How old is the Web? Characterizing the age and the currency of the European scientific Web[1]

José Luis Ortega*, Viv Cothey** and Isidro F. Aguillo*

*Cybermetrics Lab, CINDOC-CSIC, Joaquín Costa, 22. 28002 Madrid. Spain {jortega; isidro}@cindoc.csic.es
** School of Computing and Information Technology, University of Wolverhampton, Lichfield Street, Wolverhampton, United Kingdom, WV1 1SB viv.cothey@wlv.ac.uk

Abstract

The aim of this paper is to model and study the age of the Web using a sample of about four million of web pages from the 16 European Research Area countries obtained during 2004 and 2005. Web page time-stamp (date when the web pages were created or last changed for last time), format and size in bytes data have been analysed. Several indicators are introduced to measure longitudinal aspects of the Web. Half-age is proposed as a measure of the age distribution because this is found to be exponential. "Web Update Index" and "Lifespan Index" are introduced to measure the changing rate of a small sample over time. Results show that the British Web space has the youngest Web pages while the Greek and Belgian ones have the oldest. The study also compared Web pages topics and found that Biology pages are more stable than Physics pages.

Introduction

The World Wide Web is a complex system in continuous evolution. Web content is added, deleted and changed daily without any apparent pattern [ORTEGA, AGUILLO & PRIETO, 2006]. A single snapshot illustrates the scale-free topology of the Web in which a few nodes attract many links while most nodes attract only a few links [ORTEGA, AGUILLO, COTHEY & SCHARNHORST, 2008]. ROUSSEAU [1997] and BARABASI AND ALBERT [1999] were the first to identify the Web as a scale-free network in which preferential attachment or cumulative advantage [PRICE, 1976] and the power law distribution describes its inlink structure. However, this model does not explain the emergence of new nodes [ADAMIC & HUBERMAN, 2000] or their disappearance.

Several studies have shown that Web growth can be described using a power law. PENNOCK, FLAKE, LAWRENCE, GLOVER & GILES [2002] discovered that incoming links to a website increase in accordance with a power law. Other investigations show that web domains or sites have grown since 1994 with a similar power law rate [INTERNET SYSTEMS CONSORTIUM, 2004; NETCRAFT, 2007]. The OCLC Web Characterization Project [O'NEILL, LAVOIE, BENNET, 2003], carried out between 1998 and 2000, found that although the Web keeps growing, the rate of content contribution slowed by 1% during the 2001-2002 period. Later NIELSEN [2007] showed that the Web's growth rate has reduced from an explosive period in 1991-1997 (850%) to a more sedate 25% during the period 2002-2006.

The continuous disappearance and modification of Web content are key factors in understanding the evolution of this complex network. Several studies have

[1] This paper has been accepted in order to be published in the journal *Scientometrics*

investigated the ephemeral existence of incoming links in e-journals [HARTER & KIM, 1996] and scientific repositories [LAWRENCE, PENNOCK, FLAKE & AL., 2001; SPINELLIS, 2003]. Other studies using longitudinal surveys show the gradual disappearance of Web pages [KOEHLER, 1999; 2002; 2004] and digital objects [NELSON & ALLEN, 2002]. An estimate of the rate Web page disappearance is 0.25% to 0.50% per week [FETTERLY, MANASSE, NAJORK & WIENER, 2003]. Furthermore, BAR-YOSSEF, BRODER, KUMAR and TOMKINS [2004] discovered that a quarter of 200 (OK) pages are really *soft 404* or dead links which introduces more complexity to the study of the Web persistence and decay. A soft 404 is when the server redirects to a page with a user friendly message such as “the resource that you requested is no longer available” rather just presenting an automatic 404 page not found error.

A few papers have attempted to describe, calculate or characterize the age of the Web. The first was by DOUGLIS, FELDMANN and KRISHNAMURTHY [1997]. They reported that the age of a Web page varies according to both its content type and top-level domain, but not its size. BREWINGTON and CYBENKO [2000] studied 100,000 Web pages from the main search engines daily. They found a power law distribution of their age in days. About one page out of five is younger than eleven days, and they reported that about half of Web content is less than three months old. BAEZA-YATES and CASTILLO [2005, 2007] characterized the Chilean Web space during both 2004 and 2006. They discovered that in the 2004 study, 25% of Web pages were created during 2004 while for the 2006 study the proportion created during 2006 falls to 22%. The rate of new Web page creation is thus reducing. BORDIGNON, LAVALLÉN and TOLOSA [2006] investigated the Paraguayan Web and showed that 60% of Web pages were created or modified within the previous year. BORDIGNON and TOLOSA [2006] also analysed the South American academic Web space and discovered that between 37% and 70% of Web pages were created during the current year. More recently, TOLOSA, BORDIGNON, BAEZA-YATES and CASTILLO's [2007] study of the Argentinean Web found that 72% of Web pages were created within the previous year. This makes Argentina the youngest Web space in South America.

Objectives

The aim of this investigation is to study the age of the Web and the dynamics of the change in Web page content. We do this, using two samples of pages from the European Research Area (ERA) Web space. We compare the youthfulness or maturity of this European Web space with other studies. Finally we propose using models that indicate the age and currency of a Web space to assess the evolution and growth of the European academic Web space.

