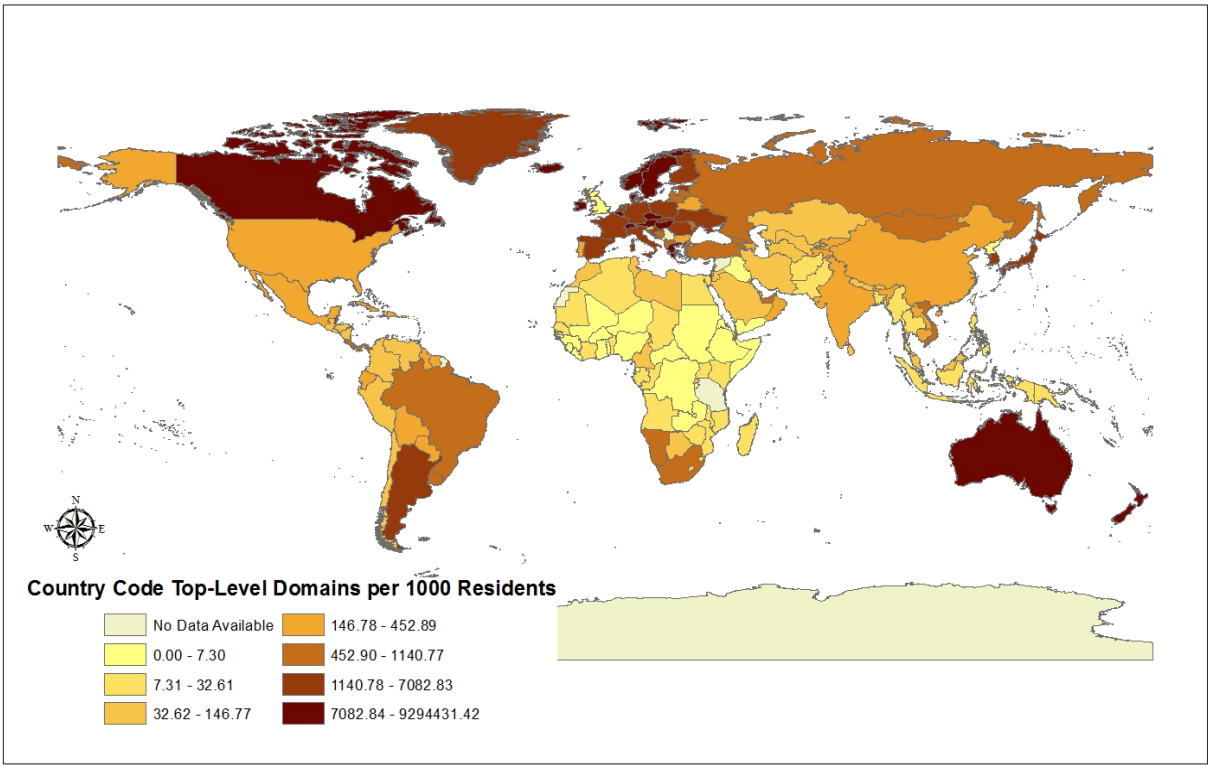
The distribution of ccTLDs per 1 000 residents is extremely skewed to the right (meaning that the average is much greater than the median). This fact should be addressed when doing econometric analysis (*i.e.* least absolute deviation regressions might be preferable to ordinary least squares). Alternatively, outliers occurring because the particular economy has sold rights to its ccTLD could be omitted from the analysis. A good example of an outlier in this category is Tuvalu, with more ccTLDs per capita than the next 50 economies combined. Please refer to Figure 2 below for a graphical depiction of Tuvalu compared with other economies.

Figure 2. Impact of outliers –"country code top-level domains" per 1000 inhabitants



Source: Google (ccTLD), Worldbank (population)

Figure 3. "Country code top-level domains" per 1000 residents



Note: .gb ccTLD used in the United Kingdom rather than .uk

Source: Google.com,

Table 3. "Country code top-level domains" in the OECD countries

Country Name	ccTLD	Country Name	ccTLD
Australia	au	Japan	jp
Austria	at	Korea	kr
Belgium	be	Luxembourg	lu
Canada	ca	Mexico	mx
Chile	cl	Netherlands	nl
Czech Republic	cz	New Zealand	nz
Denmark	dk	Norway	no
Estonia	ee	Poland	pl
Finland	fi	Portugal	pt
France	fr	Slovak Republic	sk
Germany	de	Slovenia	si
Greece	gr	Spain	es
Hungary	hu	Sweden	se
Iceland	is	Switzerland	ch
Ireland	ie	Turkey	tr
Israel	il	United Kingdom	uk
Italy	it	United States	us

Source: *www.iana.org/domains/root/db/#*

Facebook subscribers per economy

Facebook is a social network service available in most economies around the world. It was launched in 2004 and has gradually expanded the cohort of people eligible to join. In 2004, Facebook was only available to students from a handful of U.S. universities. In 2005, this eligible user base was extended to include both United States and international high schools and colleges, and finally in 2006 Facebook permitted anybody over the age of 13 to join.

In May 2011, Facebook ranked among the top ten websites in all OECD countries (based on data collected from the web traffic ranking site Alexa.com). In many of these countries, it was the second most visited website, only trailing behind Google.

Facebook's popularity makes it a facilitator of exchange in local languages. For many individuals, Facebook provides a way to be connected to news and current events. In addition to personal Facebook pages, Facebook is also a low-overhead way for small businesses to advertise and get involved in the community. The popularity of Facebook and the amount of exchange in local languages which takes place there make the number of Facebook users a good measure for local content.

Potential drawbacks

Although it has many advantages, Facebook's primary drawback is that it is not used ubiquitously. Facebook is unavailable in a number of economies such as China, for example, and instead substitute social networking sites such as RenRen and QQ are used. China is not unique in this regard; other economies such as Vietnam[5] and Iran[6] block access to Facebook as well. In addition to the complications caused by heterogeneous uptake across countries, the age distribution of Facebook users also introduces bias. A study by Gallup indicates that individuals in the United States between the ages of 18 and 29 are most likely to have a Facebook page (73%), while uptake among older age groups is significantly less. In the 30-49 year old age group, uptake is estimated to be 55%, while in the 50-64 year-old and 65-plus age groups, user-ship is only about 33% and 17% respectively[7]. Because of the skewed age distribution of

Facebook users, Facebook as a measure of local content over-represents content created by young people. It is unclear from our data, however, whether this is a problem unique to Facebook, or whether all measures of online local content favour content created by youth. The latter would seem possible if internet usage in general was more popular among younger generations than older ones.

Using Facebook as a metric for local content also poses a taxonomical issue; it is unclear how to classify subscribers who register outside their economy of origin. These people often produce content in their native language to communicate with friends and family at home, but many often produce content in the language indigenous to their place of residence. In our classification, these users are counted in the statistics for their economy of residence, although it is likely that many are producing content which could be considered local content in their economy of origin.

In order for the Facebook data to be an unbiased measure for local content, we must make several assumptions. First, Facebook's popularity relative to other social networks must be the same across economies. In the OECD member economies, Facebook is the largest social network site; however, the proportion of people who use Facebook relative to MySpace or other social networking sites must be the same as well.

