Report on Statistical Survey of World Wide Web Content

Web Content Amount as an Indicator of Internet Development in Japan

December 2004

Chigusa Saeki, Hiroya Shimada, and Shinya Tahata

Institute for Information and Communications Policy
Ministry of Internal Affairs and Communications

Contents

Figures

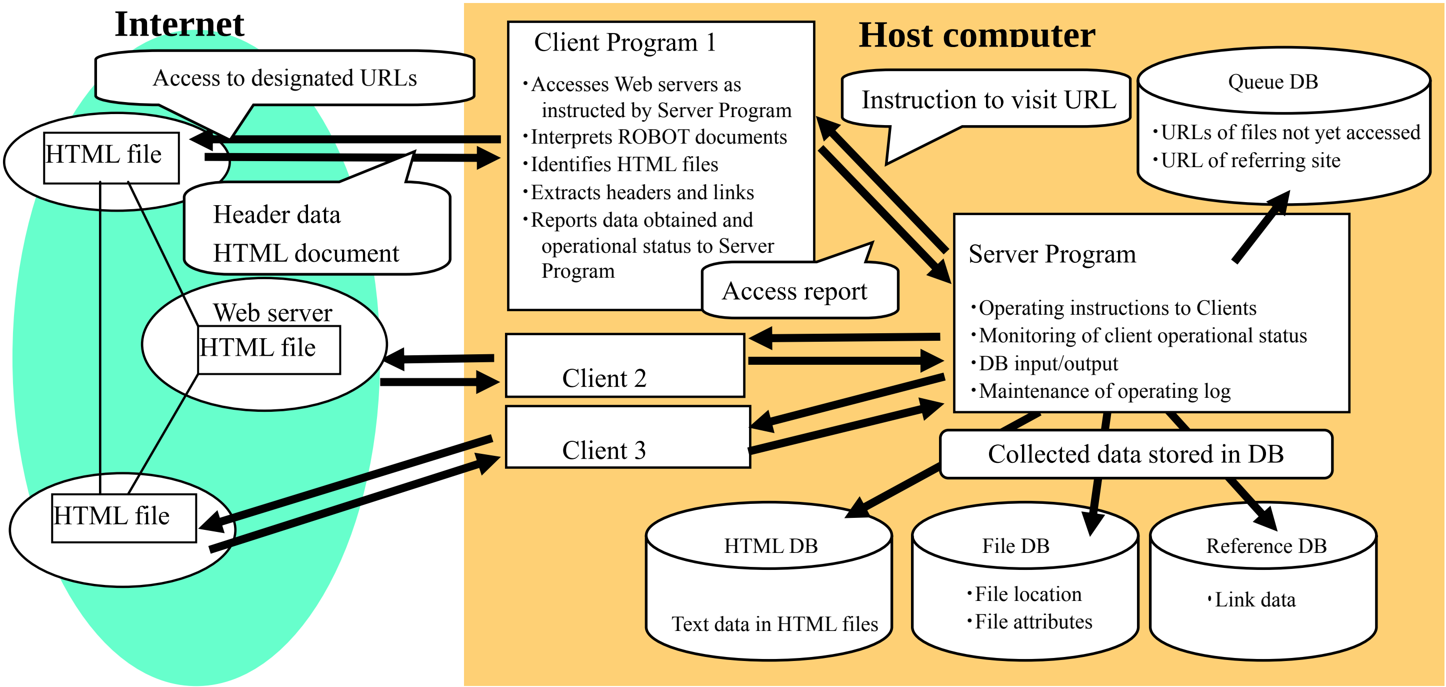
1. Overview of Survey

Since February 1998, the Institute for Posts and Telecommunications Policy (now the Institute for Information and Communications Policy, or IICP) has conducted surveys of domestic Japanese World Wide Web (hereafter “Web”) content amount. The statistical surveys adopt a unique combination of scans by search robots[1] and statistical estimations based on information collected by the scans, a method the IICP jointly developed with Allied Brains, Inc., a Japanese ICT consultancy.

Given the vast size of the Web and the speed of its expansion, it would be practically impossible, even with a sophisticated search robot, to access every Web page that has been published. Instead the search robot has been used to survey a certain number of pages on the Web in order to obtain information, such as uniform resource locations (URLs)[2] and the data volumes of the files corresponding to those URLs. After a period of scanning, the total numbers of Web servers, files,[3] and pages[4] and the total volume of data[5] are estimated by statistical means.

