

Longitudinal Study of Contents and Elements in the Scientific Web environment¹

José Luis Ortega, Isidro F. Aguillo and José Antonio Prieto

Internet Lab, Centro de Información y Documentación Científica (CSIC). Joaquín Costa, 22. 28002 Madrid (Spain). E-mail: {jortega, isidro}@cindoc.csic.es

Abstract.

The aim of this work is the longitudinal study of the evolution and the state of 738 web sites in two different points in time (1997 and 2004). It tries to establish the rate of growth and decay of the Web and all the web elements. To this end, the structure and the contents of these web sites are extracted through a crawler and compared at the two different moments in time. The main results confirm a growth of web contents and elements in the web, although there is also a high degree of web content decay. The results suggest that in the seven year period covered by this study the web is characterized by both strong dynamism and instability.

Keywords: Webometrics; Web persistence; Web growth; Web decay; Linkrot

1. Introduction

Since the beginning of the World Wide Web, different growth behavior patterns have been studied. Pennock et. al. [1] discovered that the incoming links of a web site grow with time in accordance with a power law. According to Internet Systems Consortium [2] web domains are growing since 1994 with a similar rate. However, the OCLC Web Characterization Project [3], carried out between 1998 and 2000, warns that although the WWW keeps growing, contents contribution rates slowed by 1% in 2001-2002 period.

Nevertheless, there is a bibliographic gap about web decay or the disappearance of pages in the World Wide Web. Harter and Kim [4] were the first to study the ephemeral nature of the Web, detecting that a third of the electronic citations in e-

¹ This paper is a pre-print of: Ortega, J. L.; Aguillo, I. F., and Prieto, J. A. (2006) Longitudinal Study of Contents and Elements in the Scientific Web environment, *Journal of Information Science*, 32(4):344-351.

journals were not available. Lawrence et al. [5] also studied the problems of the electronics cites obtaining similar results. Koehler [6,7,8], one of the busiest authors in this field, monitored 360 pages and 343 web sites over several years, finding that in 2001 the operative pages had reduced 34.4% and in 2003, 33.8%. Nelson and Allen [9] tested the contents of different e-libraries during one year finding only 3% of unavailable objects (*linkrot*). However, they warn that these media are more stable than the rest of the World Wide Web and that their results have to be considered carefully. Fetterly et al. [10], continuing with the work of Cho and García-Molina [11], studied the evolution and persistence of 150 million pages for 11 weeks and found that the larger pages change more often and more deeply than the smaller ones.

Bar-Ilan and Peritz [12] queried “informetrics” using the most important search engines for 5 years, with the intention of studying the evolution of that discipline in the web, finding a disappearance rate of 40%. Wouters, Hellsten and Leydesdorff [13] studied the time span features of *Google* and *Altavista* and detected a great variability. While Ortega et al. [14] also detected that the query results of *Google* decayed according to the isotope radiation decay.

2. Objectives

The aim of this paper to study the state and evolution of 738 web sites in two different moments in time, 1997 and 2004. It intends to establish the increment and decrease of several of web objects, to detect the different growth patterns in the web sites studied and to describe the persistence of these objects with time. It also tries to analyse the relationship between several web elements with the intention of finding out their behaviour in these two moments in time. These web sites were crawled in 1997 and 2004, and the results compared with intention of analysing their evolution.

3. Methodology

In 1997, web sites were analysed by NetCarta.com [15]. This web site gathered the 1000 high quality web sites in terms of importance and contents. For this reason, most of these web sites are directories, e-libraries and information resources for

scientists. These web sites were analysed with the WebMapper 2.0 software of NetCarta. 921 of these web sites were downloaded to develop this study.