Methodology

Data

The first sample consists of approximately four million Web pages collected from the ERA Web space using the *Blinker* Web crawler [COTHEY, 2004; 2005]. European Research Area (ERA) is a unified area created in 2000 where researchers and institutions from the European Union and Switzerland can share knowledge and resources. So, ERA Web space comprises the academic and scientific web content of the European Union state member in 2005 (EU-15), included Switzerland. The date of collection of each page is known and was within the period March to October 2005. Each Web page was analyzed in order to determine its file time-stamp (date when the

Web page was created or last changed), format and file-size in bytes. We calculated the age of each Web page as the difference between the crawler's time-stamp of when the Web-page was collected and the file time-stamp of when the Web page was last modified or created. This suggested that some Web pages were older than the Web itself or younger than how long ago the crawling was carried out. This is because not all Web servers correctly report file timestamps. Potentially this is a serious problem and pages older than the Web's creation in 1991[2] [CAILLIAU, 1991; BERNERS-LEE, CAILLIAU, GROFF & POLLERMANN, 1992] or dated after the crawler collection activity clearly have invalid time stamps. However the proportion of pages affected is small 9,604 (0.24%) and these were excluded from the sample. The remaining 3,942,655 Web pages with valid time-stamps formed the study's major sample.

For the second or minor sample we monitored 146 selected Web pages daily. This was in order to analyse when they were updated or modified. During the period September to November 2005, 43 pages were checked daily for 55 days while 103 pages were checked daily for 77 days. Each day each of the Web pages was collected by the Blinker crawler and analysed by calculating the page's MD5 signature [RIVEST, 1992]. The MD5 signature is designed to be a digital fingerprint for a computer file. If there is any change to a file (that is, the Web page) then the MD5 signature changes. On the other hand if the MD5 signature is the same then the file has not changed. The selected Web pages represent two research disciplines or topics, Biology and Physics. Hence as well as allowing us to study the rate of daily change overall, we can compare the two topics for any differences.

Data can be downloaded from: http://www.scit.wlv.ac.uk/~in7803/sandpit/jose/.

Half-age

Previous studies have indicated the age of a Web space simply using the proportion of Web pages younger than one year. However this measure does not characterize the whole distribution. We introduce the half-age as indicator for a Web space which, since the distribution of ages is exponential, allows us to estimate the age distribution over time. The notion of *half-age* ($t_{1/2}$), is derived from the nuclear physics term *half-life* [RUTHERFORD, 1900]. KOEHLER [1999] used the half-life concept to allude to the average persistence of a web page during a time period. However, half-life as the median of an exponential distribution was used to study the persistence of web references [SPINELLIS, 2003] and the change ration of web pages [CHO and GARCIA-MOLINA, 2000]. We introduce Half-age as the age at which we can find the half the number of Web pages in a sample. Thus, in a sample distribution of Web page ages, the half-age is the median age. Hence, the more recently Web pages are created the smaller is the half-age. This gives us an indicator for the youthfulness of a Web space.

Mathematically the half age in years is expressed as:

$$t_{1/2} = \frac{\ln 2}{\lambda}$$

where the half-age $t_{1/2}$ is the natural logarithm of 2 divided by the decay constant λ and is found from the exponential regression:

$$W_a = W_0 e^{-\lambda a}$$

where W_a is the number of Web pages which have the age a and W_0 is the number of Web pages age 0, that is less than one year.

[2] The fist web server (SLAC) outside from CERN opened in 1991 from which time the Web spread across the World.

The smaller the half-age of a collection of Web pages from a particular Web space, the younger is the Web space. Thus, by computing half-ages for each country we can compare the youthfulness of country Web spaces within the ERA

Web Update Index

The WUI expresses the daily rate of change of the URLs. CHO and GARCIA-MOLINA [2000] used the *average change interval* as number of changes during a time window period. It is measure in time term, assigning the average period in which the page remain without changes. This measure is dependent on time window used. KOEHLER [1999] proposed omega value (ω) as the average proportion of changed web pages in the analysis period over the total sample. His indicator is similar to our one; however it is dependent on the sample and it does not show individualised values for each web page. We introduce the WUI as no dependent measure because it may be compare with web pages from different time windows. To calculate WUI, a Web page is assigned a score of 1 if it had changed or 0 if it had remained the same. Hence for example, if a Web page is surveyed daily for 6 days then discounting the initial baseline survey and starting at day 2, the scores assigned are a sequence of five zeros or ones. These can be summed to get a total score varying between 0 and 5. Now normalizing this total by dividing by 5 we arrive at an index varying between WUI = 0 and WUI = 1 where 0 represents stability or no daily modification and 1 represents the maximum rate of change with modifications occurring every day.

Mathematically this is expressed as:

$$WUI_i = \frac{\sum C_i}{D_i}$$

where *C* is the value of change, and *D* is the number of days monitored. So, a Web page which changes each day has a WUI=1 and a Web page without any daily modifications has WUI=0.