1.1 Survey Structure and Principles

Figure 1: The Structure of the “Loki” Search Robot


Internet
Access to designated URLs
HTML file
Header data
HTML document
Web server
HTML file
HTML file
Host computer
Client Program 1
・Accesses Web servers as instructed by Server Program
・Interprets ROBOT documents
・Identifies HTML files
・Extracts headers and links
・Reports data obtained and operational status to Server Program
Access report
Instruction to visit URL
Queue DB
・URLs of files not yet accessed
・URL of referring site
Server Program
・Operating instructions to Clients
・Monitoring of client operational status
・DB input/output
・Maintenance of operating log
Client 2
Client 3
Collected data stored in DB
HTML DB
Text data in HTML files
File DB
・File location
・File attributes
Reference DB
・Link data

[1] *Intānetto Kontentsu Tōkei ni Kansuru Chōsa Kenkyū* [Report on Survey of World Wide Web Content Statistics], in *Yūsei Kenkyūjo Geppō* [Monthly Report of the Institute for Posts and Telecommunications Policy], September 2002 (http://www.soumu.go.jp/iicp/seika/data/research/monthly/2002/168-h14_09/168-asearch2.pdf)
Media to shite no Web no Seichō o Hakaru: Sāchi Robotto o Tsukatta Web Kontentsu Tōkei Chōsa no Kokoromi [Measuring the Growth of the Web as a Medium: An Attempt to Conduct a Statistical Survey of Content Using Search Robots], in Masu Komyunikēshon Kenkyū [Mass Communication Research], No. 62, March 2003 (http://www.a-brain.com/HP/rep/rep15/index.html)

[2] A URL is a mechanism used to indicate the location of files, including Hypertext Markup Language (HTML) files, document/data files and image files, on the Web. Its contents include the type of information, the name and port number of the server, the folder name, and the file name.

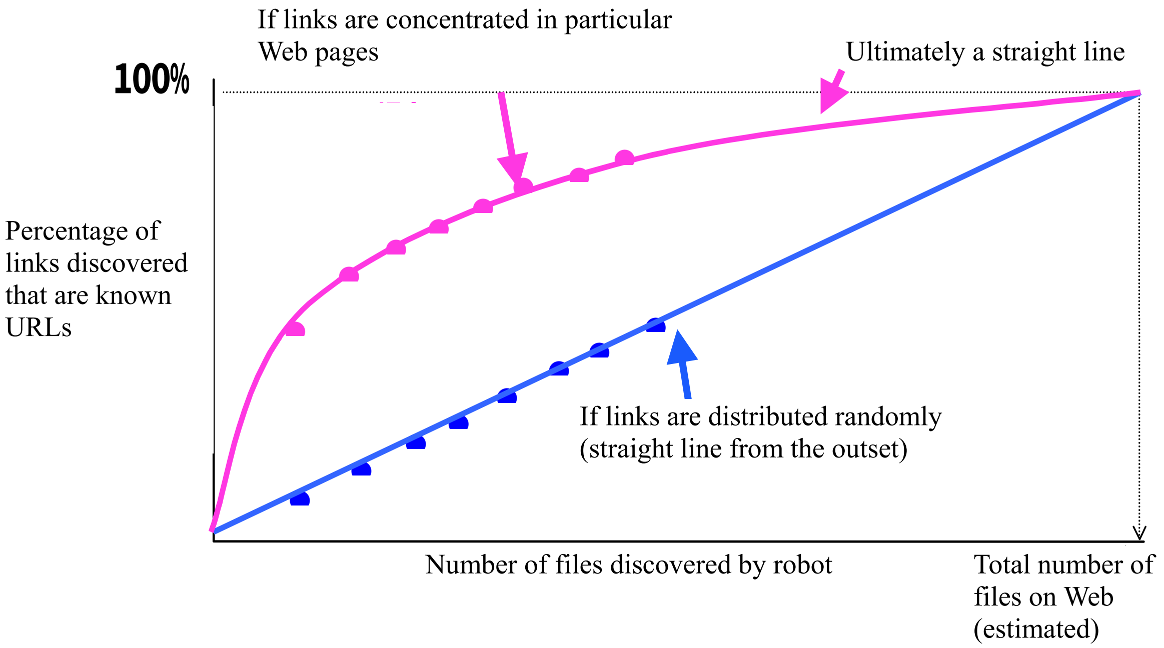
[3] For the purposes of this survey, a file is defined as an HTML file, document/data file, image file, video file, audio file, or other type of file used as content on websites with independent URLs.

[4] In the context of this survey, “page” is synonymous with “HTML file.” HTML is used to describe the logical structure, appearance, and other characteristics of a document. HTML files are called “pages” and put in a separate category from other file types because it is possible to embed various types of information, including images, audio, video, and links to other documents, in them.

[5] In addition to these totals, the amount of data in each type of file and the number of files are also estimated.

The robot program developed for the survey is named Loki. It automatically crawls along HTML hyperlinks to gather URL information about a wide range of Web pages on the Internet. To speed up the data gathering process, multiple client programs are activated to crawl the Web simultaneously. A server program meanwhile controls the operations of these client programs and organizes the data that they gather.