In 2004, with the intention of comparing the results obtained in 1997, these web sites were again analysed with the software Microsoft Site Analyst. This software was used because WebMapper was acquired by Microsoft, and merged with Site Analyst. In this way, Microsoft Site Analyst was the only software that could open the reports generated in 1997. For this reason, this study is limited to the features of this software and the elements arrangement supplied by this commercial crawler. This software works at different levels and it defines one web site according to the URL inserted. Thus, a web site can be a institutional domain, a directory or a unique page, and then it extracts information only of these unities. Table 1 shows the elements that Site Analyst generates in the crawl process and that are analysed in this study [16].

| <i>Element</i> | <i>Description</i> |
|----------------|---|
| Images | GIF, JPEG, and other types of images. |
| Gateways | Representations of CGI Scripts. |
| Internet | links to FTP, Telnet, Mailto, WAIS, NNTP, Gopher, and all other Internet services (except HTTP) |
| Applications | Java applets, executable files, PDF files, Microsoft Word documents, PostScript files, and other applications |
| Audio | WAV, AIFF, AU, and other audio files |
| Video | MPEG and other video file types. |
| Text | TXT files and other text files (other than HTML pages), including plain text. |
| Pages | Number of pages in the web site |
| Internal Links | links from the web site that point to its own pages |
| Outlinks | links from the web site that point to pages in other web sites. |

Table 1. Elements generated by Site Analyst and their description.

At first observation, less than half of these web sites had changed their address; concretely 427 (46.3%) and 183 (19.8%) had disappeared or had produced failures in the conversion to Microsoft Site Analyst, since to compare both crawls it was necessary to open again the Webmapper files in Site Analyst; and only 311 (33.7%) are remained constant. Finally, apart from the disappeared and faulty web sites, 738 web sites were

analysed. The following URL contains (<http://internetlab.cindoc.csic.es/cv/11/listado.htm>) the 921 resources obtained in 1997 and the 738 analysed in 2004.

Next, the data of each web site were extracted from the final reports of Microsoft Site Analyst through a little software programmed in VBS, and were recorded in a Microsoft Access database. Finally, they were analysed in a Microsoft Excel spreadsheet.

4. Research field

Web sites analysed are significant research web sites, which have been working from 1997 until 2004. These web sites are characterised by having a great volume of information and act as an information resource to the scientific community.

Table 2 shows the distribution of these web sites according to the institutional domain. More than half the web sites belong to the academic and scholarly domain (56.91%), followed by a considerable government presence (18.56%). Nevertheless, the economic sector only represents 10.03%. As we can see, the commercial sector was hardly present in 1997, as the Web was almost exclusively used by academics, and the non profit sector takes up the whole web.

| <i>Sectors</i> | <i>Web sites</i> | <i>Percentage</i> |
|----------------|------------------|-------------------|
| University | 420 | 56.91% |
| Government | 137 | 18.56% |
| Organisations | 107 | 14.50% |
| Commercial | 74 | 10.03% |
| TOTAL | 738 | 100.00% |

Table 2. Web sites by institutional sector.

In the following Table 3, the web sites have been presented by country, first, from the TLD of each site and then from an heuristic exploration. The web sites of United States are more than half of the sites studied (52.85%), followed at a distance by United Kingdom (7.45%) and Canada (6.23%). However, there are minor presence of French (1.22%) and Japanese (0.95%) web sites. It is understandable that the United States takes up all the net and nevertheless it is surprising that other countries, who carry a considerable weight in science, were poorly represented, such as France and Japan, which could suggest that the Web was still expanding.

| Countries | Web Sites | Percentage |
|-----------------|------------|---------------|
| USA | 390 | 52.85% |
| UK | 55 | 7.45% |
| CA | 46 | 6.23% |
| DE | 32 | 4.34% |
| IT | 25 | 3.39% |
| AU | 22 | 2.98% |
| FI | 10 | 1.36% |
| FR | 9 | 1.22% |
| NL | 9 | 1.22% |
| JP | 7 | 0.95% |
| Other Countries | 133 | 18.02% |
| TOTAL | 738 | 100.00 |

Table 3. Web sites by country TLD.

5. Results

Next, the result of the crawl process carried out in 2004 and its comparison with the initial data of 1997 is discussed.

