GLOBAL INFORMATION SOCIETY WATCH 2008

Focus on access to infrastructure

Association for Progressive Communications (APC), Hivos and the Third World Institute (ITeM)

Global Information Society Watch

2008

Editor
Alan Finlay

Assistant editor
Lori Nordstrom

Publication production
Karen Higgs

Graphic design
MONOCROMO
Myriam Bustos, Leticia da Fonte, Pablo Uribe
info@monocromo.com.uy
Phone: +598 (2) 400 1685

Cover illustration
Matias Bervejillo

Proofreading
Lori Nordstrom
Lisa Cyr

Website
www.GISWatch.org
Andrea Antelo
Ximena Pucciarelli
Monocromo

Printed by
CinnamonTeal Print and Publishing

Printed in India

Global Information Society Watch 2008
Published by APC, Hivos and ITeM
2008

ISBN: 92-95049-65-9
APC-200812-CIPP-R-EN-P-0058

Accessing content

Daniel Pimienta
Networks and Development Foundation (FUNREDES)
www.funredes.org

The world wide web is, effectively, the largest public domain of information. But while it maintains its exponential growth in content, search engines are losing their capacity to index a significant part of it; and advertising directly and more perversely extends its reach, influencing users' behaviour when accessing content. At the same time, issues such as open content and the right to access this public domain of knowledge remain important, and some progress in this regard is being made. The demography of the net is finally evolving towards greater cultural and linguistic diversity, announcing the end of an initial and transitory phase of English dominance, which was a consequence of its historic development.

This chapter alerts us to the growing risk of bias from online search services: on the one hand, a bias that is culturally sensitive, and indirectly caused by a reduction in coverage or capacity; on the other, a bias that is the direct result of the influence of advertisers, who deliberately affect search results.

The extent of the challenge facing us is clear when we consider that a legitimate objective is for people to have access to online content in their own language, and sense that the digital divide is much deeper in terms of content than in terms of access to technology. And as new users get online, fewer and fewer of them appear to be content producers, emphasising the importance of digital literacy in the struggle against the digital divide.

Nurturing a wide public domain of information, especially in the sciences, is crucial for the future of our knowledge societies, and is important to the global development divide between the North and the South. It is commonplace that cyberspace must embrace and reflect the linguistic and cultural diversity of the world. The key, however, is how quickly this happens.

Content topology

What are the characteristics of what we commonly refer to as online "content"? Consider these figures: the number of internet hosts crossed the 500-million mark in 2008,[1] while the number of internet users is estimated around 1.4 billion,[2] the number of websites around 100 million or more, and the number of visible web pages[3] at least 140 billion.[4]

Table 1: **Worldwide internet statistics**

Internet users	1.4 billion
Registered domains	140 million
Websites	100 million-170 million*
Web pages	140 billion-one trillion
Indexed web pages	20 billion-40 billion

* Differences in figures may be due to virtual sites which are hosted on servers. See news.netcraft.com/archives/web_server_survey.html

An idea of the topology of the "content universe" is obtained by constructing the following ratios:

- Three users per internet host
- One domain name for every ten users
- One website for every fourteen users
- 1,000 web pages per user, 150 of which are indexed by search engines.

These ratios have probably kept relatively stable over the years, except the last one. In recent years the percentage of indexed pages has been shrinking to less than 15% of the total, potentially making users much more vulnerable to the various biases which condition their access to content, besides being more malleable to targeted advertising strategies[5] launched on search services.

Bias in access to content

Powerful applications like Google have for years been able to keep track of our web navigation behaviour, posing a threat to our online privacy. Empirical evidence suggests that the order of presentation of search results is not only decided by the ranking algorithm which has made Google so successful, but that it feeds off our personal history of searches in order to target us with sponsored links. Furthermore, keywords are being sold to commercial interests, questioning the whole idea of "objective information retrieval". Add to this the fact that 85% of the visible web now escapes the attention of web

1 www.isc.org/index.pl
2 www.internetworldstats.com

3 The invisible web (also called "deep web") is the sum of dynamic pages produced by databases or other programmed mechanisms that produce dynamic pages. Some authors estimate it could be 100 to 500 times larger than the visible web. See Bergman (2001).
4 Today it is impossible to find data for the total number of visible web pages. This figure has been extrapolated by the author from previous years' figures.
5 Advertising is so far the main driver of the content economy.

crawlers,[6] and the situation begins to feel like a subtle form of censorship.[7] Some voices are starting to complain[8] and citizens should follow carefully the evolution which makes a company like Google an ally for open access, as in their Google Scholar initiative,[9] but also a commercially biased operator that uses its basic search interface to make money.

Content diversity

While the average figures quoted above hold interesting meaning to understand the content universe, as always with averages, they hide the diversity factor.