Additionally, by using the number of Facebook subscribers per economy as a metric for local content, we make no stipulation about whether the user is active. Some accounts are created and never used; according to our metric, these people contribute as much to local content as active users. If inactive accounts are not equally probable across economies, our results will be biased.

Available data

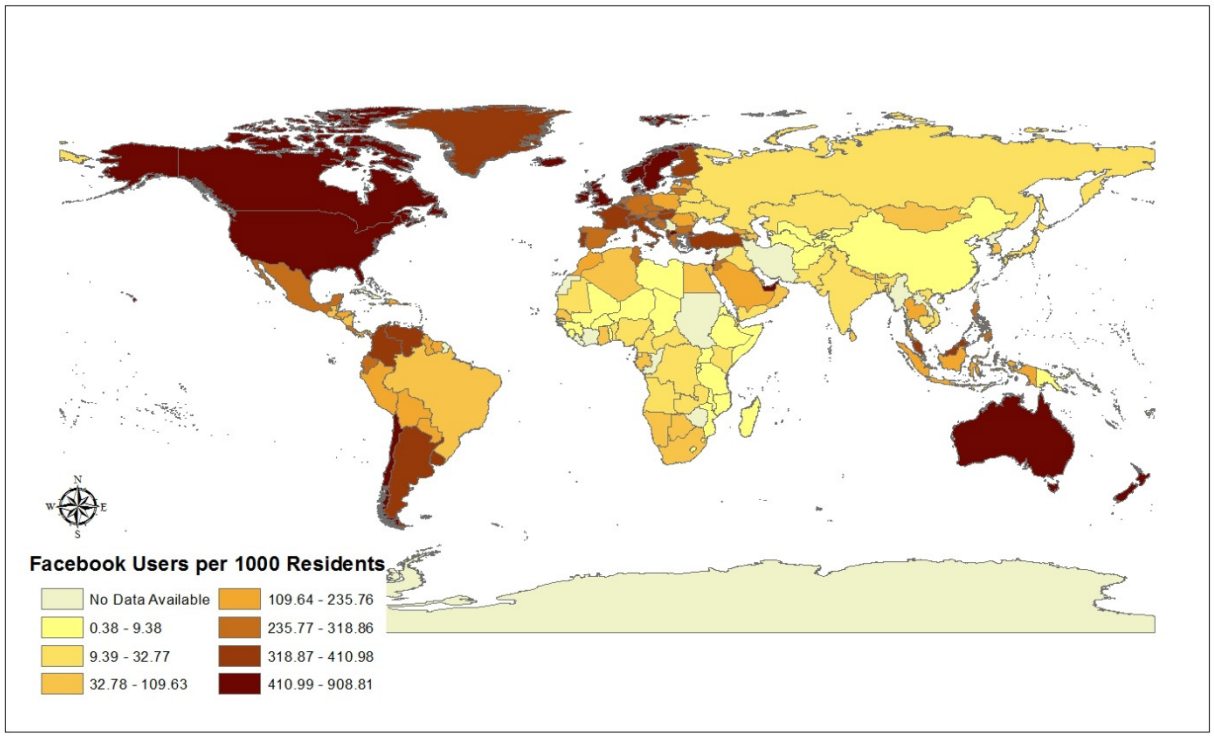
Facebook makes data on the number of users in each economy publically available through their advertising tool. According to Facebook.com, Facebook has over 750 million active users, 70 % of whom live outside of the United States[8]. Table 4 provides the number of users per 1 000 residents in the OECD member economies along with the Alexa.com in-economy ranking of Facebook relative to other sites.

Table 4. Facebook usership and web-traffic rankings in the 34 OECD member economies

Economy	Facebook users per 1000 inhabitants	Facebook ranking among websites in the economy	Economy	Facebook users per 1000 inhabitants	Facebook ranking among the websites in the economy
Australia	460	2	Japan	30	10
Austria	300	2	Korea	70	3
Belgium	420	2	Luxembourg	400	1
Canada	530	2	Mexico	220	1
Chile	490	1	Netherlands	270	3
Czech Republic	320	2	New Zealand	460	2
Denmark	500	2	Norway	550	1
Estonia	300	2	Poland	170	2
Finland	370	2	Portugal	360	2
France	360	2	Slovak Republic	340	2
Germany	230	2	Slovenia	330	3
Greece	300	1	Spain	320	2
Hungary	330	2	Sweden	460	2
Iceland	660	1	Switzerland	340	2
Ireland	460	3	Turkey	380	1
Israel	470	2	United Kingdom	510	2
Italy	330	2	United States	510	2

Source: Facebook.com (advertising tool), Alexa.com

Figure 4. Facebook subscribers per 1 000 residents[9]



Source: Facebook.com

Online newspapers per economy

Thousands of newspapers around the world provide some or all of their content online. As online sources of news crowd out printed news sources, online newspapers become an important provider of professionally-created news.

Local newspapers are the quintessential local content providers. Not only do they cover local news and culture, but they are generally written by professional staff and provide high-quality content. Transplanting this reliable source onto the Internet accurately represents an important facet of digital local content.

Potential drawbacks

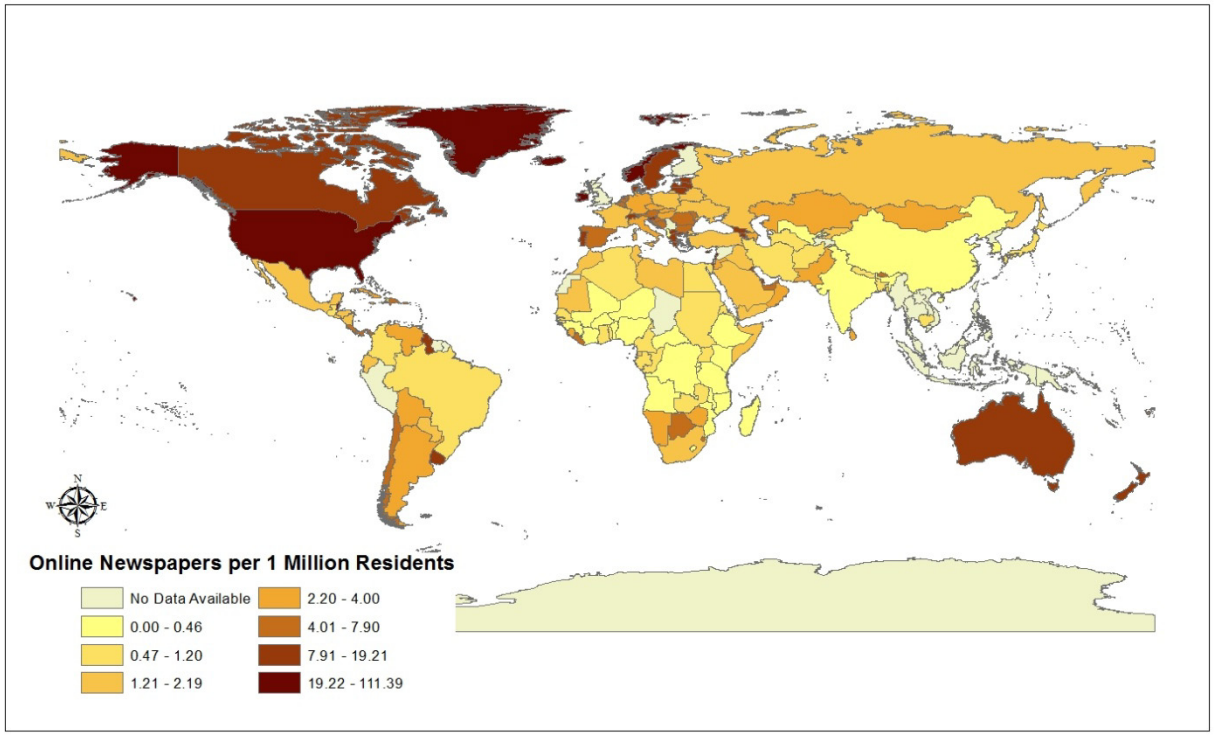
A potential source of bias using newspapers to measure local content is that the use of newspapers (print or online) might be more prevalent in certain regions than in others and may reflect other demographic characteristics such as literacy rates. If for example radio is more popular in economies with low population densities relative to areas with high population densities than this metric will be biased. Also, large newspapers might potentially have translated content in order to attract a wider readership.