Figure 2: Changes in Discovery Rate for Known URLs



When the search robot accesses a new Web page, it gathers the link data contained in the HTML document. Some of the URLs will be unknown to the robot, while others may have been discovered previously. As the search robot's scan proceeds, the number of duplicate URLs already known to the robot ("known URLs") will increase, while the percentage of newly discovered URLs will decline rapidly. When the search robot has acquired URLs from all accessible Web pages, all of the links should be known URLs. By determining the percentage of known URLs in data newly obtained by the search robot once a certain number of URLs has been gathered, it is possible to estimate the percentage of total files on the Web represented by the URLs obtained up to that point. In this way, the total number of files on the Web can be estimated without accessing every Web page.

When using this method to estimate the total number of files on the Web, it is necessary to take into account the distribution of links on the Web. If links between Web pages were randomly and evenly distributed across the entire Web, the rate of discovery of known URLs by the search robot would be proportionate to the total number of files discovered, and the discovery rate for known URLs would rise in a straight line, as shown in Figure 2. In reality, however, there is significant bias in the distribution of URLs, with large numbers concentrated on a small number of popular Web pages. There is a high probability that the search robot will discover these popular pages early in its search, and once such pages have been discovered the discovery rate for known URLs will rise sharply. In sum, the discovery rate for known URLs is unlikely to rise in a straight line in actual surveys. The ratio can instead be expected to rise rapidly during the early stages of the survey and then shift to a more gradual rate of climb. Finally, as the distribution of links becomes more uniform, a pattern of gradual, straight-line increase should emerge. On this basis, the linear approximation method[6] is used to estimate the total number of files on the Web at the point when the line on the graph becomes straight.

1.2 Scope of Survey

The survey covered publicly accessible Web pages with .jp domains[7]. Recently there has been an increase in the number of domestic companies and other organizations in Japan that have sites with generic top level domains[8](hereafter "generic domains"), such as ".com". However, generic domains are not in the scope of this survey.[9]

The search robots cannot access every Web page. The following Web pages and files were excluded from the survey because they are inaccessible to the search robots.

- Web pages to which there are no external links
- Web pages where access is limited under membership schemes, etc.
- Web pages that use a "robots.txt" file to prohibit access by robots
- Web pages generated automatically by scripts, such as CGI programs[10]
- Web pages generated automatically by search engines
- Video files with embedded "Play" and "Stop" buttons
- Streaming video and audio files[11]

[6] The IICP (formerly the Telecommunications Economics Research Department of the Institute for Posts and Telecommunications Policy) and Allied Brains, Inc. have jointly filed a patent application.

[7] .jp" is the country-code top-level domain (ccTLD) for Japan. Second-level domains include organizational second-level domains, which indicate types of organizations, and general-use second-level domains, the names for which are chosen and registered by the person who acquires the domain.

[8] Generic top-level domain (gTLD) names have no country codes and can be acquired anywhere in the world. Examples include ".com", ".net", and ".org".

[9] The reason for limiting the survey to .jp domains is described in Section 3.1.1. Generic domains were included in the November 2002 survey on a single-year basis.

[10] Common Gateway Interface. This is a mechanism used to start programs in response to requests from a Web browser. In the past, Web servers simply transmitted stored documents. By using CGI, it is possible to dynamically create and transmit documents based on the results of program operations.

[11] Examples include files for Windows Media Player and RealPlayer.

2. Survey Results

2.1 Survey Record

The first search robot survey was implemented in February 1998. Thereafter surveys were carried out at half-yearly intervals (February and August), until the eighth survey in August 2001. Since 2002 the surveys have been conducted on an annual basis.

Figure 3: Crawling Surveys

	Period
1st	February 10–26, 1998
2nd	August 3–September 7, 1998
3rd	February 16–March 11, 1999
4th	August 4–September 26, 1999
5th	January 17–March 7, 2000
6th	August 30–September 27, 2000
7th	February 10–March 19, 2001
8th	July 20–October 30, 2001
9th	February 3–April 23, 2002
10th[12]	October 22–November 6, 2002
11th	January 5–February 26, 2004

[12] For the 10th survey, a search engine belonging to NTT-X, Inc. (now NTT Resonant Inc.) was used in the search robot.

2.2 Survey Results

2.2.1 Results of 11th Survey (February 2004)

2.2.1.1 Total Content Amount in .jp Domains

Substantial increase in total number of servers and total data volume, slow growth in total number of files

The total numbers of servers, files, and pages and the total volume of data were all higher than those of the previous survey (November 2002). The total data volume was estimated at 13,609 gigabytes (GB).[13]

The rates of increase in the total number of servers (37.99%) and the total data volume (34.08%) were greater than the rate of increase in the total number of files (6.39%). This suggests that there are now numerous servers that contain only a small number of files. The large increase in the total data volume compared with the increase in the total number of files also indicates that the data volume per file has risen.