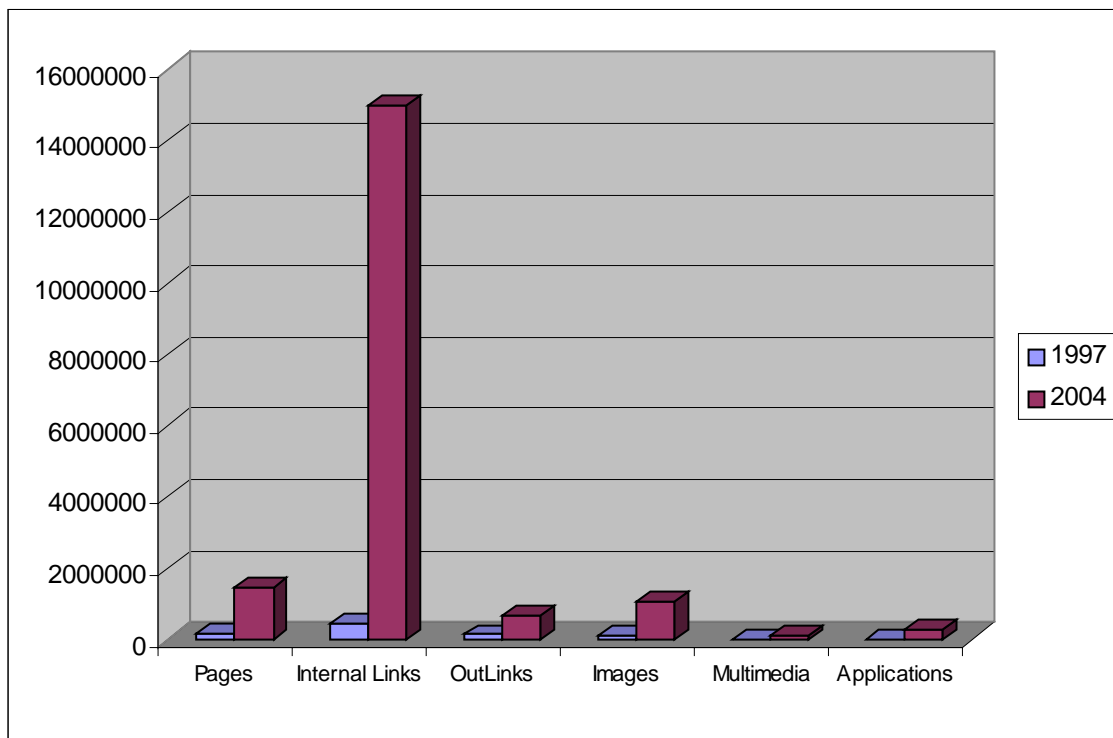


Figure 1. General evolution in the elements of the web sites.

| <i>Web elements</i> | <i>1997</i> | <i>2004</i> | <i>Growth</i> |
|---------------------|----------------|-------------------|---------------|
| Pages | 183,488 | 1,444,156 | 7.87 |
| Internal Links | 458,456 | 15,000,486 | 32.72 |
| OutLinks | 145,092 | 677,533 | 4.67 |
| Images | 102,504 | 1,076,383 | 10.50 |
| Multimedia | 14,401 | 92,663 | 6.43 |
| Applications | 10,523 | 266,572 | 25.33 |
| TOTAL | 914,464 | 18,557,793 | 20.29 |

Table 4. General growth of elements in web sites.

Table 4 shows the growth in the number of web elements substantially, up by a factor of 20.29. The elements that show the highest rate of growth are Internal Links (32.72 times) and Applications (25.33 times), and the elements that show a lower growth rate are Outlinks (4.67 times) and the Multimedia element (6.43 times). It is significant that the number of pages, the main element in a web site, only increased by 7.87 times.

Figure 1 shows the high number of Internal Links with regards to the other elements in 2004. This could be due to the improvement of the pages navigability, due to both an improvement in the quality of the information architecture and the web design, because, as Koehler [8] saw, the percentage of navigational pages increases with respect to the number of content pages over time, confirming the proliferation of internal links for navigational reasons.

Another element with a significant increase is Applications. This suggests a growing use of scripts and programming languages used to build web pages such as ASP or PHP. It is necessary to say that Applications contain text formats such as .pdf (Portable Document Format) and .ps (PostScript), which are the formats used mainly in the Web for the spreading of scientific results (articles, informs, reports, etc.), and suggests the adoption of new formats to disseminate the knowledge in the web. It can be seen that in 1997 the number of Images was smaller than the number of Outlinks, and are now almost double, which confirms the great weight of the graphical elements in the present web design.

| Web elements | 1997 | 2004 | Growth |
|----------------|---------------|----------------|-------------|
| Gateways | 4,735 | 30,382 | 6.42 |
| Other Protocol | 52,089 | 143,658 | 2.76 |
| Audio | 667 | 3,484 | 5.22 |
| Video | 856 | 5,885 | 6.88 |
| Text | 9,321 | 51,711 | 5.55 |
| Other Media | 12,878 | 83,294 | 6.47 |
| TOTAL | 80,546 | 318,414 | 3.95 |

Table 5. Increase of multimedia and other elements.