Finally, the collection process for sites aggregating newspaper links may introduce bias into the data. For example, the site *www.onlinenewspapers.com* is largely monitored by readers who can edit data and add new papers to the list so the collection might favour areas with a more active user base.

Available data

Data on the number of online newspapers in an economy was collected from *www.onlinenewspapers.com*. This site provides links to online newspapers in hundreds of economies around the world. Because the site is frequently updated, it will be possible to slowly build a time-series of newspapers in the available economies.

Figure 5. Online newspapers per 1 million residents



Source: OnlineNewspapers.com

Online radio stations per economy

Radio is a vital source of information for millions of people around the world. In many parts of the developing world, radio has filled the information vacuum, giving locals access to knowledge which has directly improved their quality of life. Because of the perceived importance of radio, large amounts of aid money have been spent establishing and maintaining local radio stations. For example, a project sponsored by the Bill & Melinda Gates Foundation endeavours to use radio to provide information to farmers in developing economies, thereby increasing agricultural security. Academic research has shown that access to radio and other media in the developing world improves agricultural efficiency, public health, and other aspects of life such as social equality.[10]

As the Internet penetration in many developing areas increases, local radio content is finding a second home online. The growing number of online radio stations creates more opportunity for locals to get local information from increasingly diverse sources. Additionally, the Internet is lowering entry cost into the radio business, removing the need to install large broadcasting facilities.

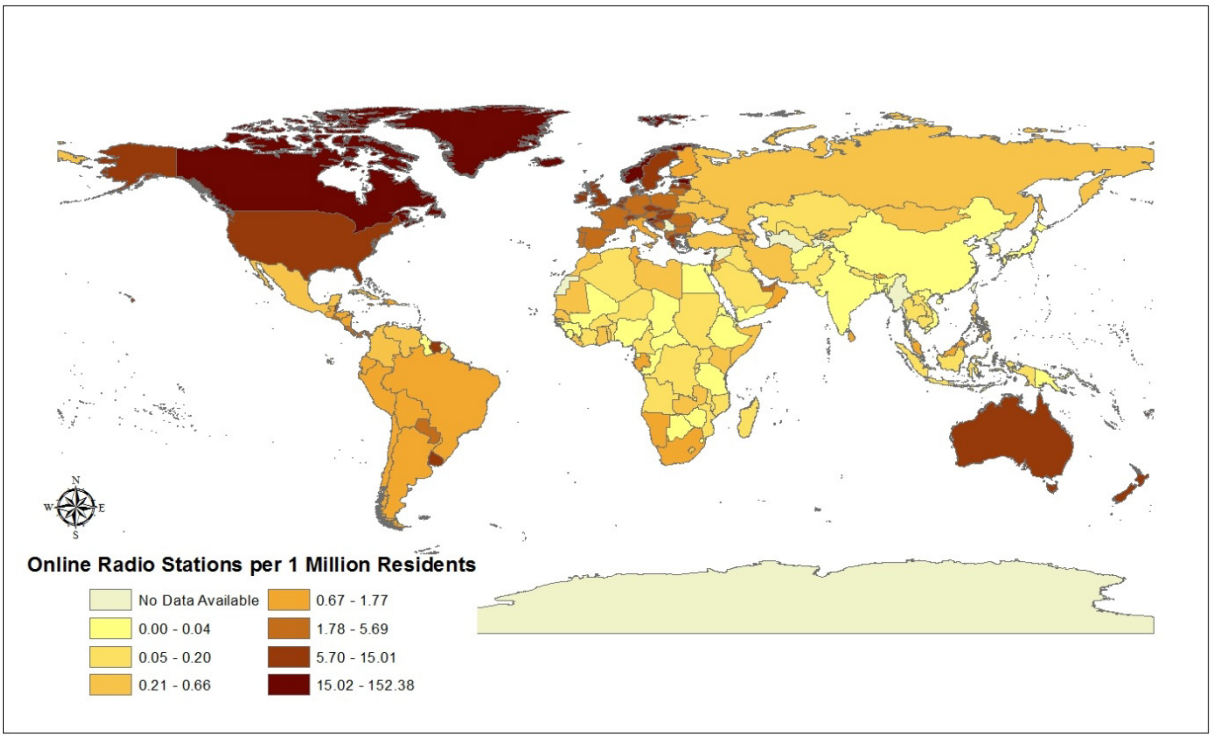
Potential drawbacks

Using the number of online radio stations as a measure for local content raises a few challenges. The ease with which international programming can be rebroadcast to an audience which does not speak the given language (especially true in the case of music) means that in some cases local radio might fail to qualify as purely local content. Additionally, regulatory differences across economies might make it relatively more difficult to start an online radio station in some places than in others. This would lead to systematic bias against the more heavily regulated areas.

Available data

Our data are drawn from two different sources: live-radio.net and radio-locator.com. In total information was collected for over 220 economies. Although minor differences exist between the two sources, both give a similar picture of the economies with many radio stations per capita and those with few. The average number of stations per million residents from Live-Radio.net is slightly higher (8.8) than from Radio-Locator.com (7.7). Because there are no systematic differences between the two sources, Figure 6 uses the more comprehensive Live-Radio.net data.

Figure 6. Online radio stations per economy



Source: Live-Radio.net

Geotagged Flickr photos per economy

Flickr is a photo sharing site which was launched in 2004. In addition to simply uploading and commenting on photos, Flickr permits users to "geotag" (attach geographic metadata to) photos. This is done simply by dragging and dropping photos onto the appropriate spot on a map, and does not require any special equipment or software.

Presently, there are over 150 million geotagged photos on Flickr. These photos catalogue a wide range of local art, architecture, geography, culture, and activities. Although many photographs might be taken by visitors, we include this measure in our analysis as photos are understandable and appreciable by locals. This is the only measure discussed in this paper where the language criterion is relaxed and the broader UNESCO description is used.

