Artif Life Robotics (2011) 16:311–314
DOI 10.1007/s10015-011-0937-5

ORIGINAL ARTICLE

**Katsuko T. Nakahira · Yoshiki Mikami · Hiroyuki Namba
Minehiro Takeshita · Shigeaki Kodama**

Country domain governance: an analysis by data-mining of country domains

Received: March 14, 2011 / Accepted: March 17, 2011

Abstract Along with protocols, domain names are critical internet resources. To build a safe, secure, and globally ubiquitous internet, all entities responsible for the governance of domains must allocate their lower-level domains with competence and discipline following the basic principle of decentralized governance. The governance and operation of a country code top-level domain is delegated to the governing and operating entity of the relevant country or territory. There are great differences among these entities in domain allocation policy, fees, discipline, and technical competence. The undertaking of governing a global infrastructure requiring high integrity, such as the internet, on the basic principle of decentralized governance is unprecedented. To enable this governance to work well, it is necessary to set up proper objectives for the governance of country domains, and to use highly transparent criteria for a multi-faceted evaluation of how far these objectives have been achieved. This article presents a set of country domain governance indicators which is intended to be used for the construction of a safe and secure ubiquitous global network.

Key words Internet governance · Country code top-level domains (ccTLDs) · Indicator · Openness of network

1 Introduction

Since the 1990s, along with the progress of the IT revolution that started with the use of the internet, it has been pointed out that a gap continues to exist between countries in the enjoyment of the benefits.[1] This gap is termed the "digital divide," but this divide is classified into several layers according to the object of research. For example, one article[2] identifies the differences in the "place" in which IT technology is used. Among these reports, the digital divide between countries is addressed. This is caused by economic disparities amongst other things, and often starts with the language barrier (education gap). The need to eliminate the digital divide is also discussed. However, these studies are all based on statistical values related to the physical means of communication, and do not give a view of the bigger picture of the overall disparity.

Mikami et al.[3] established the Language Observatory (LOP) with the intention of carrying out observations of the digital divide between languages on the internet. LOP combines Web crawling technology and a language identification engine[4] which can retrieve up to 10 multi-lingual Web pages. In addition, it automatically retrieves data included in the Web pages, such as URL information, tags, the main body, and the Web server response time. By combining the results of these observations and existing statistical data, it is possible to obtain a large amount of information, such as the number of Web pages per capita, the number of Web pages by language, the ratio of native language pages, the degree of mixing-up of character codes, and basic research concerning the infrastructure. Some of the results have been published by Nakahira and Mikami,[5] but their report does not extend to an interpretation of the results obtained, or to the development of an index to handle them systematically. Here, we consider country domain governance based on a distribution/growth chart of out-degree, which is one of the data-mined LOP observation results, and on an investigation of the number of links.

2 Country domain governance

While the 250 or so countries and territories differ considerably in the size of population, economic power, and the extent of the use of, and experience in, information technology, any country code top-level domain (ccTLD) governing

K.T. Nakahira (✉) · Y. Mikami · H. Namba · M. Takeshita · S. Kodama
Nagaoka University of Technology, 1603-1 Kamitomioka-machi, Nagaoka, Niigata 940-2188, Japan
e-mail: katsuko@vos.nagaokaut.ac.jp

This work was presented in part at the 16th International Symposium on Artificial Life and Robotics, Oita, Japan, January 27–29, 2011

Table 1. CDG (country domain governance) index

Category	Indicator	Description	
Access	Accessible and affordable	(A1) RPDM	Relative price of DNs compared to monthly income
		(A2) RPDG	Relative price of DNs compared to the global average price
		(A3) NDPP	Number of DNs (issued) per head of population
	Openness of network	(A4) RSLO	Ratio of number of servers located outside of the jurisdiction to the total number of servers
		(A5) NOLM	Number of outgoing links to global news media
Linguistic diversity	IDN service	(L1) NIDN	Number of active international domain names (IDNs)
	Local language use	(L2) LLPP	Number of local language pages per head of population
		(L3) RPLL	Ratio of pages in local languages to the total pages
		(L4) LDLI	Linguistic diversity measured using the Lieberson index
Security and trust	Security, stability and resilience (SSR)	(S1) SSMO	Share of spam mail originations
		(S2) RDRA	Ratio of domain names whose registrants are anonymous to the total number of domain names
	Trusted content	(S3) ADRP	Availability of a dispute resolution process

entity needs to satisfy a certain level of country domain governance (CDG) if a safe and secure ubiquitous global network is to be built.

The functions of a domain administrator can generally be divided into three areas: (a) technology (to ensure uniqueness in the name space), (b) economy (to allocate domain names, which are a limited resource, in a manner which is not wasteful), and (c) policy (to coordinate the rights of the various parties regarding domain names).[6] In considering the responsibilities of the administrator, we can refer to the framework given in the Affirmation of Commitments (AoC) made between ICANN and the US Department of Commerce.