Figure 4: Results of 11th Survey, February 2004

(Total Numbers of Servers, Files, and Pages, Total Data Volume)

	November 2002 (for reference)	February 2004
Servers	308,000	425,000
Increase (%)		37.99
Pages (millions)	74.38	85.90
Increase (%)		15.49
Files (millions)	274.21	291.73
Increase (%)		6.39
Data (GB)	10,150	13,609
Increase (%)		34.08

[13] A byte is a unit of information consisting of eight bits, each of which can express either a “0” or a “1.” A kilobyte (KB) consists of 1,024 bytes, and 1,024 kilobytes make a megabyte (MB), 1,024 megabytes a gigabyte (GB), and 1,024 gigabytes a terabyte (TB). To facilitate statistical processing, a kilobyte is defined here as 1,000 bytes, a megabyte as 1,000 KB, a gigabyte as 1,000 MB, and a terabyte as 1,000 GB.

2.2.1.2 Breakdown of Total Number of Files in .jp Domains by File Type

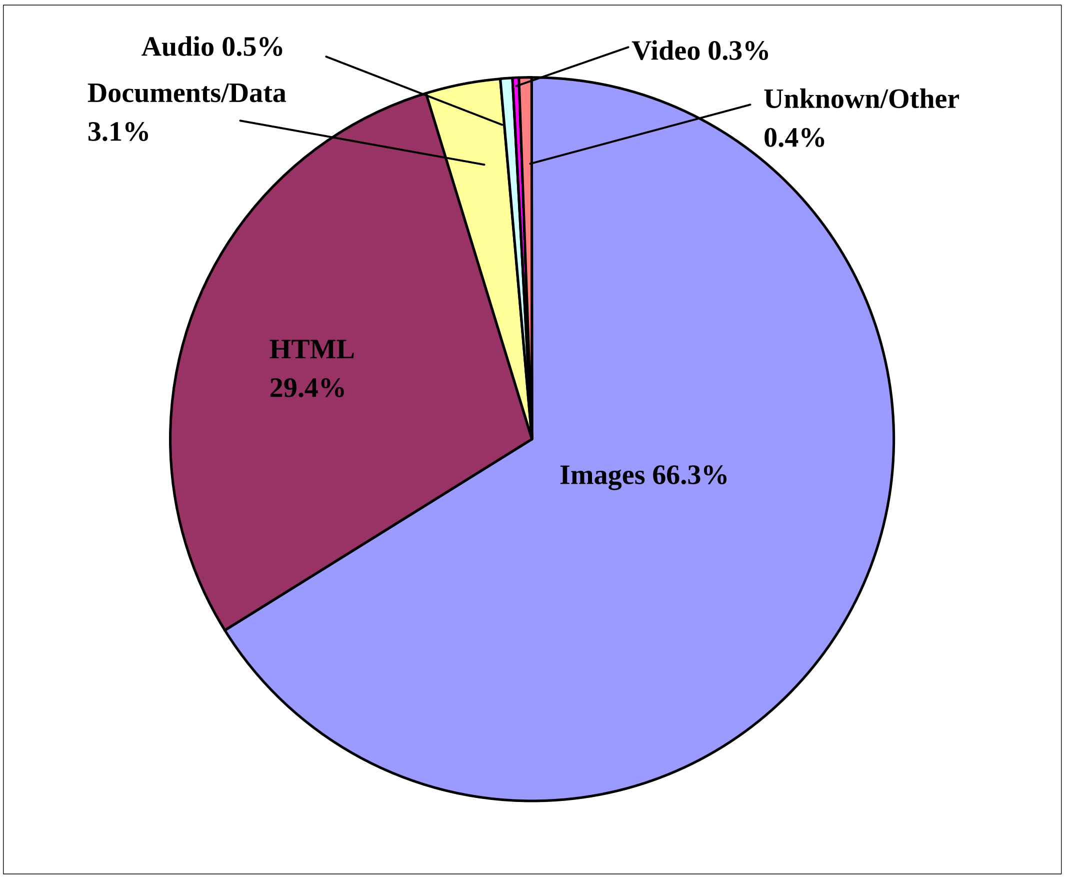
Image files the largest category, followed by HTML files

An analysis of the total number of files by file type[14] shows that image files are the largest category, accounting for 66.3% of the total, followed by HTML files (29.4%). Video and audio files make up 0.3% and 0.5% respectively of the total.

Figure 5: Breakdown of Total Number of Files by File Type (Based on Results of February 2004 Survey)

(%)

Images	HTML	Documents/Data	Audio	Video	Unknown/ Other
66.3	29.4	3.1	0.5	0.3	0.4



[14] Files are classified according to the extension at the end of the URL. The main file types and their extensions are as follows.