Table 5 shows the increase of the remaining elements. Note that the Audio, Video and Other Media are included in the Multimedia element of the previous table. The increase of these other elements is much lower than the elements studied before (3.95 times). The element with the highest growth is Video (6.88 times) and the lowest is Other Protocol (2.76 times). Thus, the multimedia elements (audio, video, etc.) have not increased much, probably due to the low use of these formats to diffuse scientific results, although they were already introduced in the web some time ago.

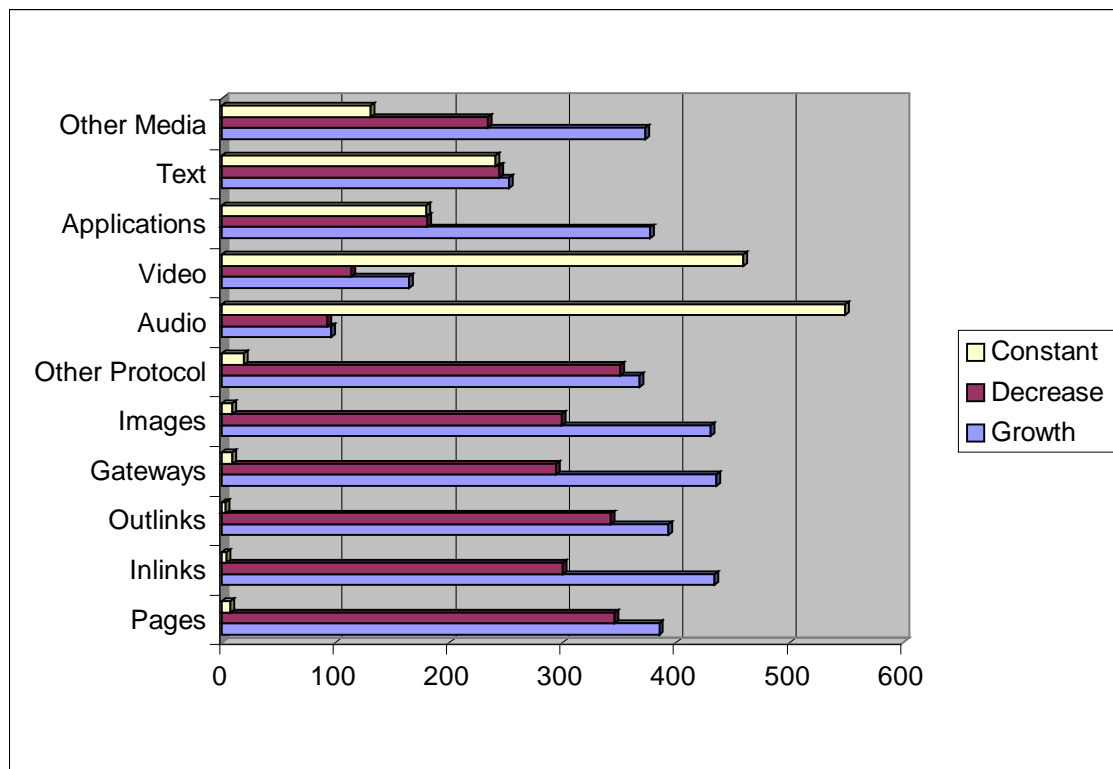


Figure 2. Percentage and increase of web sites by type of element.