The split of global internet users between regions[10] shows Africa with 4% and an internet penetration rate of only 5%, while Europe accounts for 27% of internet users and a penetration rate of 48%. The split of users per language shows English at 30%, followed by Chinese (17%), Spanish (9%), Japanese (7%), French (5%) and German (5%). As for the split of content on the web by language, there is no single source, and there are divergent figures for English.[11]

Indeed, the digital divide is not only a question of access: it is also, and even more, a question of content. FUNREDES studies[12] have shown that, for instance, more web pages are being produced in French by the United Kingdom (0.4% of the total) and Germany (0.5%) than the whole of Africa (0.3%), and that France is producing more English pages (0.7% of the total) than the whole of Africa (0.3%; 80% of them from South Africa). Furthermore, the trends observed show absolutely no improvement in the last five years. Language Observatory Project (LOP)[13] studies in Asia and Africa demonstrated than local languages accounted for a percentage of web pages in the order of 1%, 0.1% or 0.01% compared to cross-border languages (English, French, Russian or Arabic).

The depth of the digital divide as it is reflected through the lens of content appears much greater than the access gap: 4% of internet users are in Africa, while African languages account for less than 0.4% of all content, and less-spoken African languages for less than 0.04%!

And this is based on the 5% of the world's languages that have a digital existence – meaning that there is a codification scheme to transcribe their alphabets in digital form. Human beings have created some 40,000 different languages throughout history, of which some 7,000 are still used. Of these, only about 350 have a digital existence.

For the internet to be a resource for everyone, it will take much more than connecting everybody. It will mean allowing everyone to relate to the net in his/her mother tongue; which implies, obviously, the balanced existence of content in everyone's language.

The so-called pragmatics who believe this goal is unreachable – and therefore that it is acceptable to force people to work online in a non-native language and/or that English is the natural lingua franca of cyberspace – should consider the following: UNESCO studies[14] have shown that not being educated in a mother tongue is a significant handicap for children. And Wikipedia linguistic statistics[15] show the presence of articles in 264 languages, offering a reason to believe another world of content is possible…

The Internet Governance Forum (IGF)[16] in November 2007 started to take note of linguistic and cultural issues, as witnessed by a roundtable chaired by Gilberto Gil, the Brazilian minister of culture at the time, with the president of the World Network for Linguistic Diversity (MAAYA),[17] Adama Samassekou, invited as one of the speakers. Yet these efforts are focusing on the tip of the iceberg: while the internationalized domain name (IDN)[18] system will certainly mean progress when it allows users to navigate the web with links written in other character sets, this still falls short of confronting the challenges of cyberspace reflecting the genuine cultural and linguistic diversity of our planet.

At the time of writing this report, China has just passed the United States in terms of internet users (258 vs. 220 million)[19] – just one sign of the acceleration in the pace of change in internet demographics. What happens is a simple question of inflexion in curves getting closer to saturation when the penetration into a segment gets very high (the US has an internet penetration rate above 70%, and the figure for English speakers connected to the internet is above 50%).

6 At the time of writing, a new service, cuil.com, has been released. It claims to not retain the user search history, which is a good move, and also to index close to the whole web (122 billion pages). Unfortunately, the results so far are contradicting these claims.

7 FUNREDES studies have shown in particular that English content is over-represented in the search engines' indexes. See Observatory of Linguistic and Cultural Diversity on the Internet: funredes.org/lc

8 See in particular www.iicm.tugraz.at/iicm_papers/dangers_google.pdf

9 scholar.google.com

10 www.internetworldstats.com/stats.htm

11 Some sources claim that the percentage of web pages in English has been above 70% over the last ten years, in spite of the drastic change in user demographics, while others, such as FUNREDES, quote figures of less than 50% (funredes.org/lc). See UNESCO (2005).

12 See the above-mentioned studies at funredes.org/lc

13 www.language-observatory.org

14 portal.unesco.org/education/en/ev.php-URL_ID=21260&URL_DO=DO_TOPIC&URL_SECTION=201.html

15 en.wikipedia.org/wiki/Wikipedia:Multilingual_statistics

16 See www.intgovforum.org and in particular www.intgovforum.org/Rio_Meeting/IGF2-Diversity-13NOV07.txt

17 The World Network for Linguistic Diversity, like the IGF, was also a product of the World Summit on the Information Society (WSIS) process. See www.maaya.org

18 en.wikipedia.org/wiki/Internationalized_domain_name

19 See Barboza (2008). Note that the latest figures from Internet World Stats are different (May 2008: US=220, China=210). This serves to remind us that apart from the number of hosts the figures are not 100% reliable.

The above-mentioned FUNREDES studies have shown that initially there was a link between the growth of users and content in a given language. However, over time less content is produced proportionally to the number of users signing on: new users behave more like consumers than producers. The missing link is probably digital literacy, which includes sensitising users to the importance of content production.