HTML:	.htm, .html
Images:	.jpg, .gif, .bmp, .pict, .tif, .eps, .png
Video:	.mpg, .avi, .mov
Audio:	.au, .ra, .midi, .mp3, .rmi, .wav
Documents/data:	.pdf, .txt, .doc, .jbw, .lzh, .tar, .xls, .exe, .java

2.2.1.3 Breakdown of Total Data Volume in .jp Domains by File Type

Video files make up three-tenths of data volume

A breakdown of total data volume by file type reveals that multimedia files, such as video and audio files, and document and data files account for large shares. Video files lead with a 29.1% share, followed by document and data files (26.0%), image files (25.4%), and audio files (12.1%).

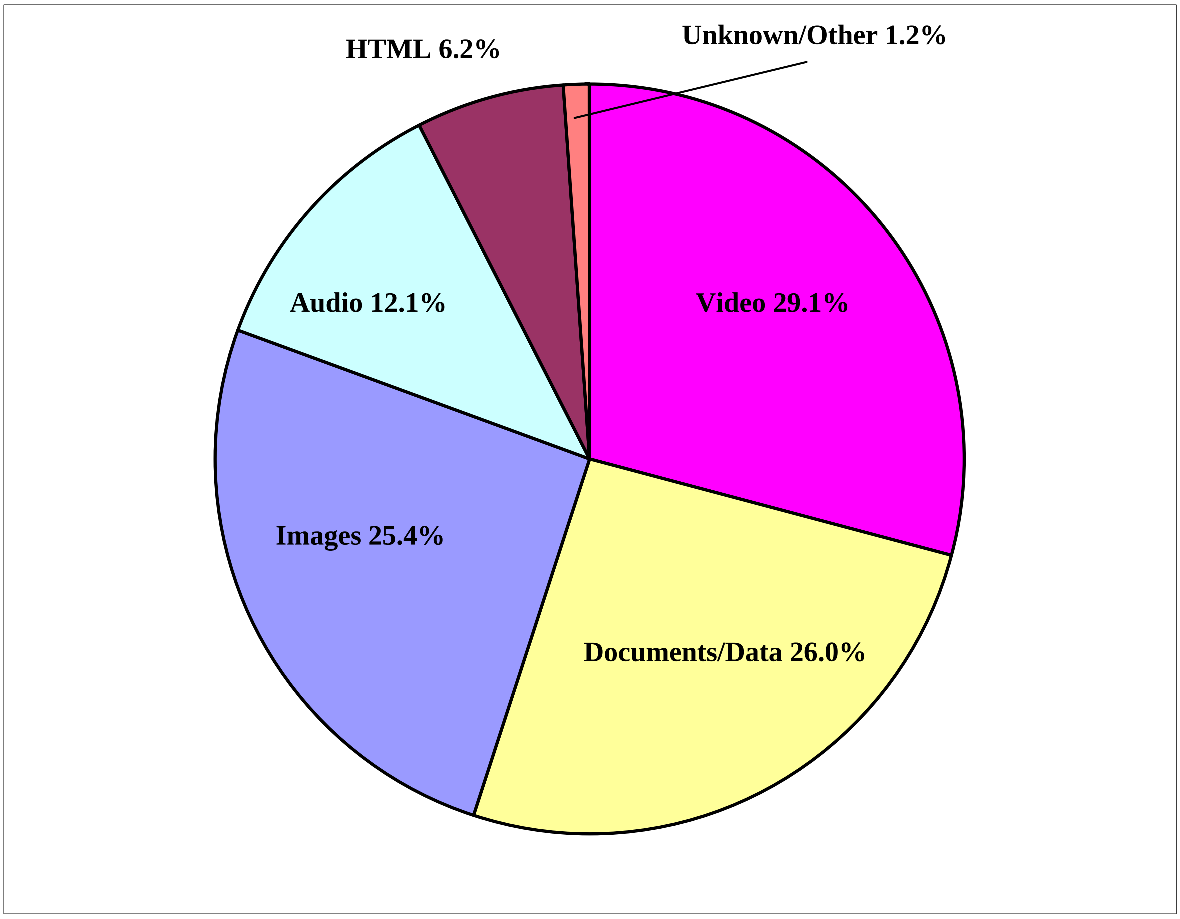
Though video and audio files account for over 40% in total data volume terms (Figure 6), they make up just 0.8% of the total number of files (Figure 5). This reflects the fact that video files are by far the biggest type in terms of data volume per file. Future growth in the percentage of video files and other types of multimedia files can be expected to result in sustained growth in the total volume of data on the Web.

Figure 6: Breakdown of Total Data Volume in .jp Domains by File Type

(Based on Results of February 2004 Survey)

(%)

Video	Documents/Data	Images	Audio	HTML	Unknown/ Other
29.1	26.0	25.4	12.1	6.2	1.2



2.2.2 Comparison with Past Survey Results

2.2.2.1 Trends in Total Number of .jp Domain Servers

Sustained growth in total number of servers

The total number of .jp servers has grown steadily since the first survey in 1998. Of particular interest is the sustained acceleration of the rate of increase over the past two years.