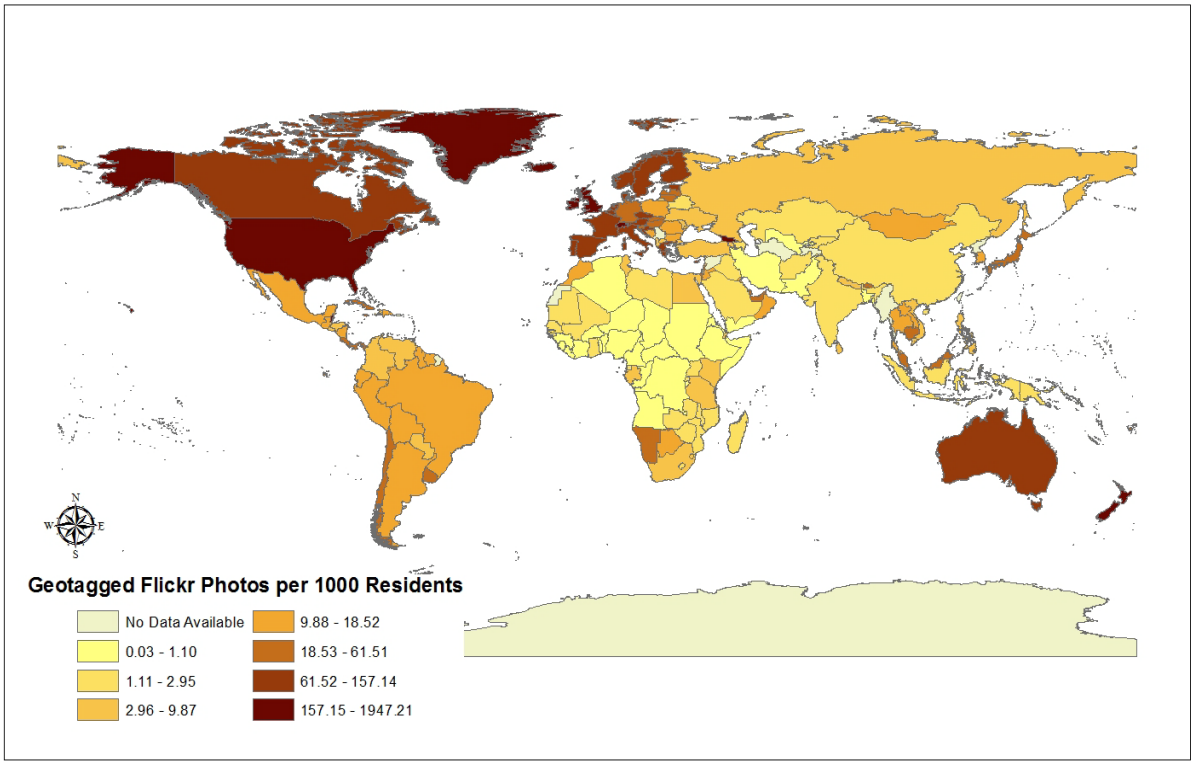
Potential drawbacks

The primary drawback of using Flickr is that there is no way to distinguish photos intended for and appreciated by locals, and photos taken by tourists to show off to friends. This might inflate the measured amount of local content in popular tourist destinations. Additionally, since not all photos are geotagged, we must assume that propensity to geotag photos in all economies is equal. Finally, we must assume that Flickr usage and, more generally, photography is equally popular in all economies.

Available data

Using the Flickr API we collected data on the number of geotagged photos in 221 economies. Although this data only represents a snapshot at the time this paper was written, it will be possible to build a monthly time series in the future.

Figure 7. Geotagged Flickr photos per economy



Source: Flickr.com

YouTube uploads per economy

YouTube is an online video sharing service launched in 2005. It permits individuals as well as corporations to upload videos, which can then be viewed by visitors to the site. In 2011, YouTube.com was ranked by Alexa.com as the third most popular website worldwide, behind Google and Facebook. Estimates from YouTube.com indicate that 70% of site traffic comes from outside the United States.[11] In

addition to home videos, YouTube currently contains content produced on major Television networks, including sports, news, and other programming. Presently, sharing and viewing YouTube content is free for non-commercial users.

Although YouTube has been intermittently criticised for allowing users to upload copyrighted content, much of the video available on the site can be classified as user-created content. A study by Kansas State University indicates that user-created content accounts for more than 80% of uploads to the site.[12]

Potential drawbacks

YouTube is blocked in a number of areas, making it an imperfect measure of local content. In China, other sites such as YouKu and TuDou are popular substitutes. In order to be an accurate measure for local content (which is comparable between economies), usage must be independent of location. In most areas, the popularity of YouTube makes this assumption plausible, but in areas where its usage is blocked or where substitute websites have a large market share, the estimate of local content will be biased downward.

Available data

Since YouTube does not release usership information, no comprehensive YouTube dataset is currently available. Nonetheless, several private entities have used statistical sampling techniques to approximate the number of YouTube uploads by economy. The Kansas State University study cited above collected data on the 20 most recent videos posted to YouTube every 2 hours for a 24-hour period. They estimate that the five economies which upload most actively to YouTube are the United States (34.5%), the United Kingdom (6.9%), the Philippines (3.9%), Turkey (3.4%), and Spain (3.4%).

Measures by language

Number of websites per language

According to OECD statistics, 16 % of OECD Internet users created a web page in 2010 (OECD, 2011). Although this figure represents less than a one percentage-point growth from 2005, it still is indicative of the fact that a large amount of content is being created by Internet users.

The number of web pages per language offers a few advantages over ccTLDs as a measure for local content. First, the number of websites per language encompasses a broader class of local content than ccTLDs. In addition to sites with a ccTLD, this metric captures .com, .net, .org, and other popular top-level domains which may contain local content. This permits the inclusion of a large amount of user-created content in addition to business and government-created content. Additionally, collecting data by language rather than by economy permits analysis of local content development in regions where languages other than the national language are spoken. Wonderful examples exist in China, India and the African continent, where hundreds of region-specific local dialects coexist with the national languages. National level data does not permit this finer-scale analysis.

Potential drawbacks

Although measuring local content on a language basis rather than a country basis provides several advantages, this technique is much more difficult than using ccTLDs. Rigorously defining and categorising websites by language is challenging, as some sites have content in two or more languages. This could lead to measurement error.

Another problem using measures collected on a language by language basis is that additional assumptions must be made before this content can be attributed to a particular region. To illustrate, assume there are 100 websites in a particular language, and that the language is spoken in two economies. In order to divide these 100 sites between the two economies, the researcher must know the propensity of a speaker of the given language in the first economy to create a website relative to a speaker of the same language in the second economy. For the purposes of this study, it is assumed that speakers of a given language in all economies are equally likely to create a website (or other local content). This permits the straightforward approach of assigning content to economies (or regions) based solely on the percentage of speakers residing there. Because this is a two-step process requiring additional data on speakers of a given language per economy, measurement bias might be exaggerated.

Available data

We have collected panel data on 43 languages from 2000-2010. The same difficulties with Google's proprietary algorithm exist in these data as in the ccTLD data. The proprietary search algorithm makes it impossible to verify the accuracy of the data collected.