On the basis of the functions and responsibilities identified above, we have built a CDG index system consisting of the 12 indicators shown in Table 1. These indicators were identified by analyzing cost information published by the domain administrators, and data collected by Web crawlers. They are highly transparent and can be observed on a continuous basis. We have confirmed that these indicators are effective for a multi-faceted evaluation of the state of governance.

In principle, these indicators can be derived objectively from open information sources on a continuous basis. This property is essential because the indicators should be shared by stakeholders on a continuous basis, and should be reflected in the actions of domain administrators. In fact, all the indicator data, other than the prices of domains (A1, A2) and the availability of a dispute resolution process (S3), can be derived by automatically collecting Web data and analyzing it with, for example, UbiCrawler.[7] The data for these indicators can be produced on a continuous basis as long as there are the required computing resources. Although domain prices and the availability of a dispute resolution process are openly available data, it is difficult to collect them mechanically because they are the type of information provided to users. Although, in our project, the team in GLOCOM, International University of Japan, collected this data manually, in a somewhat laborious process, the data obtained is objective enough to satisfy the requirements for the indicators mentioned above.

Table 2. Details of Web data collected so far

Year	Region	Max. number of pages/host	Number of pages	Data size (bytes)
2005	Africa	61 970	62 183 975	517 418 908 968
2005	Asia (excluding CJK)	34 466	51 416 890	490 253 207 703
2007	Africa	10 000	14 629 405	197 735 652 710
2007	Asia (excluding CJK)	10 000	22 827 036	144 587 836 525
2009	Africa	80 000	26 494 210	209 099 050 984
2009	Asia	10 000	46 962 508	464 656 628 810
2009	Caribbean	10 000	58 351 106	316 389 465 077

3 Example of the observed CDG index

We collected Web data from some 140 domains in the Asia–Pacific area and Africa. We had to exclude domains in Europe and North America because the sheer volume of data that exist in the domains in these regions exceeds the capacity of the computing resources of the research team. We studied by how much the sample size can be reduced without the risk of losing the significance of the survey results. We applied analysis of variance (ANOVA) to domains in the Asia–Pacific area and Africa, and found that the number of Web pages per site can be reduced considerably without any problem, but that a certain number of sample sites per region is essential. Therefore, we expect that by pursuing this study further, we should be able to develop indicator data from a minimal volume of data from big domains that exist in the major countries in Europe. Table 2 shows details of Web data collected so far by the research team. Table 3 shows the indicator data that have been collected so far for each region. Using these data, we show examples of our observations.

3.1 Openness of network: RSLO

This indicator was originally conceived to evaluate how extensively the internet infrastructure is built in the terri-

Table 3. Indicator data collected for each region (as of October 2010)

Region	Number of ccTLDs	(A1) RPDM	(A2) RPDG	(A3) NDPP	(A4) RSLO	(A5) NOLM	(L1) NIDN	(L2) LLPP	(L3) RPLL	(L4) LDLI	(S1) SSMO	(S2) RDRA	(S3) ADRP
Africa	59	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Asia–Pacific	82	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Europe	50	✓	✓	✓							✓		✓
Latin America	47	✓	✓	✓							✓		✓
North America	5	✓	✓	✓							✓		✓
Other	7	✓	✓	✓							✓		✓

Table 4. Location of Web servers and the languages used in Web pages of the domains of small island countries in the Pacific Ocean

ccTLD: country name	Percentage of servers installed overseas (%)	Main locations where servers are installed	Major language(s) used
fj: Fiji	52.70	USA 40%, Australia 6%	English 87%
fm: Federated States of Micronesia	98.00	Austria 43%, USA 21%	English 60%, Japanese 8%
ki: Kiribati	99.60	Germany 99%	English 55%, German 12%
nr: Republic of Nauru	100.00	Belgium 96%	English 96%
pw: Republic of Palau	100.00	USA 100%	English 87%
sb: Solomon Islands	41.90	Australia 23%, USA 14%	English 98%
to: Tonga	99.80	Japan 38%, Taiwan 29%	Japanese 54%, English 17%
tv: Tuvalu	99.90	Japan 46%, USA 23%	English 41%, Japanese 32%
vu: Vanuatu	99.10	Germany 96%, Canada 2%	German 71%, English 19%
ws: Samoa	99.70	USA 91%, Hong Kong 2%	English 73%, Russian 5%

tory concerned, but later we decided to adopt it to evaluate the openness of the network. Web servers are not necessarily located in the country represented by the domain name. Servers are located outside the country to escape from a variety of problems within the country, such as inadequate network infrastructure, non-availability of sufficient technical staff to manage servers, and non-technical restrictions imposed in the country.

A research representative once reported that more than 80% of Web servers under African country domains were not located on the African continent, but in Europe, the US, or Japan. This was mainly due to the desire to use the good network environment available in these regions. The preliminary survey of this research project also found that most of the Web servers under the domain names of small island countries are located overseas (Table 4).