Web page lifespan indicator

WLI helps to study the change interval or lifespan of a Web page. CHO and GARCIA-MOLINA [2000] also used the *visible life-span*, which measures the number of changes during a time window period. As the previous indicator, visible life-span is dependent on time window. WLI tries to solve this drawback. To calculate it, a Web page url is assigned a score each day starting at 1 and increasing by 1 each day if there is no modification to the Web page. If the Web page has changed then the score assigned resets to 1. If the Web page is not accessible then a score of 0 is given. Hence for example over a six day period if the Web page at a particular url remains the same then its scores will be the sequence 1, 2, 3, 4, 5, and 6. The total here is 21. If the Web page is modified once then its WLI sequence might be 1, 2, 3, 1, 2 and 3 or 1, 1, 2, 3, 4 and 5 giving totals of 12 and 16 respectively. These are normalized by dividing (in this case) by 21 so that the WLI is 1 for the longest lifespan or period of no modification, and 12/21 or 16/21 for the other examples. The larger the index the longer is a Webpage's lifespan of no modification.

This formula shows how this indicator is calculated:

$$WLI_i = \frac{n_i(n_i + 1)}{2\sum N_i}$$

n_i is the number of observations in a time period where 0 is no server response and N_i is all observations over the period.

Results

We present two sets of results. The first relates to the static analysis of the distribution of Web page ages. The second is the dynamic analysis of the rate of change or updating of Web pages.

Age distributions

Figure 1 shows the distribution of the number of Web pages overall according to their age when sampled or collected. The age of a Web page is the time interval between when it was last modified and when it was collected. The distribution is exponential (y = 3314.8x^-0.527) over all the age range (15 years) back to the oldest Web pages created in 1991 (R^2=0.92). 53.2% of Web pages were created or modified within a year of their collection and a further 16.6% were between one and two years old when collected (Figure 1). When either just the younger half or older half of the sample is considered then the exponents are -0.545 and -0.527 showing close agreement and a uniformity across the whole sample.

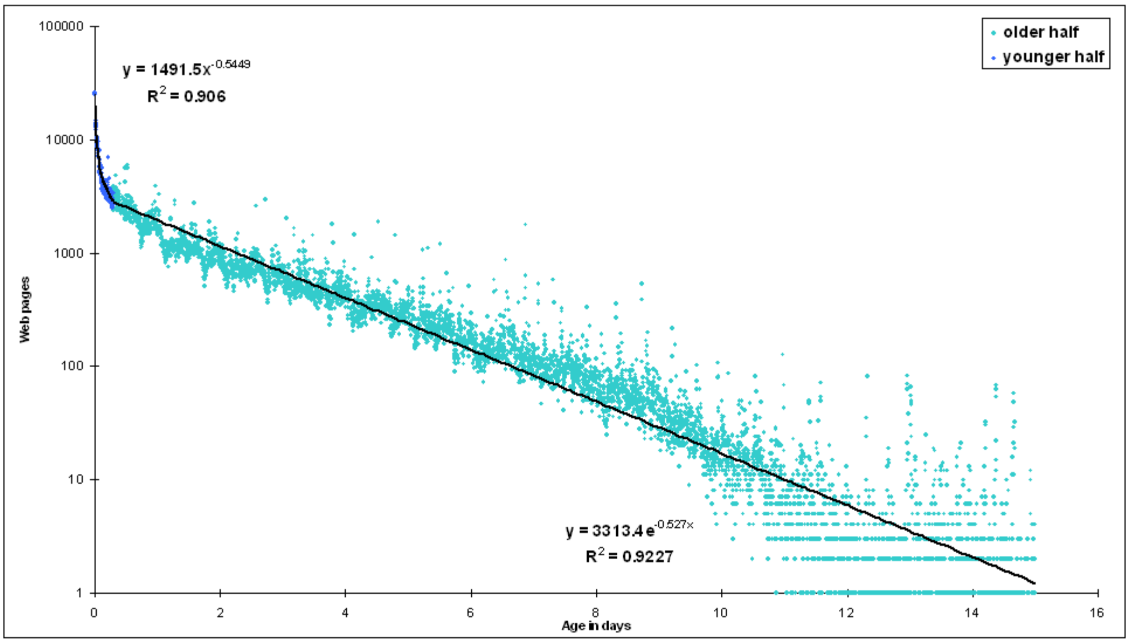

Figure 1. Years distribution of web pages (log-normal scale)

This is shown in Table 1. The United Kingdom (UK) and Switzerland are the most youthful since half of their Web pages are no more than 1.32 years (=481 days) old and 1.42 years (=518 days) old respectively. We note that the result for Luxembourg is not reliable because of the small number of Web pages in the by-country sample and the poor fit to the exponential age model (R^2=.01).

Countries	Pages in 0 year	Pages in 0 year %	Total pages	λ	Half-age (in years)	Half-age (in days)	R^2
Luxembourg	234	34.36	681	-.109	6.36	2321	.01
Greece	29,051	47.7	60,908	.313	2.21	808	.37
Belgium	90,393	75.42	119,849	.367	1.89	689	.54
Austria	62,603	68.54	91,342	.369	1.88	686	.62
Italy	162,464	59.82	271,572	.374	1.85	676	.78
Ireland	36,215	67.66	53,525	.389	1.78	650	.62
France	155,925	63.67	244,884	.395	1.75	641	.75
Finland	89,069	52.91	168,351	.403	1.72	628	.79
Sweden	151,593	71	213,500	.405	1.71	625	.82
Denmark	99,817	78.17	127,693	.417	1.66	607	.67
Portugal	48,264	61.24	78,799	.422	1.64	600	.54
Netherlands	84,338	70.03	120,425	.429	1.62	590	.61
Spain	141,756	59.92	236,575	.458	1.51	552	.77
Germany	789,505	72	1,096,474	.478	1.45	529	.91
Switzerland	159,124	79.84	199,314	.488	1.42	518	.79
UK	605,867	70.61	858,087	.526	1.32	481	.89
ERA	2,706,218	68.64	3,942,655	.531	1.31	476	.92

Table 1. Age distribution by countries ranked by half-age

The countries in Table 1 are shown ranked by their decay constant λ. For example, the half-age of Austria is:

$$Austria_{1/2} = \frac{\ln 2}{.369} = -1.878 \text{ years}$$

That is half of the Austrian pages are younger than 1.88 years (or 686 days) and half are older.