Figure 7: Total Number of .jp Domain Servers

	1998		1999		2000		2001		2002	2002	2004
	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Nov	Feb
Servers	36,000	54,000	75,000	85,000	95,000	120,000	152,000	177,000	197,000	308,000	425,000
Increase (%)	－	50.00	38.89	13.33	11.76	26.32	26.67	16.45	11.30	56.35	37.99

Decline in percentage of .ac domains, increase in .co domains

A breakdown of second-level domain categories shows that there has been a decline in the percentage of .ac domains, which were the largest category at the time of the first survey. The number of .co domains has meanwhile increased over the past four years, and this category now makes up the majority. This trend is indicative of a gradual shift in the role of the Web, from an information exchange tool for academic institutions when the surveys began, to a tool for business activity.

One conspicuous trend over the past few years has been the rapid growth of domains in the "other" category. In February 2004 domains in this category accounted for approximately 19.6% of the total number. The increase reflects the introduction and spread of general-use second-level domains[15] since 2001.

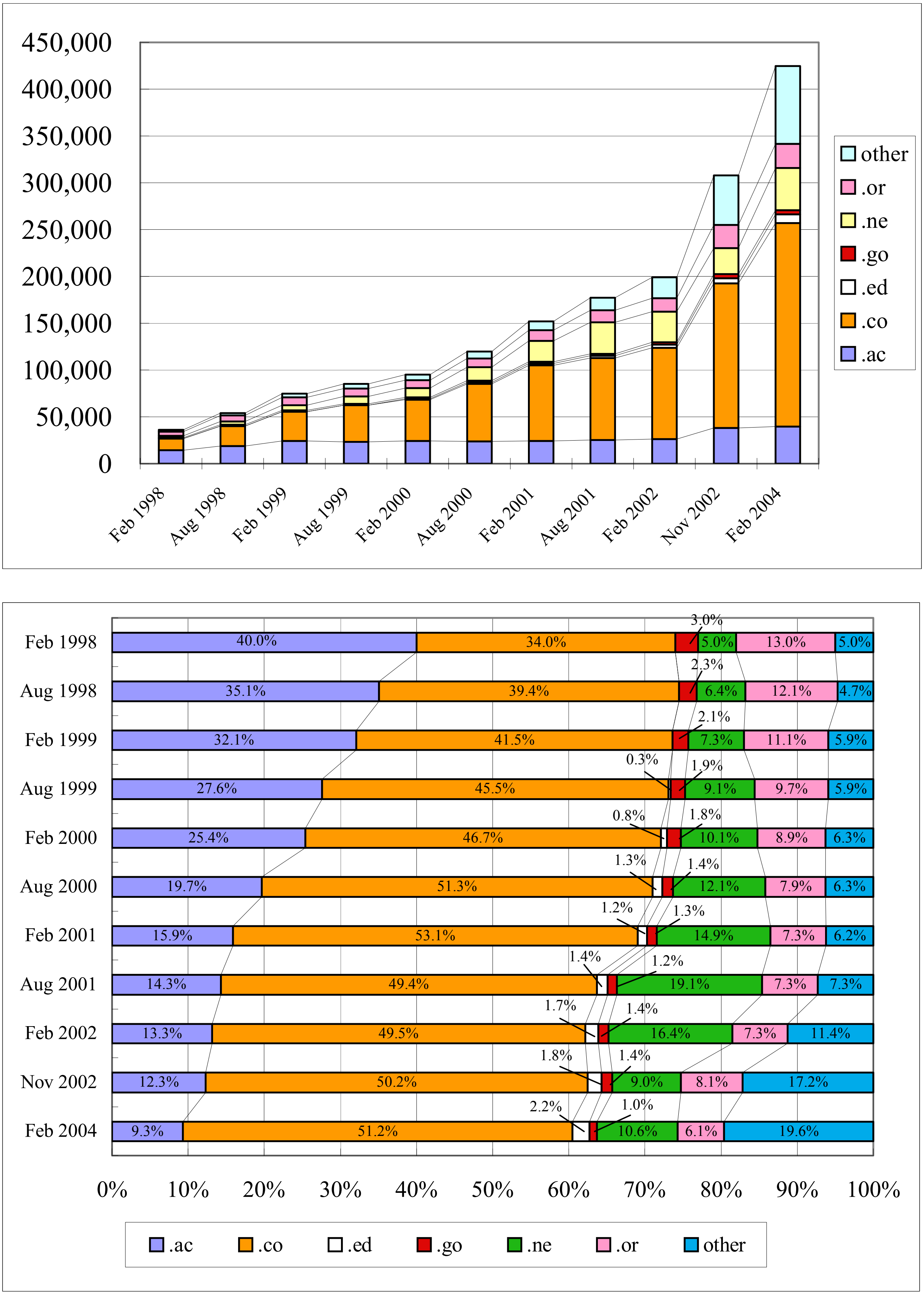
Figure 8: Principal Second-Level Domains

.ac	University-type educational institutions
.co	General businesses, etc.
.ed	Elementary schools, junior/senior high schools
.go	Government agencies
.ne	Network services, etc.
.or	Organizations, etc., other than businesses

[15] Organizational second-level domains are subject to several limitations, including the fact that only one can be used per organization. Any number of general-use second-level domains can be registered, provided there are no existing domains with the same names.