Wikipedia entries by language

Wikipedia is a free, web-based, collaborative, multilingual encyclopaedia project supported by the non-profit Wikimedia Foundation. Its 18 million articles (over 3.6 million in English) have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone with access to the site. Wikipedia was launched in 2001 by Jimmy Wales and Larry Sanger and has become the largest and most popular general reference work on the Internet, ranking around seventh among all websites on Alexa and having 365 million readers.

Wikipedia's departure from the expert-driven style of encyclopaedia building and the large presence of unacademic content has been noted several times. Time magazine recognised "You" as its Person of the Year for 2006, citing Wikipedia as an example of online collaboration and interaction by millions of users around the world.

Although the policies of Wikipedia strongly espouse verifiability and a neutral point of view, critics of Wikipedia accuse it of systemic bias and inconsistencies (including undue weight given to popular culture), and allege that it favours consensus over credentials in its editorial processes[13]. Its reliability and accuracy are also targeted. Other criticisms centre on its susceptibility to vandalism and the addition of spurious or unverified information; however, scholarly work suggests that vandalism is generally short-lived. An investigation in Nature found that the science articles they compared came close to the level of accuracy of Encyclopædia Britannica and had a similar rate of "serious errors."[14]

Wikipedia's user-created nature makes it a nice measure for local content. There are no barriers to entry for individuals wishing to post information on Wikipedia, and Wikipedia articles are reflective of the overall Internet trend away from a single dominant language. Figure 3 shows the evolution of the share of English Wikipedia articles relative to other languages from 2001 to 2010.

Potential drawbacks

Creation of Wikipedia articles is easily automated by computer programs. While this makes constructing and updating of dynamic pages (*i.e.* pages containing frequently changing population figures) much more efficient, it also permits large-scale creation of pages with very little information. Box 1 describes an extreme example of using a computer program to distort the ratio of Wikipedia articles to speakers of a language. A better way to measure contribution to local content on Wikipedia might be to factor in the average length of articles or the number of contributors for each language.

Box 1. Wikipedia articles in Volapük

Volapük is a constructed language, created in 1879-1880 by a catholic priest named Johann Martin Schleyer. Although once popular in Paris and Munich, today it is estimated that there are only 20 speakers of Volapük worldwide. Nonetheless, Volapük speakers have been very prolific in their contributions to Wikipedia; in 2010, there were over 118,000 Wikipedia articles written in Volapük, or nearly 6,000 for every speaker. Most of these articles were created by a single user who utilized a computer program to automatically create Volapük-based stubs (incomplete articles) primarily based on existing census databases. Ostensibly, these pages were created to draw public attention to this dying language. Other languages on Wikipedia have also witnessed a proliferation of computer-generated Wikipedia articles.

Source : *http://en.wikipedia.org/wiki/Volapuk*

Available data

We have collected panel data from Wikipedia.org which contains information on 274 languages represented on the site. The data are recorded at monthly intervals from 31 January, 2001 until 31 December, 2010. The number of languages and the monthly 10-year time series is the ideal dataset to perform econometric analysis. Table 5 below contains summary statistics for our data.

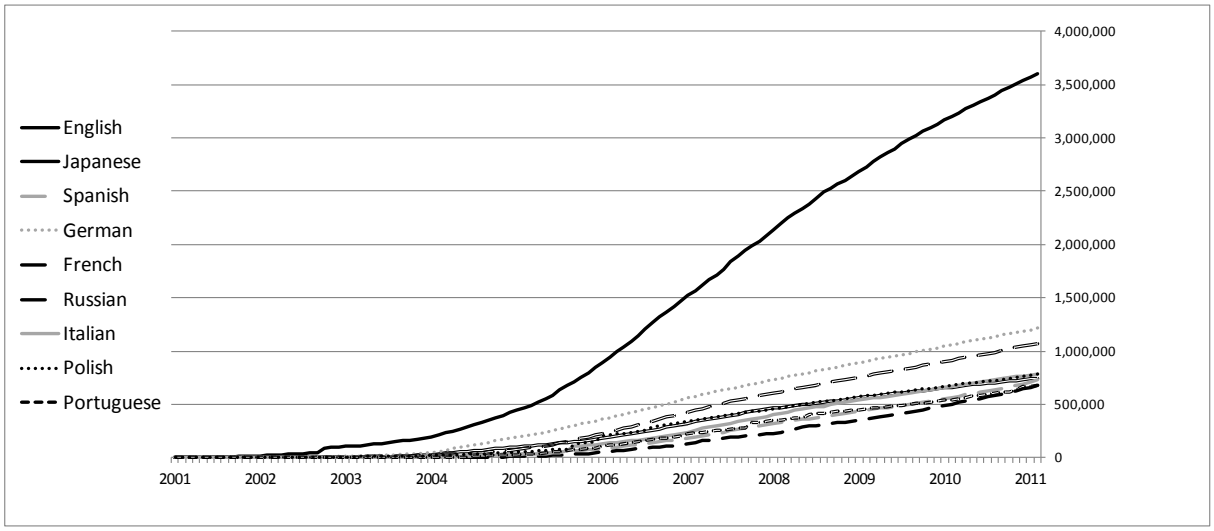
Table 5. Wikipedia summary statistics

Date	Number of languages	Mean no. of articles	Std. Dev of no. of articles	Min. no. of articles	Max. No. of Articles
Dec-10	273	64 591.0	260 045.5	11	3 536 986
Dec-09	273	53 782.5	226 526.2	3	3 127 804
Dec-08	272	43 987.5	190 933.9	1	2 643 417
Dec-07	267	34 286.4	151 679.3	1	2 082 419
Dec-06	256	23 044.6	108 500.6	1	1 468 828
Dec-05	213	14 121.5	67 415.2	1	844 743
Dec-04	180	7 042.1	35 860.8	1	423 505
Dec-03	103	3 852.0	19 016.5	1	184 817
Dec-02	46	2 956.6	14 494.7	1	98 197
Dec-01	15	1 211.3	4 069.8	3	15 893