At the other extreme of maturity we found that half of the Web pages from Greece and Belgium are 2.21 years (=808 days) and 1.89 years (=689 days) old respectively. This makes them the most mature country Web spaces in the ERA. Overall the ERA Web space has a half-age of 1.31 years (=476 days).

The distribution of Web pages by age fits an exponential distribution similar to that illustrated in Figure 1 for every country except Luxembourg. R^2 is between 0.37 (Greece) and 0.91 (Germany)

We therefore conclude that the age distribution of Web pages is exponential unlike other many Web phenomena (Web pages per site, in-links per site, etc.) which generate power law distributions [ADAMIC & HUBERMAN, 2001; ALBERT, JEONG & BARABASI, 1999].

We found that a very high proportion (36.3% overall) of Web pages are created the same day as they are collected by the crawler. Like BORDIGNON and TOLOSA [2006] we believe that these pages are mostly dynamically created when they are requested by the user. The countries with highest proportion of dynamic pages are Denmark (59.2%) and Switzerland (55.2%), while Finland (16.8%) and Greece (17.6%) are the countries with the smallest proportion of dynamic pages.

In addition to analysing the distribution of Web pages by country (as indicated by the domain country code e.g. .es is Spain) we examined the age distribution of the ERA Web pages according to the main types of Web file extensions found. Three categories are considered. Web page hypertext formats such as .htm, .html, .shml, .php or .asp are grouped together as “hypertext”; “applications” means formats such as .pdf, .ps, .rtf and .doc; and “images” groups together formats such as .jpg, .gif, .bmp or .tif.

These three format categories account for 97.93% (3,870,403) of all file extensions found in the sample. The Web page half-age was computed for each category of format type. The findings are shown in Table 2.

The "hypertext" format category has a half-age of 458 days (1.25 years) which is similar to that for "applications", while for the "images" category the half-age is 778 days (2.13 years). We conclude that "image" format Web pages are least likely to be created or modified compared with "hypertext" or "applications" format Web pages. That is, the textual content of a Web site is modified more frequently than is the graphical content. However this conclusion is tentative because the low coefficient of determination (R^2=.7) makes the estimation of half-age images less precise.

	Web pages	Web pages %	λ	Half-age	R^2
hypertext	3,376,154	85.42	.553	458	.91
application	385,781	9.76	.542	467	.92
images	108,468	2.74	.325	778	.7
TOTAL	3,870,403	100			

Table 2. Age distribution by formats ranked by half-age in days

We also investigated the relationship between the web page age and size in bytes. Only the hypertext format category is analyzed in detail. The frequency distribution of hypertext pages by size does not follow a specific distribution (Kolmogorov-Smirnov Z=889, p-value=.000), although the best fit is obtained with a power law (R^2=.73). Figure 2 shows the initial hook-shape of the distribution in which it starts with a normal trend and finish with a power law one. This shape is usual in distribution dominated by a mean value because is exceptional to find both very little pages and very large pages. We hypothesized a difference between the distributions of the younger and the older hypertext pages in order to examine whether or not Web pages were becoming larger or smaller. We split the sample into two equal parts: the first part is the younger one and the second part is the older one. This is called a split-half procedure. We then carried out the Two-Sample Kolmogorov-Smirnov test comparing the means of the two parts. This statistical test is a non-parametric test suitable to compare two data samples come from the same distribution. There is no difference (Kolmogorov-Smirnov Z=204.21, p-value=.000) so we can conclude that the size distribution of Web pages does not change in young and old web pages.