| <i>Element</i> | | <i>sites</i> | <i>%</i> | <i>1997</i> | <i>2004</i> | <i>Increment</i> |
|----------------|--------|--------------|----------|-------------|-------------|------------------|
| Pages | < 1997 | 385 | 52.17 | 52336 | 1421896 | 27.17 |
| | > 1997 | 346 | 46.88 | 131107 | 22215 | -0.17 |
| | = 1997 | 7 | 0.95 | 45 | 45 | 0.00 |
| Internal Links | < 1997 | 434 | 58.81 | 175605 | 14954372 | 85.16 |
| | > 1997 | 300 | 40.65 | 282850 | 46113 | -0.16 |
| | = 1997 | 4 | 0.54 | 1 | 1 | 0.00 |
| Outlinks | < 1997 | 393 | 53.25 | 41885 | 650202 | 15.52 |
| | > 1997 | 342 | 46.34 | 102301 | 26425 | -0.26 |
| | = 1997 | 3 | 0.41 | 906 | 906 | 0.00 |
| Gateways | < 1997 | 435 | 58.94 | 1274 | 29695 | 23.31 |
| | > 1997 | 294 | 39.84 | 3383 | 609 | -0.18 |
| | = 1997 | 9 | 1.22 | 78 | 78 | 0.00 |
| Images | < 1997 | 430 | 58.27 | 34637 | 1062773 | 30.68 |
| | > 1997 | 299 | 40.51 | 67553 | 13296 | -0.20 |
| | = 1997 | 9 | 1.22 | 314 | 314 | 0.00 |
| Other Protocol | < 1997 | 368 | 49.86 | 8434 | 138605 | 16.43 |
| | > 1997 | 351 | 47.56 | 43583 | 4981 | -0.11 |
| | = 1997 | 19 | 2.57 | 72 | 72 | 0.00 |
| Audio | < 1997 | 96 | 13.01 | 68 | 3447 | 50.69 |
| | > 1997 | 93 | 12.60 | 599 | 37 | -0.06 |
| | = 1997 | 549 | 74.39 | 0 | 0 | 0.00 |
| Video | < 1997 | 165 | 22.36 | 72 | 5746 | 79.81 |
| | > 1997 | 114 | 15.45 | 784 | 139 | -0.18 |
| | = 1997 | 459 | 62.20 | 0 | 0 | 0.00 |
| Applications | < 1997 | 377 | 51.08 | 2630 | 265448 | 100.93 |
| | > 1997 | 181 | 24.53 | 7871 | 1102 | -0.14 |
| | = 1997 | 180 | 24.39 | 22 | 22 | 0.00 |
| Text | < 1997 | 253 | 34.28 | 329 | 51150 | 155.47 |
| | > 1997 | 244 | 33.06 | 8980 | 549 | -0.06 |
| | = 1997 | 241 | 32.66 | 12 | 12 | 0.00 |
| Other Media | < 1997 | 373 | 50.54 | 1414 | 82347 | 58.24 |
| | > 1997 | 234 | 31.71 | 11419 | 902 | -0.08 |
| | = 1997 | 131 | 17.75 | 45 | 45 | 0.00 |

Table 6. Percentage and increase of web sites by type of element.

Figure 2 and Table 6 show the behaviour of the web sites according to the elements studied. Table VI illustrates the number of web sites where each element have increased, decreased or remained constant since 1997. For instance, there are 253 (34.28%) web sites that have increased their number of plain text files (Text element)

by 155.47 times since 1997, 244 (33.06%) web sites that have decreased the number by 0.06 times and 241 (32.66) sites that have the same number of files as in 1997. In this way, it is seen how the increase of different elements affects certain web sites. Figure 2 also shows the infrequent use of the Video, Audio and Text elements since 1997.

The number of sites which show increases in the most important formats such as Pages, Internal Links, Outlinks, Images, Gateways and Other Protocol is similar to the number of sites showing decreases. Thus, although all the elements have grown, there is a significant percentage of sites in which some elements have decreased. This allows us to observe that the widespread growing seen in Table V is not present in all web sites studied, but the increase and decrease pattern is irregular. From this we can say that there is not a unique pattern to the evolution of the different elements of a web site.

| | <i>Added</i> | <i>Changed</i> | <i>Vanished</i> |
|----------------|-----------------|----------------|-----------------|
| Pages | 1521.32% | 17.09% | 80.67% |
| Images | 2160.77% | 11.17% | 80.34% |
| Gateways | 566.15% | 4.32% | 65.08% |
| Media | 2678.75% | 7.56% | 65.49% |
| Internet | 901.89% | 10.62% | 77.72% |
| Average | 1566.16% | 10.76% | 75.22% |

Table 7. Average of added, changed and missed elements.