Source: Wikipedia.org

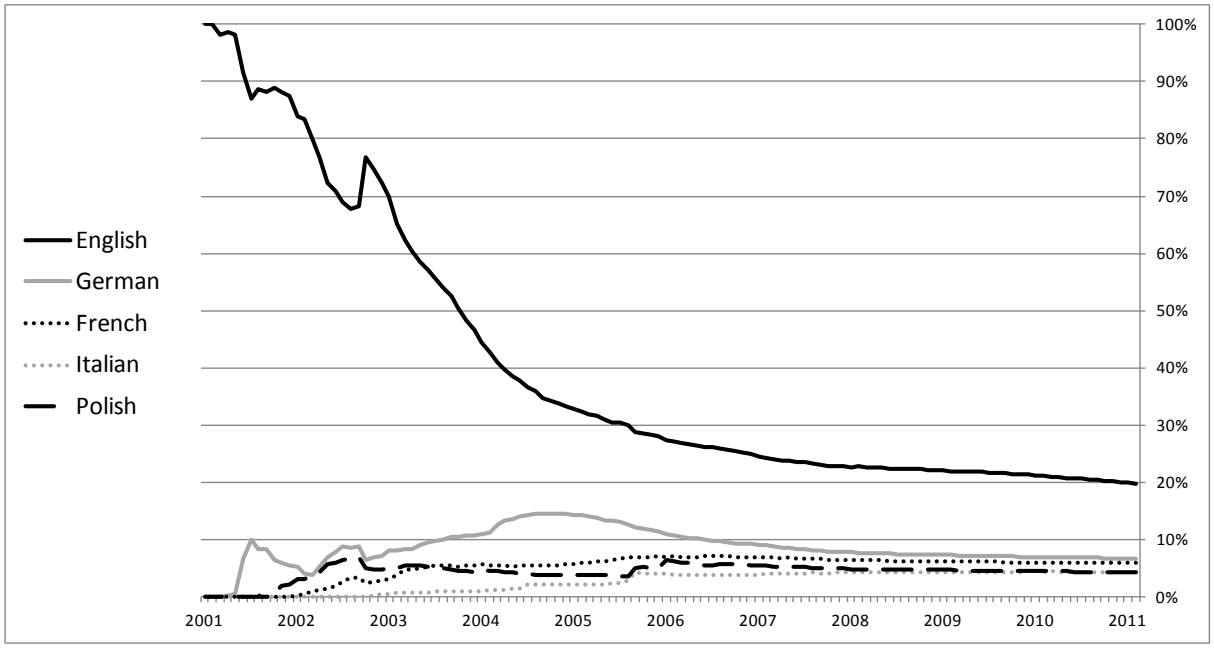
As mentioned in the introduction, the population of English-speaking Internet users relative to other languages has been declining. Wikipedia mirrors the overall move of Internet users away from English, as depicted by the figure below. For the first few years after its inception in 2001, Wikipedia was primarily dominated by articles in English. By 2010, only about 20% of Wikipedia articles were in English, while it was estimated that 27% of Internet users were English speakers.[1]

Figure 8. Total Wikipedia articles by language



Source: Wikipedia.org

Figure 9. Proportion of Wikipedia articles by language (Top-5 languages)



Source: Wikipedia.org

The median growth for all of the languages in the data is 89% per year; this translates to a doubling time of just under 13 months. Average growth statistics, especially for languages with a small number of articles might be misleading as the addition of just a few articles could potentially translate into growth of several hundred percent.

Blogs by language

Although the precise definition of the word 'blog' is still disputed, loosely defined a blog is a website containing date-stamped entries in reverse-chronological order (OECD, 2007). Blogs serve many purposes, including sharing information (*e.g.* news blogs, political blogs, etc), a platform for self-expression (*e.g.* personal blogs containing journal-like content), and social networking (*e.g.* interest blogs) (OECD 2006).

A large number of blogs are written by academic or professional experts, and are considered credible sources of information. For example, Paul Krugman, a Nobel Prize-winning economist, maintains a blog which was ranked the 69th most popular blog in May 2011 by a blog search engine.[15] Krugman's blog generates large amounts of commentary and engenders frequent debate. In addition to this type of informational blog, many blogs are used by local and national officials to reach out to the public (the official Whitehouse blog in the United States was also in the top 100).

According to the 2006 OECD Information Technology Outlook, blogs were among the most important early developments in the participative web (OECD, 2006). As an easily available, all-purpose platform for expression, blogs lend themselves well to the dissemination for local content. In fact, creating a blog does not require any special software and is as easy as using a word processor. Blogs are also interactive, generally permitting viewers to post comments and engage in debates. Blogging has also encouraged the creation of articles in minority languages. Because of the small readership of minority language blogs, publication of this material in other formats would be too expensive. The free and easily accessible online format of blogs makes them effective platforms for local language content.

Potential drawbacks

Blogs fit the criteria for local content well, but measuring them is difficult. Our Google search often returns blogs which are clearly not of the language of interest. This could be due to the fact that snippets of multiple languages appear in the same blog. In this case, it is difficult to know how to classify the content. Additionally, the Google search only allows us to sample a few of the interesting languages which we would ideally like to measure.

For the data to be unbiased, we rely on the assumption that the Google search engine detects an equal proportion of blogs in all languages. For example, if Google is able to find 75 % of blogs in French, it must also detect 75 % of blogs in other languages of interest. If however, the Google algorithm is relatively more efficient in certain languages, statistical analysis of the dataset will be biased. Along similar lines, to accurately measure local content, blog use must be equally popular in the languages of interest. Previous research indicates that this might not be the case, at least in the earliest years of our dataset[16] (OECD, 2006).

Finally, blogs have two inherent biases as measures of local content. It is possible for one person to create several blogs simultaneously, thus distorting the usefulness of the metric to judge the number of different contributing voices. Additionally, not all blogs are created equal: while some blogs have over a million followers, other blogs are read only by their authors. On the production side, some blog authors post frequently (sometimes multiple times a day) while other authors write infrequently or not at all. The different degree of importance introduces the need for a subjective weighted index: we simply count the number of blogs (weight = 1), but other researchers might decide to use a more sophisticated approach which depends on blog readership. The subjectivity involved in any approach is an unavoidable drawback of blogs as a measure for local content.

Available data

The estimated number of blogs varies between sources. The discrepancy can be attributed to the fact that blog search engines use the 'number of links and the perceived relevance of blogs' to tabulate the numbers (OECD, 2006). Survey and sampling methods have also been employed to estimate the number of blogs, but these estimates vary as well.

In this paper, we use Google blog search to collect data on the number of blogs created in 46 languages from 2001-2010. Google blog search was chosen over other available search engines for *i)* consistency with the search engine used for ccTLDs and websites per language, and *ii)* Google search has a language option with 46 available languages. Figure A4 in the Annex contains a screenshot with our search parametres.

Number of tweets per language:

Twitter is a social networking site which was launched in 2006. Twitter permits users to share short text-based messages with a 140 character limit. Because of its availability to those with Internet connections, Twitter has increased in popularity worldwide. Today, it is estimated that there are over 106 million accounts on Twitter, and that this number grows by 300 000 every day.[17] Although the majority of tweets are in English, a market research firm has estimated that 11 % are in Portuguese, 6 % in Japanese, 4 % in Spanish, and 18 % in other language. Small local languages such as Haitian Creole, Maori, and Wolof are even present on Twitter[18]. By permitting exchange and even revival of small local languages, Twitter is a good metric of local content as defined by UNESCO.

Twitter provides an outlet from which people can express themselves in a wide variety of languages. Its growth internationally is testament to its value as a platform for the production of local content. Twitter permits people to share with a wide audience of speakers of the same language and discuss topics which are of interest to the community.

The Indigenous Tweets blog and companion website *indigenoustweets.com* are evidence that online-communities of speakers of local languages are present. This anecdotal evidence suggests that Twitter is becoming an important location for the creation of language-preserving local content.

Box 2. Indigenous tweets

Dr. Kevin Scannel, a professor of mathematics and computer science at Saint Louis University, maintains a blog called Indigenous Tweets. The purpose of the site is to help speakers of endangered languages find each other on Twitter. This enables young people, especially, to use social media available on the Internet to connect and form online language communities. Dr. Scannel hopes that his site will help to preserve and even revive many endangered languages.