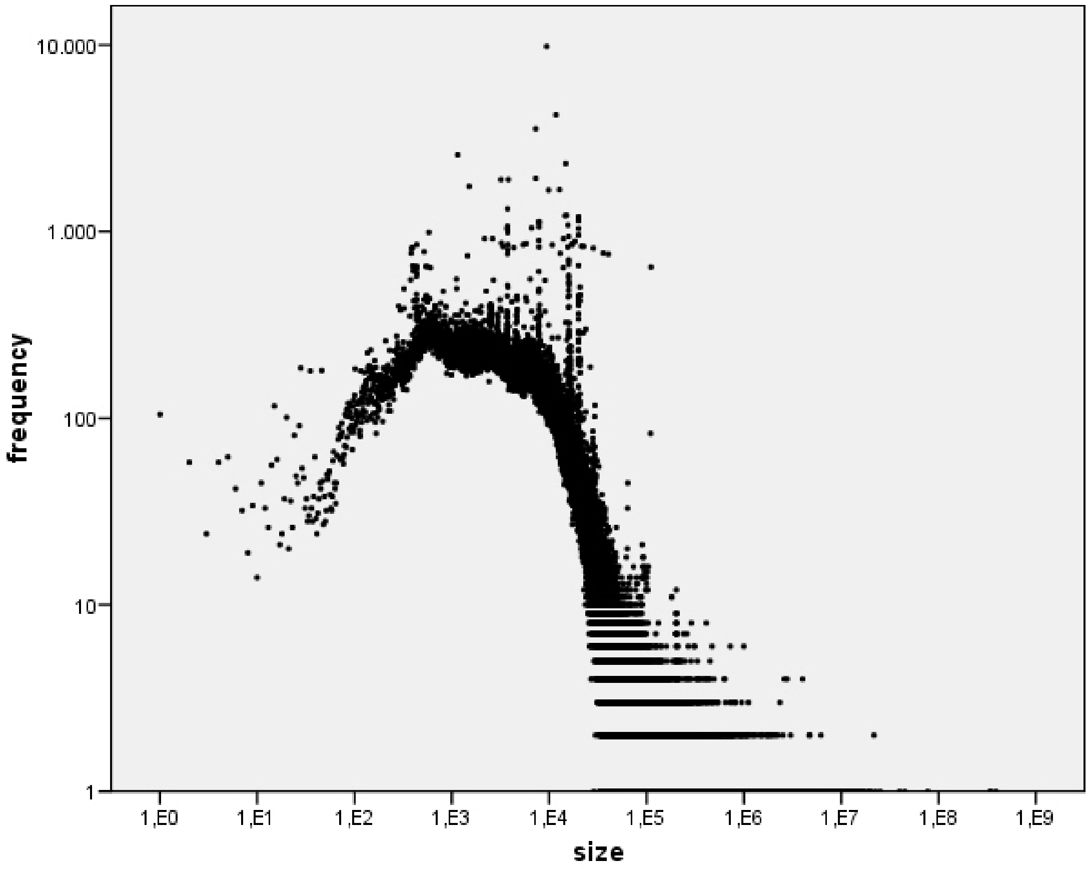

Figure 2. Size frequency distribution (log-log scale)

Update rate

The minor sample was a longitudinal study of daily samples of the same 146 Web page urls over a period of 55 (43 pages) and 77 (103 pages) days.

The objective was to monitor the daily change in Web pages. Each day the Web page from each url was compared with the Web page seen the previous day. Over the survey period 68 (46%) remained the same, 63 (43%) changed at least once and 16 (11%) became inaccessible during the survey period and dropped out of the study.

Two indicators were developed. The first of these is a Web update index (WUI) which measures how dynamic or subject to change is the Web page. The second is a Web page lifespan indicator (WLI) which measures the durations between changes.

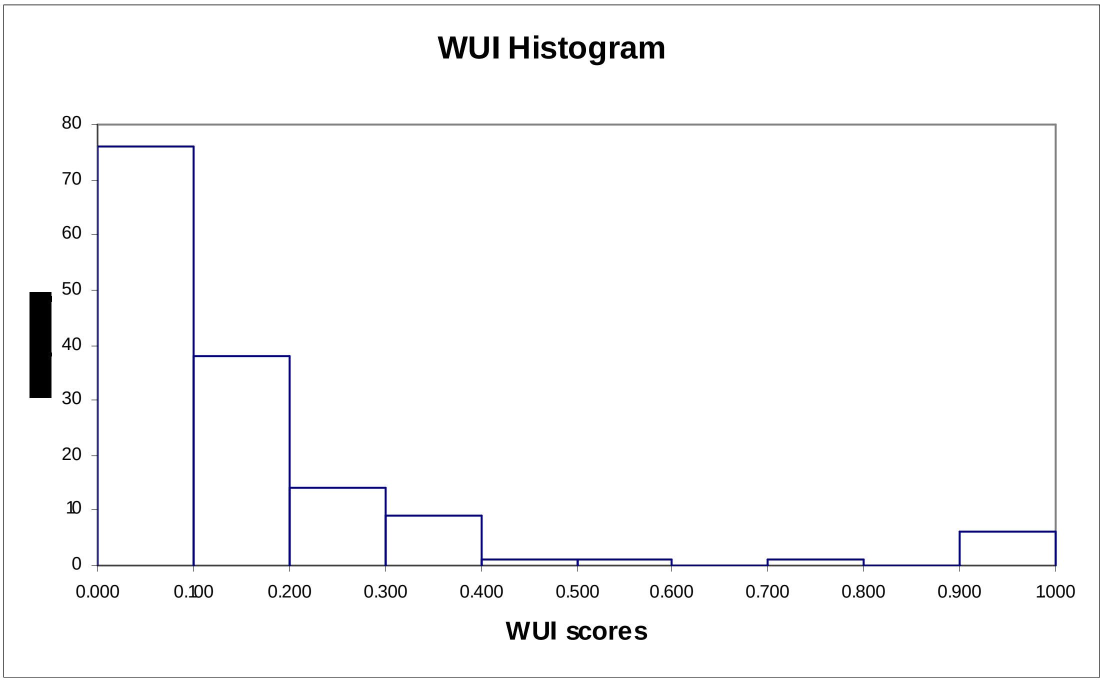

Figure 3. WUI Histogram.

Figure 3 shows the Histogram distribution of the WUI indicator. We find that most pages are very stable with a small WUI however there are a few with a WUI approaching 1 that change nearly every day. Inspection shows that these pages contained calendar information.