Figure 9: Breakdown of Total Number of .jp Domain Servers

by Second-Level Domain Type



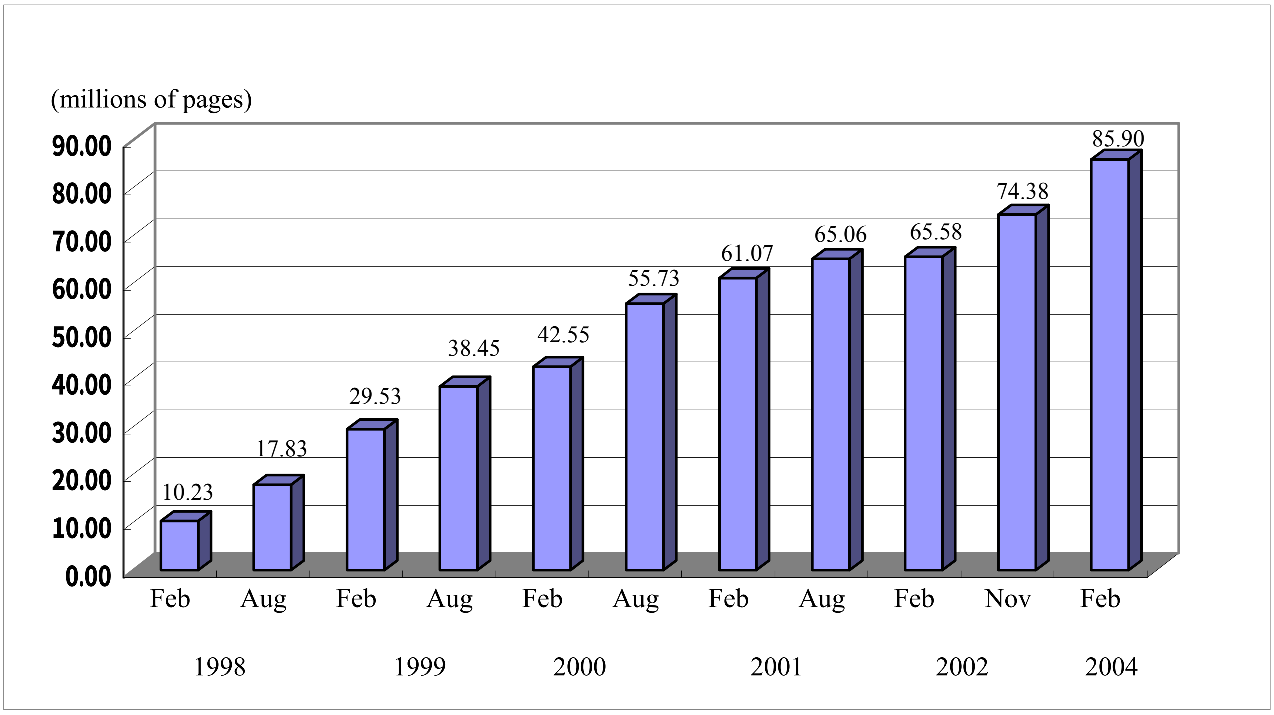
2.2.2.2 Trends in Total Number of Pages in .jp Domains

Temporary deceleration in growth of total number of pages followed by recovery in recent years

The rate of increase in the total number of pages in .jp domains began to decelerate in 2001. However, the pace of growth has risen over the past two years.

Figure 10: Trends in Total Number of Pages in .jp Domains

	1998		1999		2000		2001		2002	2002	2004
	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Nov	Feb
Total number of pages (millions)	10.23	17.83	29.53	38.45	42.55	55.73	61.07	65.06	65.58	74.38	85.90
Increase (%)	－	74.29	65.62	30.21	10.66	30.98	9.58	6.53	0.80	13.42	15.49