Open content, the main global issue

The ultimate goal for universal access to telecommunication services is to allow citizens to communicate and access information and knowledge. The empowerment of all citizens through knowledge is indeed the essence of the information society, and the largest public domain of information shall be then considered a basic human right, and linked to social cohesion and economic development.

The Creative Commons[20] initiative offers a range of possibilities for legally protecting content in such a way that it becomes open content in the public domain, and it poses a significant challenge to traditional copyright protection. The point is to try to reverse a tendency of people overprotecting their content, and encouraging a more open approach, benefiting the general interest without causing harm to any particular interest.

Public domain information, also known as the "information commons", refers to freely accessing intellectual work, or the media on which this is stored, the use of which does not infringe on any intellectual property right, or breach any other communal right (such as indigenous rights) or any obligation of confidentiality.[21] The knowledge society must be built on the widest public domain to achieve its ambition.

Open access to content: An emblematic theme showing progress

Open access refers to scientific publications being placed in the public domain instead of being held by editors or publications. The current status quo is essentially the following: public money funds researchers, but the product of those researchers ends up being privately owned by publishers who legally take the intellectual property from the researcher, and indirectly from public administration and from the taxpayer. This is done in order to finance the editing and publishing system, which includes a peer-review system. The latter secures the prestige of a publication, on which researchers depend for academic recognition and credits. This suggests something of the challenge at stake in aiming to change a copyright regime – there are many interests involved!

It would be an extraordinary effort in worldwide collaboration, and a boost to research in a developing world which can hardly afford the high price of scientific publications, to see this wealth of scientific knowledge freely accessible at a click. Unfortunately, the complex resistance ingrained in a system created for the age of print prevents this from happening.

It is not that the scientific world has not tried to push the issue, as witnessed by the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities[22] in 2003 and initiatives such as the Public Library of Science (PLoS),[23] which are offering concrete solutions. The Scholarly Publishing and Academic Resources Coalition (SPARC)[24] is developing advocacy strategies in support of public policies on open access, and is reporting progress.

Again, the subject of linguistic and cultural diversity is not neutral to the struggle for open access, as the dominant system has played an important role in making English the language for scientific communication in most instances. ■

References

Barboza, D. (2008) China Surpasses U.S. in Number of Internet Users. *The New York Times*, 26 July. Available at: www.nytimes.com/2008/07/26/business/worldbusiness/26internet.html?_r=1&oref=slogin

Bergman, M. (2001) *The Deep Web: Surfacing Hidden Value*. Ann Arbor: Scholarly Publishing Office, University of Michigan Library. Available at: quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0007.104

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities: oa.mpg.de/openaccess-berlin/berlindeclaration.html

Creative Commons: creativecommons.org

FUNREDES Observatory of Linguistic and Cultural Diversity on the Internet: funredes.org/lc

Internet Governance Forum: www.intgovforum.org

Internet Systems Consortium: www.isc.org/index.pl

Internet World Stats: www.internetworldstats.com

Netcraft (2008) September 2008 Web Server Survey. Available at: news.netcraft.com/archives/web_server_survey.html

Public Library of Science (PLoS): www.plos.org

Scholarly Publishing and Academic Resources Coalition (SPARC): www.arl.org/sparc

UNESCO (2005) *Measuring Linguistic Diversity on the Internet*. Available at: portal.unesco.org/ci/en/ev.php-URL_ID=20804&URL_DO=DO_TOPIC&URL_SECTION=201.html

World Network for Linguistic Diversity: www.maaya.org

20 creativecommons.org
21 unesdoc.unesco.org/images/0012/001297/129725e.pdf
22 oa.mpg.de/openaccess-berlin/berlindeclaration.html
23 www.plos.org
24 www.arl.org/sparc

GLOBAL INFORMATION SOCIETY WATCH 2008

GLOBAL INFORMATION SOCIETY WATCH 2008 is the second in a series of yearly reports critically covering the state of the information society from the perspectives of civil society organisations across the world.

GLOBAL INFORMATION SOCIETY WATCH or **GISWatch** has three interrelated goals:

- **Surveying** the state of information and communication technology (ICT) policy at the local and global levels
- **Encouraging** critical debate
- **Strengthening** networking and advocacy for a just, inclusive information society.

Each year the report focuses on a particular theme. **GISWatch 2008** *focuses on access to infrastructure* and includes several thematic reports dealing with key access issues, an analysis of where global institutions stand on the access debate, a report looking at the state of indicators and access, six regional reports and 38 country reports.

GISWatch 2008 is a joint initiative of the Association for Progressive Communications (APC), the Humanist Institute for Cooperation with Developing Countries (Hivos) and the Third World Institute (ITeM).

GLOBAL INFORMATION SOCIETY WATCH
2008 Report
www.GISWatch.org

ISBN 92-95049-65-9

APC Hivos ITeM