	WUI Average	*Life-span Average*	*URLs*	*%*
Biology	.178	.72	101	69.17
Physics	.193	.68	45	30.82
TOTAL			146	100

Table 3. Comparative table between topics

Table 3 shows the different indicator values for two different disciplines. The Student's t-test for dependent samples was made to check if there is any difference between Physics and Biology web pages. The test showed that there are differences between them (two-tailed p-value>.01). We can see that Physics pages (.193) have a slight higher WUI than Biology (.178). This means that Physics pages are a bit more updated than the Biology ones. The Life-span indicator also shows that Biology pages are more stable (.72) than the Physics ones (.68).

Discussion

Two main data extraction methods have been used in Cybermetric research: crawlers and search engines. Search engines have a great coverage and are better tools to perform macro level analyses, while the crawlers are a more precise tool and may be customised. However, some studies have shown that the search engines are not suitable for longitudinal analyses. LEYDESDORFF and CURRAN [2000] show distrust with the data reported by AltaVista when they retrieve the date of the web pages in several queries. They argue that the constant update of web pages distorted the view of the past. BAR-ILAN and PERITZ [1999] found that the rank dynamics considerably affects to the longitudinal retrieval of data. BAR-ILAN [2001] also observed that AltaVista assigns the date it visited the page for the last time. This suggests that crawler's data are more suitable because it avoids these miscounting and let us to reliably compare several populations along the time [ORTEGA et al., 2008; PAYNE and THELWALL, 2008].

This study has enabled us to characterize the age distribution of a sample of Web pages from the ERA Web space. The half-age indicator allows us to characterize the youthfulness of a Web space more effectively than by using just the percentage of web pages created or modified in the first observation year. For example, Switzerland is the country with the highest proportion of pages in younger than a year (79.84%) but its half-age is 518 days. The UK has a comparable proportion of 70.61%, but a half-age of 481 days old.

We are able to compare our results with previous studies. BORDIGNON and TOLOSA [2006] found that the proportion of pages created within the previous year in the South American academic web space varies from the 37% of Bolivia to the 70% of Chile. In comparison the ERA web space is more dynamic and younger since the proportion of pages created within the previous year ranges from 47.7% for Greece and to 79.8% of Switzerland (excluding Luxembourg at 34.4%). TOLOSA, BORDIGNON, BAEZA-YATES and CASTILLO [2007] revealed the youthfulness of the Argentinean web finding that 72% of pages were created during the previous year and a half-age for the age distribution of 301 days. This is younger than the ERA web space. However the results presented by BAEZA-YATES and CASTILLO [2005, 2007] regarding the Chilean web show that only 20-25% of pages are created during the previous year which is quite different to either the ERA or Argentine Web.

We also found that the web pages with an age of 0 days were an outliner in the distribution. This suggests that this type of pages is qualitatively different and is, for example, a dynamic page. This phenomenon was also found by [BORDIGNON & TOLOSA, 2006]. Our examination of the distribution of web pages therefore prompts us to suppose that there are three web spaces with different behaviour. In the first one might group dynamic pages (ERA=37.3%) which change their content daily. The second one could be the younger half (ERA=31.38%) which might represent the constantly updated web pages, which probably corresponds to dynamic content, whose levels of updates descend more accentuated. This type of contents could be considered as live or current contents. Finally, the older half (ERA=31.36%) might be web pages less updated, whose contents do not change for a wide time period and their update rates descend less marked. These pages may correspond to old and outdated contents.

These two last classes, the younger half and the older half, have been detected in the age distribution (Figure 1). In that figure we can observe two different trends. The young half follows a power law distribution, while the older one follows an exponential fit. A power law distribution with negative exponent has stronger decay process than the exponential one. This may be because the younger half is characterized by a great number of dynamic pages and more active pages which update rate is high, while the older half is set up by pages with a low update rate and non update pages. This trend has been also detected by BREWINGTON and CYBENKO [2000], which noticed that the older half of the distribution has a very long tail. NIELSEN [2007] points to this same fact in the growth of the Web, where the growth follow a power law increase in the initial years and them an exponential growth in the mature years of the Web. Exponential distributions are usual in scientometrics studies as scientific literature obsolescence [LINE, 1970; 1993] and e-journal usage along the time [NICHOLAS et al., 2005]. Although some papers have shown similar singularities in the fit of these distributions [EGGHE & RAVICHANDRA, 1992; BURRELL, 2002].

The split-half procedure allowed us to test whether age may be a factor that affects the size distribution in two samples. No difference was found between so age is not an important variable in the size distribution of web pages. This result is similar to the obtained by DOUGLIS, FELDMANN and KRISHNAMURTHY [1997]. They conclude that the age distribution is related to the content type, but not to the size which agrees with our results, because we have found difference in age distribution by formats, but not with size.

In the update part of the study we introduced two new indicators. WUI measures the change rate, i.e. the number of changes in a time period, while the WLI measures the lifespan of a web page, i.e. the time passed between changes. We think that both indicators allow us to quantify the dynamics of a web page or a web space. The indicators show that there is different update behaviour by discipline topic. We thus relate our findings to the differences by disciplines or “communities” found by FLAKE, LAWRENCE and GILES [2000] in inlink distributions. This suggests that web communities would show differences in their age and update distribution as well. It would be valuable to extend the study of the Web using these indicators in order to test their reliability over a longer time window and larger sample size.