Presently, Dr. Scannel's software and his blog track 82 minority languages. The largest languages followed by his software are Haitian Creole, Welsh, and Castilian. Smaller languages, such as Wolof, only have two Twitter users tracked by the software for the moment. With help from speakers of these small languages, Dr. Scannel hopes to expand the number of languages supported on his site.

Source : *http://indigenoustweets.blogspot.com/*

Potential drawbacks

Although Twitter is a diffusion medium for local content, in order for it to be useful in cross-economy comparisons of local content we must make several assumptions. Since Twitter is designed to be an online form of SMS or texting (*i.e.* there is a character limit for each message), a specialised Internet vocabulary including abbreviations is frequently used. For the number of 'Tweets' to be an accurate metric for local content, this specialised texting vocabulary has to be a part of the local language. If English abbreviations are common, this metric will not conform to our definition of local content. Also, controlling for demographic, religious, and educational characteristics, use of Twitter must be independent of economy. Specifically, this assumption bars the use of substitute websites such as QQ (China). Finally, the short length of Twitter makes it a popular rebroadcasting tool; links to other popular media sources are often sent as Twitter messages. By using Twitter as a measure for local content, we must accept that Twitter is not solely a creative platform for new content, but also a means for retransmission of extant content.

Available data

Data on the percentage of twitter messages in 61 languages is available from a French social media research firm named Semiocast. Although the data provides a reasonable approximation to the number of users, sampling error might be introduced as the company does not continuously track tweets; instead, it monitors for two-day periods. Other sources of data, such as Dr. Scannel's web-crawling software are also available. Due to restrictions by the Twitter site on the number of searches which can be performed by a user in a single day, Dr. Scannel's software is best suited to tracking small languages, ideally with only a few thousand speakers.

General trends

The dissemination of digital local content is changing rapidly as new disruptive technologies and ideas become mainstream. Measures suggested in this study provide a nice illustration: Country Code Top-Level Domains have been around since the mid-1980s,[19] Google since 1998,[20] Facebook since 2004,[21] and Twitter in 2006.[22] Thus the only platforms for online local content only 15 years ago have been completely eclipsed by sites less than 10 years old. It is not clear how long Google, Facebook and Twitter will remain some of the most popular sites on the internet (and hence the most popular locations for local content creation), but it does seem clear that the internet is dynamic. Cheaper and faster smartphones and tablet computers are bringing internet access to more people in more places. These new enabling technologies will almost surely be a hub that gives access to tomorrow's creators of local content.

In part due to the rapidly changing technologies and ideas that permit the creation of local content, it is difficult to find a single representative channel. Even the most popular websites such as Google and Facebook are young enough that their long-term viability is unclear. Additionally, these platforms are biased towards certain languages and age groups. Nonetheless, by combining many proxies, it can be established that local content creation is growing worldwide.

The OECD member countries were some of the earliest to witness a proliferation of local content. Currently, all of these countries have well-established sources of local content that continue to grow. Specifically, the top languages on Wikipedia are all official languages of OECD members, and sites such as Google and Facebook are among the most popular in the OECD. Growth of these platforms is still strong, although it has slowed from its early boom.

Worldwide growth in the development of local content (as measured by the above proxies) remains high, in part thanks to countries in the developing world. Omitting outliers such as Tuvalu and the Federated States of Micronesia, growth in ccTLDs, Facebook users, Wikipedia articles, etc. has not slowed

with the world's financial markets since 2008. Further, new languages are available on these platforms each year, bringing greater opportunity for self-expression and participation in the internet community to those who previously did not have access.

Although the internet is dynamic and it is impossible to precisely quantify the opportunities for local content creation that it creates, the measures of local content discussed in this paper suggest that the internet is a valuable outlet for local content creation. It is not only permitting content to be created by small groups who speak endangered languages, but it is permitting individuals of every nationality and speakers of every language to share. Local content, and internet content in general, are becoming decentralised and empowering the individual.

Conclusions

Local content can be defined in a number of ways, but in general it is intended for an audience which speaks the same language as the author. There are a number of measures for local content which meet these criteria, but each measure is an imperfect estimate as to the true amount of local content. Ideally, the researcher will combine several of the measures discussed in this paper in order to get an unbiased estimate of the quantity of local content across regions. In estimating the potential biases associated with the measures, it is important to consider *i)* the availability of the measure in an economy (*i.e.* regulatory environment and other barriers to entry); *ii)* the popularity of that measure relative to substituted products; and *iii)* how accurately the measure can be quantified. This paper has attempted to provide a brief discussion of these three points for each of our proposed measures. Future research involving econometric analysis with local content will also need to consider potential instruments (Instrumental Variables) for these measures as all are endogenous to local Internet infrastructure.

ENDNOTES

1 Miniwatts Marketing Group *www.internetworldstats.com/stats7.htm*

2 *www.iana.org/domains/root/cctld/*

3 *www.iana.org/domains/root/db#*

4 *http://en.wikipedia.org/wiki/CcTLD*

5 "Facebook opens Hong Kong office in Asia push", Sydney Morning Herald, 10 February 2011, at: *http://news.smh.com.au/breaking-news-technology/facebook-opens-hong-kong-office-in-asia-push-20110210-1an9c.html*

6 "Iran tightens online censorship to counter US 'shadow Internet", The Guardian, 143 July 2011 at: *www.guardian.co.uk/world/2011/jul/13/iran-tightens-online-censorship*

7 See: *www.gallup.com/poll/146159/facebook-google-users-skew-young-affluent-educated.aspx*

8 See: *www.facebook.com/press/info.php?statistics*

9 The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

10 See for example: *www.measuredhs.com/pubs/pdf/AR10/AR10.pdf*

11 See: *www.youtube.com/t/press_statistics*

12 See: *http://ksudigg.wetpaint.com/page/YouTube+Statistics*

13 See*: http://en.wikipedia.org/wiki/Reliability_of_Wikipedia*

14 See: *www.nature.com/nature/journal/v438/n7070/full/438900a.html*

15 See: *http://technorati.com/blogs/top100/*

16 This study indicated that blogs were disproportionately used by Japanese and Koreans, relative to the number of speakers of these languages. A 2010 Technorati report on the state of the blogosphere indicates that 49% of bloggers are located in the United States (relative to 25% of Internet users). Thus while blogs overestimate Japanese and Korean local content in 2006, they underestimate it in 2010.

17 See: *www.onlinemarketing-trends.com/2011/03/twitter-statistics-on-its-5th.html*

18 See*: http://indigenoustweets.blogspot.com/*

19 See: *www.ccwhois.org/ccwhois/cctld/ccTLDs-by-date.html*

20 See: *http://en.wikipedia.org/wiki/Google*

21 See: *http://en.wikipedia.org/wiki/Google*

22 See: *http://en.wikipedia.org/wiki/Twitter*

REFERENCES

OECD (2006), *OECD Information Technology Outlook 2006*, OECD, Paris, *www.oecd.org/sti/ito.*

OECD (2007), *Participative Web and User-Created Content; Web 2.0, Wikis, and Social Networking*, OECD, Paris.