Table 7 shows the persistence of several web elements relative to the crawl carried out in 1997. First of all, the percentage of added elements in all cases is very high, confirming the strong increase of the WWW in these seven years. 17% of Pages element has changed their URL over the seven years, which is close to the rate of 2,2% per year detected by Koehler [6]. The average of all elements that have changed their URL represents only a small percentage (10,76%). This indicates that the level of content reorganisation is low unless this leads to a modification of the page content, although the unchanged elements (24,88%) have a large redistribution. However, the percentage of vanished elements (75,22%) is very high because only 2 of 10 elements remain unchanged since 1997, indicating the low level of contents persistence in the web. The percentage of unchanged pages 19,3%, (2,7% per year) is also in line with Koehler [6].

The relationship between different elements have been studied, with the aim of detecting how the evolution of certain element affects others. Tables 8 and 9 show two correlation matrices among the most significant elements.

| 1997 | Images | Internet | Pages | Outlinks |
|----------|----------------|----------------|----------------|----------------|
| Images | 1.000 | 0.079 | 0.684** | 0.231 |
| Internet | 0.079 | 1.000 | 0.116 | 0.680** |
| Pages | 0.684** | 0.116 | 1.000 | 0.195 |
| OutLinks | 0.231 | 0.680** | 0.195 | 1.000 |

** Correlation is significant at the 0.01 level (2-tailed).

Table 8. Correlation matrix between the main elements in 1997.

For 1997 (Table VIII), there is a high correlation degree between the rest of the services of Internet and Outlinks ($\rho=0.68$), that suggest that the use of Outlinks was only designed to connect the Web with other Internet services, confirming that this was a period when the WWW had not yet absorbed the rest of the services (FTP, Telnet, Mail, etc.). Also there is correlation between Pages and Images ($\rho=0,684$), that suggest that the graphics were an important part in the design of the web pages at that time.

| 2004 | Images | Gateways | Internet | Applications | Text | Pages | Internal Links | Outlinks |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Images | 1.000 | 0.297 | 0.167 | 0.704** | 0.266 | 0.938** | 0.487 | 0.273 |
| Gateways | 0.297 | 1.000 | 0.189 | 0.278 | 0.104 | 0.406 | 0.411 | 0.464** |
| Internet | 0.167 | 0.189 | 1.000 | 0.201 | 0.165 | 0.232 | 0.317 | 0.445** |
| Applications | 0.704** | 0.278 | 0.201 | 1.000 | 0.713** | 0.748** | 0.372 | 0.262 |
| Text | 0.266 | 0.104 | 0.165 | 0.713** | 1.000 | 0.368 | 0.152 | 0.147 |
| Pages | 0.938** | 0.406 | 0.232 | 0.748** | 0.368 | 1.000 | 0.632** | 0.323 |
| Internal Links | 0.487 | 0.411 | 0.317 | 0.372 | 0.152 | 0.632** | 1.000 | 0.297 |
| Outlinks | 0.273 | 0.464** | 0.445** | 0.262 | 0.147 | 0.323 | 0.297 | 1.000 |

** Correlation is significant at the 0.01 level (2-tailed).

Table 9. Correlation matrix between the main elements in 2004.

In 2004, there are changes in the correlations. Table 9 highlights that Images have increased their correlation with Pages ($\rho=0.938$), indicating the heavy presence of the graphical formats in the present web design and that the creation of pages runs parallel to the growth in images. Moreover, in 2004 there is a new correlation between Pages and Applications ($\rho=0.748$), because the Applications element contains textual formats such as pdf and ps suggesting that there is a meaningful relationship between pages and the supply of new scientific contents in different formats. On the other hand, the Application element also contains dynamic pages (ASP, PHP) which reinforce this relationship because these are one type of web pages. There is also an important correlation between Outlinks and Gateways ($\rho=0.464$), that demonstrates the proliferation of web-based databases. In the same sense, the correlation detected in 1997 between Internet and Outlinks decreased in 2004 ($\rho=0.445$), confirming the hegemonic presence of the Web with respect to the rest of the Internet services, because the

outlinks point now to the Web more than other Internet services. Finally, the correlation between Internal Links and Pages ($\rho=0.632$) confirms the spread of navigational pages because the Internal Links act as structural elements which organise the pages of a web site. Therefore the more pages there are, the more internal links will exist to arrange these contents.