Conclusions

This study characterizes the age distribution of the ERA Web space. The main finding is that the distribution can be described by an exponential decay. Taking this into account we have used an indicator called half-age. This indicator discriminates the behaviour of different European country Web domains. The UK (half-age=209 days) and Switzerland (half-age=518 days) Web spaces are the most youthful while the Greece (half-age=808 days) and Belgium (half-age=689 days) domains are the least youthful. The half-age indicator is a suitable tool to assess and value the freshness of a Web space because it is based on the whole of the age distribution unlike simpler indicators based on proportions of the age distribution.

The distribution of different file formats (hypertext, application and images) was also studied and they follow an exponential trend as well. However, there are differences between formats which reflect how their different utilities and uses in the web environment are appreciated in the age distribution of these formats. Hypertext formats have a shorter half life (458 days) than images (778 days) reflecting the essentially dynamic nature of text as the primary medium of dynamic Web page information compared to images which have a complementary or secondary role providing information that is comparatively more stable.

The update indicators have also measured the changing rate and dynamism of two set of web pages, detecting that Biology (WUI=.178) web pages are more stable than Physics ones (WUI=.193).

Assessing the dynamic nature of the Web is important. It allows us to discriminate between different Web environments. This in turn can lead to improvements in crawler and harvesting processes that can, for example, eliminate dead web pages. Comparing Web environments with stable and normalized indicators also allows proper comparison over time and between different countries and disciplines.

References

ADAMIC, I. A., HUBERMAN, B. A. (2001), The Web's Hidden Order, *Communications of the ACM*, 44(9): 55-59.

ADAMIC, L. A., HUBERMAN, B. A. (2000), Power-Law Distribution of the World Wide Web, *Science*, 287(5461): 2115.

ALBERT, R., JEONG, H., BARABASI, A. L. (1999), Internet - Diameter of the World-Wide Web, *Nature*, 401(6749): 130-131.

BAEZA-YATES, R., CASTILLO, C. (2005), Características de la web chilena 2004, Technical Report, Center for Web Research, University of Chile, Santiago de Chile. http://www.ciw.cl/webcl2004/Web_Chilena_2004.pdf

BAEZA-YATES, R., CASTILLO, C. (2007), Características de la web chilena 2006, Technical Report, Center for Web Research, University of Chile, Santiago de Chile. http://www.ciw.cl/material/web_chilena_2006/index.html

BAR-ILAN, J. (2001), Data collection methods on the Web for informetric purposes - A review and analysis, Scientometrics, 50(1): 7-32.

BAR-ILAN, J., PERITZ B. C. (1999), The life span of a specific topic on the Web; the case of 'Informetrics': A quantitative analysis, *Scientometrics,* 46(3): 371-382.

BAR-YOSSEF, Z., BRODER, A. Z., KUMAR, R., TOMKINS, A, (2004), Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay, In: S. I. FELDMAN, M. URETSKY, M. NAJORK, C. E. WILLS (Eds), *Proceedings of the 13th international conference on World Wide Web, WWW 2004*, ACM Press, New York, USA, pp. 328-337.

BARABASI, A. L., ALBERT, R. (1999), Emergence of Scaling in Random Networks, *Science*, 286(5439): 509-512.

BERNERS-LEE, T., CAILLIAU, R., GROFF, J. F., POLLERMANN, B. (1992), World-Wide Web: the information universe, *Internet Research*, 2(1): 52-8.

BORDIGNON, F. R. A., LAVALLÉN, P. J., TOLOSA, G. H. (2006), El Estado de la Web de Paraguay y la Sociedad de la Información, In: *Proceedings of the I Congreso Internacional y VI Congreso Nacional de Bibliotecarios, Documentalistas y Archivistas del Paraguay*, Asunción, Paraguay. http://eprints.rclis.org/archive/00007704/01/webpy.pdf

BORDIGNON, F. R. A., TOLOSA, G. H. (2006), Characterization of South American Educational Web Domains, In: *Proceedings Congreso Argentino de Ciencias de la Computación. CACIC 2006*, Potrero de los Funes, Argentina. http://eprints.rclis.org/archive/00007705/01/676-Educational_Webs___CACIC__English____v4.pdf

BREWINGTON, B. E., CYBENKO, G. (2000), How dynamic is the Web? *Computer Networks*, 33(1-6): 257-276. http://www9.org/w9cdrom/264/264.html

BURRELL, Q. (2002), The nth-citation distribution and obsolescence, *Scientometrics*, 53(3): 309-323(15)

CAILLIAU, R. (1995), A short history of the Web. Netvalley.com http://www.netvalley.com/archives/mirrors/robert_cailliau_speech.htm

CHO, J., GARCIA-MOLINA, H. (2000), The Evolution of the Web and Implications for an Incremental Crawler. In: *Proceedings of the 26th International Conference on Very Large Data Bases*, san Francisco, USA

COTHEY, V. (2004), Web-Crawling Reliability, *Journal of the American Society for Information Science and Technology*, 55(14): 1228-1238.

COTHEY, V. (2005), Some preliminary results from a link-crawl of the European Union Research Area Web. In: P. INGWERSEN, B. LARSEN (Eds), *Proceeding of the 10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden.