3.2 Security, stability, and resilience: SSMO

In November 2009, some time before this research was started, a McAfee report said, under the shocking heading of "spam island-hopping," that the domains of small island countries were used as stepping stones for launching spam mail.[8] This report was one of the triggers that led to our decision to undertake this study.

The indicator we finally adopted was the percentage of domains originating spam mail. This figure was produced and analyzed in the way described below from 160 thousand mails that were found to be spam mails by SpamAssassin from among spam mails received in the mail accounts of the staff members of this university. The items of information listed below, which were included in the headers of sent mail, were used in this analysis.

1. Domain name of the sendmail server (mail server TLD, or "unknown" if unknown).
2. IP address of the sendmail server (IP address).
3. The TLD of the country where the server is located as identified from the IP address of the mail server using GeoIP[9] (which we refer to as the GeoIPTLD, or IPunknown if the given IP address is arrogated).
4. The TLD written in the sender address (mail address TLD).
5. Arrival date of mail (arrival date).

There are many levels of spam mail arrogation. While it is relatively easy for a malicious user to arrogate (i.e., use falsely) a mail server TLD or mail address TLD, it is difficult to associate such acts with problems of domain governance. In our analysis, it would be necessary to identify spam originators using their IP addresses. Therefore, we first traced daily changes in the GeoIPTLD derived from the IP address of the mail server concerned. The variance in the countries identified as spam mail originators was small between our staff members. Therefore, as far as mails received by the staff members of our university are concerned, the result can be considered to have a certain degree of generality.

The coincidence of the GeoIPTLD concerned with the mail server TLD concerned suggests that spam mails are originated openly. If they do not coincide with each other, one of them is false information; i.e., one of them is arrogated.

Table 5 shows the result of our analysis for cases where the GeoIPTLD coincides with the mail server TLD.

Table 5. Top 10 originating sources of spam mail (based on GeoIPTLD)

GeoIPTLD	Number of spam mails	Mail server TLD		
		Coincidence with GeoIPTLD (%)	Other ccTLD (%)	gTLD (%)
Greece (gr)	2182	85.7	2.0	12.4
Italy (it)	3660	84.8	1.8	13.5
Poland (pl)	3228	72.9	21.3	5.9
Brazil (br)	3211	62.8	32.3	4.9
Germany (de)	5619	47.7	1.2	51.0
Spain (es)	2357	35.4	3.6	61.0
France (fr)	5948	28.1	0.9	71.1
Russia (ru)	3139	27.4	10.6	62.0
Ukraine (ua)	3249	3.5	13.7	82.8
UK (gb)	7935	2.0	0.7	97.3
USA (us)	19185	0.03[a]	0.5	99.5

[a] For this "us," GeoIPTLD = mail server TLD = ccTLD

Although the degree of coincidence varied from day to day, we found that one-third to a half of spam mails used an arrogated IP (i.e., the country where the mail server concerned is located cannot be determined by using the IP address of the mail server). Table 5 shows the coincidence of the mail server TLD with the GeoIPTLD in data in which the above cases are excluded. The upper five country domains in the table are those with a high percentage of coincidence. This may suggest that the effort to control the originating sources of spam mail is not very strong. In the country domains in the lower half of the table, there was a high percentage of cases where the mail server TLD was gTLD ("com" in most cases).

4 Conclusions

We have proposed the CDG index as an observation index by which to home in on the reality of country domain governance. This allows the indexing of the reality of the digital divide in a form that is closer to the conditions of use to be undertaken, along with carrying out data mining by region in a form in which the various pieces of Web information obtained from the Language Observatory observation results are merged with various statistical data. Several examples of these observations have been provided here. In the future, we plan to continue our observation of Web space based on this index.

References

1. Ikuo O (2009) Toward the construction of a global information society. International trends and issues surrounding the digital divide, and creation of intellectual property, vol 3, pp 6–36
2. NTT C&C Foundation (2002) The digital divide (in Japanese). NTT Foundation
3. Mikami Y, et al (2005) The language observatory project. Proceedings of the 14th International World Wide Web Conference 2005, pp 990–991
4. Nandasara ST, et al (2008) An analysis of Asian language Web pages. Int J Adv ICT Emerging Regions 1(1):12–23
5. Nakahira KT, Mikami Y (2008) Measuring language diversity in cyberspce. International Conference on Linguistic and Cultural Diversity in Cyberspace, Yakutsk, Russia, July 2–4
6. Mueller M (2002) Ruling the root. MIT Press, Cambridge
7. Boldi P, et al (2004) UbiCrawler: a scalable full distributed web crawler. Software Practice Experience 34(8):711–726
8. McAfee Report (2006) November 7
9. GeoIP, http://www.maxmind.com/app/country, accessed: 2011.3.13