6. State and Permanence

The permanence and stability of the outlinks of the 738 web sites in 1997 were studied. 145,092 outlinks were counted and checked with the software Xenu's Link Sleuth [17]. In Table 10 we can show the distribution of the outlinks according to their status in 2004.

| <i>Status</i> | <i>Frequency Percentage</i> | |
|-------------------|-----------------------------|----------------|
| not found | 36908 | 25,44% |
| Ok | 36752 | 25,33% |
| object moved | 28363 | 19,55% |
| no such host | 24026 | 16,56% |
| Timeout | 13351 | 9,20% |
| Forbidden request | 2558 | 1,76% |
| no connection | 2264 | 1,56% |
| server error | 177 | 0,12% |
| invalid path | 121 | 0,08% |
| Redirection | 103 | 0,07% |
| Other | 469 | 0,32% |
| TOTAL | 145092 | 100,00% |

Table 10. Outlinks status.

The percentage of valid links, if we consider the redirections, is very low (25,40%) while the group of broken links (*linkrot*) or no operative is almost three quarters of all outlinks (74,28%). This percentage is similar to the average number of missed elements (75,22%) as shown in Table 7. This suggests that the number of missed elements and the percentage of broken links have a similar relationship with respect to stability, because the more elements disappear the more broken links will exist.

7. Conclusions

This study shows two different moments in the evolution of the Web. On the one hand, in 1997 the Web was a young service that was yet to gain prominence. On the

other hand, in 2004 this service was consolidated in Internet as the main gateway to access to the net. This longitudinal view demonstrates that the Web, since 1997, has been characterised by an exponential growth, although the rate of growth of different web elements (pages, links, formats, etc.) is not the same. As we have seen in Table VI, the growth pattern for every element in a web site is yet to be determined. Certain web sites increase in one element or decrease in others in similar percentages. This is why we consider that it is hard to know the evolution of the Web in general because each web site evolves in a particular form. The high standard deviation in the objects distribution detected by Koehler [6], confirms our assessments.

We think that to estimate the evolution of the Web is a very complex task and that in order to do this it is necessary to take a wide and heterogeneous sample to obtain satisfactory results. This sample is limited to the scientific field and it can not be extended to the whole Web. Moreover, this sample represents directories and information sources, which is why the results are only representative of this type of web pages.

However, and according to the results obtained in this survey, we can claim that the observed growth is due to the high contribution of contents which hides the substantial elimination rate of the Web, or, phrasing it differently, the Web grows at the expense of the deleting of previous contents. This fact is reinforced by the contents and URLs persistence in these seven years. 75,22% of the original contents have disappeared and the broken links have increased in a similar percentage (74,28%). This fact is disguised by the strong contribution of new contents (1568%) in these seven years. In the future, we can ask if this rate of contents contribution could increase or decrease, if the Web will stop growing or if the contents will be more stable. We encourage future works to answer these questions.

On the other hand, This study has try to know the relationships between different web elements. The found correlations allow us to see how these relationships have developed (or not) between 1997 and 2004. For instance, the significant correlation between Pages and Image, which showed an increase over time (1997, $\rho=0.684$; 2004, $\rho=0.938$), suggests that the images are used more as a graphic element in the web pages design than instead of content itself. Nevertheless, the gradual lost of correlation between Outlinks and Internet (1997, $\rho= 0.68$; 2004, $\rho=0.445$) suggests that the links from the Web to other Internet services are disappearing due to large amount of Web contents and a gradual absence of the remaining services such as Telnet, Wais, Gopher

etc., in favour of the Web. The correlations between web elements show how these elements interact between themselves and how they structured one web site.

Both the growth of Internal Links (32.72 times) and Applications (25.33 times) demonstrate that the tested web sites have reached (a period of) maturity. On the one hand, the growth of Internal Links means more complexity in the design and structure as well as more content arrangement of one web site. On the other hand, the growth of Applications means there is a higher proportion of science-related contents, because these formats (pdf, ps, etc.) are used to publish final structural contents such as articles and reports. However, the low use of multimedia elements (6.43 times), suggests that many web sites use the web in a traditional way and do not fully exploit the facilities that the technology offers. We think that the Web is the best vehicle to disseminate scientific results in ways that are not easily done using more traditional methods, e.g the use of audio via the web by the Physics and Biology communities and the use of video via the web in Psychology or Surgery.