DOUGLIS, F., FELDMANN, A., KRISHNAMURTHY, B. (1997), Rate of change and other metrics: a live study of the World Wide Web, In: *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, Monterrey, USA, pp. 147-158 http://www.usenix.org/publications/library/proceedings/usits97/full_papers/douglis_rate/douglis_rate_html/douglis_rate.html

EGGHE, L., RAVICHANDRA, R. I. K. (1992), Citation age data and the obsolescence function: fits and explanations*, Information processing and management*, 28(2): 201-217.

FETTERLY, D., MANASSE, M., NAJORK, M., WIENER, J. L. (2003), A large-scale study of the evolution of web pages. In: *Proceedings of the 12th International World Wide Web Conference*, ACM Press, Budapest, Hungary, pp. 669-678.

FLAKE, G.W., LAWRENCE, S., GILES, L. (2000), Efficient Identification of Web Communities. In: *Proceeding of the Sixth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000)*, Boston, USA, pp. 150-160.

HARTER, S., KIM, H. (1996), Electronic journals and scholarly communication: a citation and reference study, *Information Research*, 2(1): paper 9a. http://informationr.net/ir/2-1/paper9a.html

INTERNET SYSTEMS CONSORTIUM (2004), Domain Survey Information. http://www.isc.org/index.pl?/ops/ds/.

KOEHLER, W. (1999), An Analysis of Web Page and Web Site Constancy and Permanence, *Journal of the American Society for Information Science*, 50(2): 162-180.

KOEHLER, W. (2002), Web Page Change and Persistence - a Four-Year Longitudinal Study, *Journal of the American Society for Information Science and Technology*, 53(2): 162-171.

KOEHLER, W. (2004), A Longitudinal Study of Web Pages Continued: a Consideration of Document Persistence, *Information Research*, 9(2): paper 174 http://informationr.net/ir/9-2/paper174.html

LAWRENCE, S., PENNOCK, D. M., FLAKE, G. W., KROVETZ, R., COETZEE, F. M., GLOVER, E., NIELSEN, F. A., KRUGER, A., GILES, C. L. (2001), Persistence of Web References in Scientific Research, *Computer*, 34(2): 26-31.

LEYDESDORFF, L., CURRAN, M. (2000), Mapping University-Industry-Government Relations on the Internet: The Construction of Indicators for a Knowledge-Based Economy, *Cybermetrics*, 4(1): paper 2

LINE, M.B. (1970), The 'half-life' of periodical literature: Apparent and real obsolescence, *Journal of Documentation*, 26, 46–54.

LINE, M.B. (1993), Changes in the literature with time—Obsolescence revisited, *Library Trends*, 41, 665–683.

NELSON, M., ALLEN, B. (2002), Object persistence and availability in digital libraries. *D-Lib Magazine*, 8(1). http://www.dlib.org/dlib/january02/nelson/01nelson.html

NETCRAFT LTD (2007), Web Server Survey. http://news.netcraft.com/archives/web_server_survey.html

NICHOLAS, D., HUNTINGTON, P., DOBROWOLSKI, T., ROWLANDS, I., JAMALI, H. R., POLYDORATOU, P. (2005), Revisiting `obsolescence' and journal article `decay' through usage data: an analysis of digital journal use by year of publication, *Information Processing & Management*, 41(6): 1441-1461.

NIELSEN, J. (2007), 100 Million Websites. http://www.useit.com/alertbox/web-growth.html

O'NEILL, E. T., LAVOIE, B. F., BENNET, R. (2003), Trends in the Evolution of the Public Web 1998-2002, *D-Lib Magazine*, 9(4). http://www.dlib.org/dlib/april03/lavoie/04lavoie.html

ORTEGA, J. L., AGUILLO, I. F., COTHEY, V., SCHARNHORST, A. (2008), Maps of the academic web in the European Higher Education Area - an exploration of visual web indicators. *Scientometrics*, 74(2): 295-308.

ORTEGA, J. L., AGUILLO, I. F., PRIETO, J. A. (2006), Longitudinal Study of Contents and Elements in the Scientific Web environment, *Journal of Information Science*, 32(4): 344-351.

PAYNE, N., THELWALL, M. (2008), Longitudinal trends in academic web links, *Journal of Information Science*, 34(1): 3-14.

PENNOCK, D. M., FLAKE, G. W., LAWRENCE, S., GLOVER, E. J., GILES, C. L. (2002), Winners Don't Take All: Characterizing the Competition for Links on the Web, *Proceedings of the National Academy of Sciences of the United States of America*, 99(8): 5207-5211.

PRICE, D. D. S. (1976), A general theory of bibliometric and other cumulative advantage processes, *Journal of the American Society for Information Science*, 27(5): 292-306.

RIVEST, R. (1992), The MD5 Message Digest Algorithm, *Internet RFC 1321*. http://people.csail.mit.edu/rivest/Rivest-MD5.txt

ROUSSEAU, R. (1997), Sitations: an exploratory study, *Cybermetrics*, 1(1): paper 1. http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

RUTHERFORD, E. (1900), A Radioactive Substance emitted from Thorium Compounds, *Philosophical Magazine*, 49(5): 1-14.

SPINELLIS, D. (2003), The decay and failure of Web references, *Communications of the ACM*, 46(1): 71-77.

TOLOSA, G., BORDIGNON, F., BAEZA-YATES, R., CASTILLO, C. (2007), Characterization of the Argentinian Web, *Cybermetrics*, 11(1): paper 3. http://www.cindoc.csic.es/cybermetrics/articles/v11i1p3.html