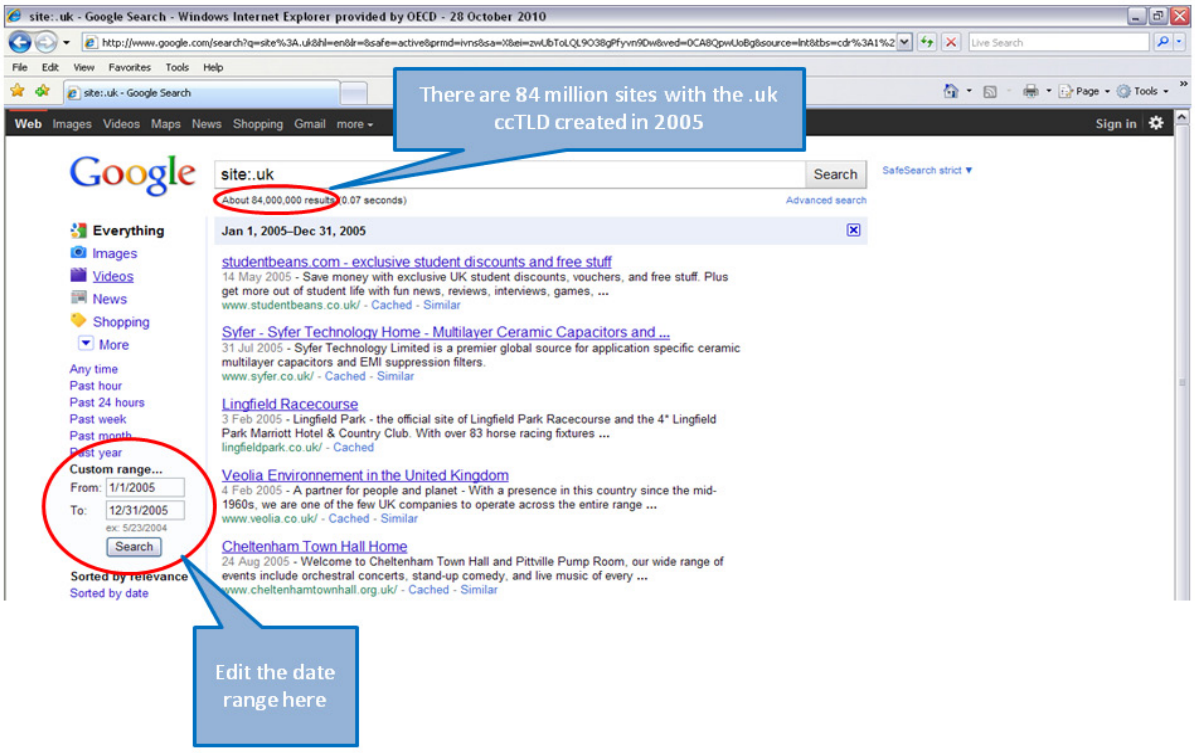
OECD (2011), "The Future of the Internet Economy: A Statistical Profile", OECD, Paris.

UNESCO (2001), "Public Service Applications of the Internet in Developing Countries, Promotion of Infrastructure and Use of the Internet in Developing Countries.", UNESCO, Paris.

ANNEX

The Google.com search features permit searches for web pages created in a certain time window and under a certain top level domain. To replicate our results, navigate to *www.google.com*, enter 'site:.xx' in the search field (where 'xx' denotes the ccTLD of interest), and click search. Now click on show search tools, and choose the date range of interest. Click on search again. Highlighted at the top of Figure A1 is the number of pages under the .uk ccTLD in 2005. It is worth noting that the Google algorithm is proprietary, subject to frequent changes, and not consistent in the number of returns for searches conducted at different points in time.

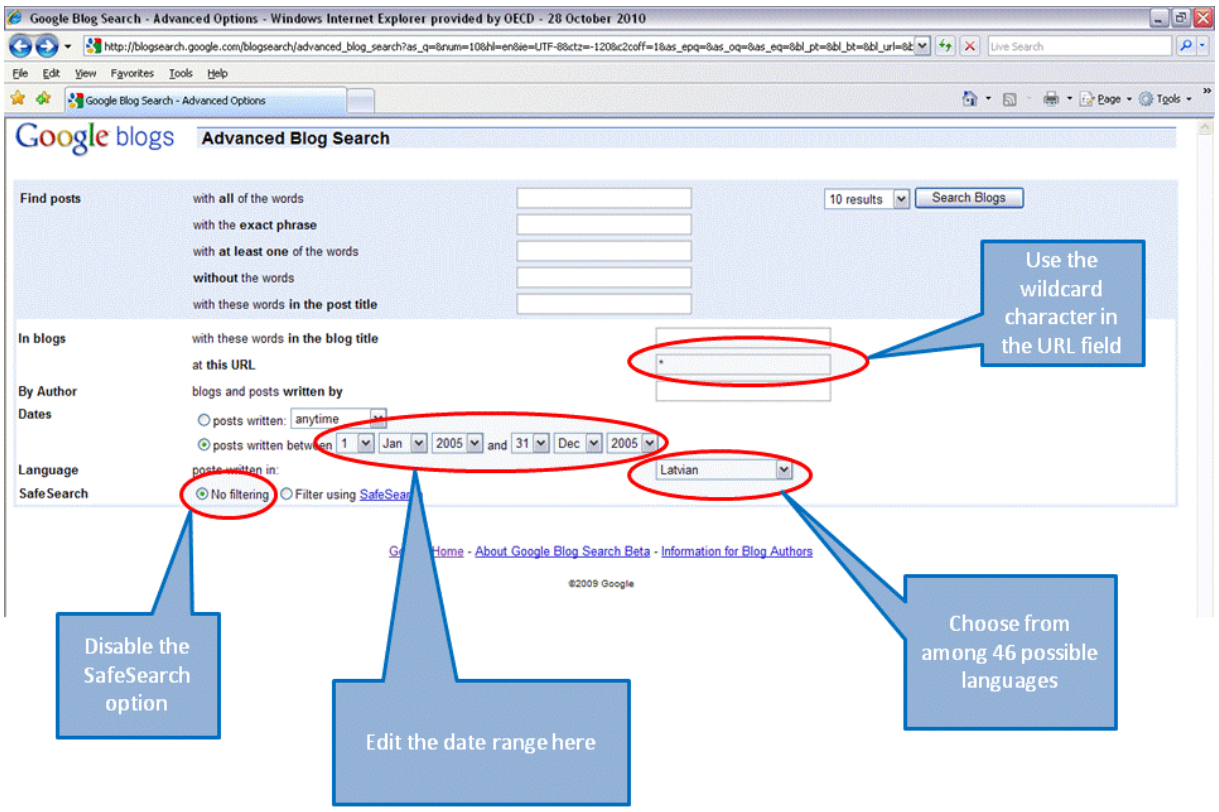
Figure A1. Parameters for Google ccTLD Search



Source: Google.com

The Google blog search is not performed at Google.com as Google has a special blog search engine: http://blogsearch.google.com/blogsearch/advanced_blog_search?hl=en. To replicate our results, use an asterisk (the wildcard character) in the URL field and turn the search filter off. Figure A2 highlights the remaining fields which need to be modified.

Figure A2. Parameters for Google blog search



Source: Google.com