8. References

- [1] D. Pennock, G.W. Flake, S. Lawrence, E.J. Glover, C.L. Giles, Winners don't take all: Characterizing the competition for links on the web, *Proc. Natl. Acad. Sci. USA* 99 (8) (2002) 5207-5211 Available at: <http://www.pnas.org/cgi/reprint/99/8/5207> (accessed 28 October 2005).
- [2] Internet Systems Consortium, Inc, Redwood, CA. (2004). Available at: <http://www.isc.org/index.pl?/ops/ds/> (accessed 28 October 2005).
- [3] E.T. O'Neill, B.F. Lavoie, R. Bennet, Trends in the Evolution of the Public Web 1998-2002, *D-Lib Magazine* 9 (4) (2003).
- [4] S. Harter, H. Kim, Electronic journals and scholarly communication: a citation and reference study, *Information Research* 2 (1) (1996) paper 9. Available at: <http://informationr.net/ir/2-1/paper9a.html> (accessed 28 October 2005)
- [5] S. Lawrence, F. Coetzee, E. Glover, D. Pennock, G. Flake, F. Nielsen, B. Krovetz, A. Kruger, L. Giles, Persistence of Web References in Scientific Research, *IEEE Computer* 34(2) (2003) 26-31
- [6] W. Koehler, An Analysis of Web page and Web site constancy and permanence, *Journal of the American Society for Information Science*, 50 (2) (1999) 162-180
- [7] W. Koehler, Web page change and persistence – a four-year longitudinal study, *Journal of the American Society for Information Science and Technology*, 53 (2) (2002) 162-171

- [8] W. Koehler, A longitudinal study of Web pages continued: a report after six years, *Information Research*, 9 (2) (2004) paper 174. Available at: <http://informationr.net/ir/9-2/paper174.html> (accessed 28 October 2005)
- [9] M. Nelson, B. Allen, Object persistence and availability in digital libraries, *D-Lib Magazine* 8 (1) (2002). Available at: <http://www.dlib.org/dlib/january02/nelson/01nelson.html> (accessed 28 October 2005)
- [10] D. Fetterly, M. Manasse, M. Najork, J.L. Wiener, A Large-Scale Study of the Evolution of Web pages, *Software Practice and Experience* 1 (1) (2003) 1-27
- [11] J. Cho, H. García-Molina, The evolution of the web and implications for an incremental crawler, *Proceeding of the 26th International Conference on Very Large Databases*, (2000)
- [12] J. Bar-Ilan, B.C. Peritz, Evolution, Continuity, and Disappearance of Documents on a Specific Topic on the Web: A Longitudinal Study of 'Informetrics, *Journal of the American Society for Information Science and Technology*, 55 (11) (2004) 980-990
- [13] P. Wouters, I. Hellsten, L. Leydesdorff, Internet Time and the reliability of Search Engines, *First Monday*, 9 (10) (2004) Available at: http://www.firstmonday.org/issues/issue9_10/wouters/ (accessed 28 October 2005)
- [14] J.L. Ortega, J. A. Prieto, N. Arroyo, V.M. Pareja, I.F. Aguillo, Análisis de la persistencia y del estado de páginas web en los resultados de Google, *9^a Jornadas Españolas de Documentación FESABID 2005, Madrid, 14 y 15 de Abril* (2005). Available at: <http://internetlab.cindoc.csic.es/cv/11/Ortega2005.pdf> (accessed 28 October 2005)
- [15] NetCarta.com (1997) NetCarta WebMap Library. Available at: <http://www.netcarta.com/> (accessed 16 April 1997)
- [16] N. Arroyo, V. Pareja, I. Aguillo, Description of Web Data in D3.1. Deliverable. IST-1999-20350 (2003) Available at: http://www.eicstes.org/EICSTES_PDF/Deliverables/Web_Data_description.pdf (accessed 28 October 2005)
- [17] Xenu's Link Sleuth. Ver. 1.2f [s. l.]: Tilman Hausherr, c1997-2004. Software. Available at: <http://home.snafu.de/tilman/xenulink.html> (accessed 28 October 2005)