2.2.2.3 Trends in Total Number of Files in .jp Domains

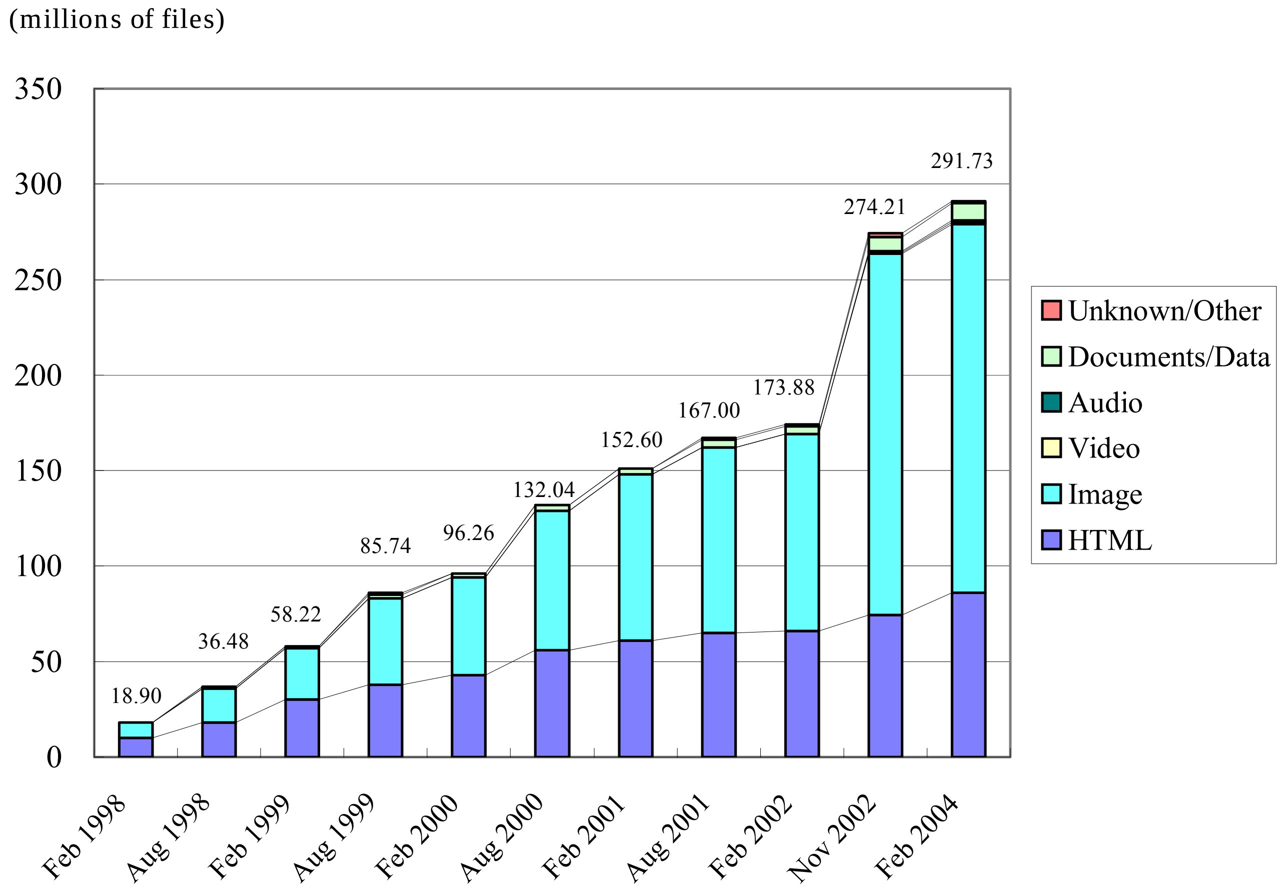
Rapid increase in second half of 2002

The total number of files in .jp domains has also grown consistently. Though the rate of increase slowed around 2001, the second half of 2002 brought a shift to rapid growth. Image files have accounted for a large share of the total number of files since the time of the first survey.

Figure 11:Trends in Total Number of Files in .jp Domains

	1998		1999		2000		2001		2002	2002	2004
	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Nov	Feb
Total number of files (millions)	18.90	36.48	58.22	85.74	96.26	132.04	152.60	167.00	173.88	274.21	291.73
Increase (%)	−	92.91	59.59	47.27	12.27	37.17	15.57	9.44	4.12	57.70	6.39

Figure 12: Trends in Total Number of Files in .jp Domains by File Type



	1998		1999		2000		2001		2002	2002	2004
	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Nov	Feb
HTML	10.23	17.84	29.53	38.46	42.55	55.73	61.07	65.06	65.58	74.38	85.89
Image	8.27	17.75	27.27	44.69	51.03	72.77	87.04	97.07	102.88	189.18	193.39
Video	0.02	0.04	0.05	0.07	0.08	0.10	0.12	0.12	0.14	0.61	0.81
Audio	0.03	0.10	0.11	0.25	0.30	0.34	0.40	0.37	0.36	0.89	1.41
Documents/ Data	0.25	0.61	1.16	1.73	1.98	2.70	3.48	3.88	4.36	7.20	8.99
Unknown/ Other	0.11	0.14	0.11	0.54	0.32	0.40	0.49	0.51	0.55	1.95	1.25
Total	18.90	36.48	58.22	85.74	96.26	132.04	152.60	167.00	173.88	274.21	291.73

(millions of files)

2.2.2.4 Trends in Total Volume of Data in .jp Domains

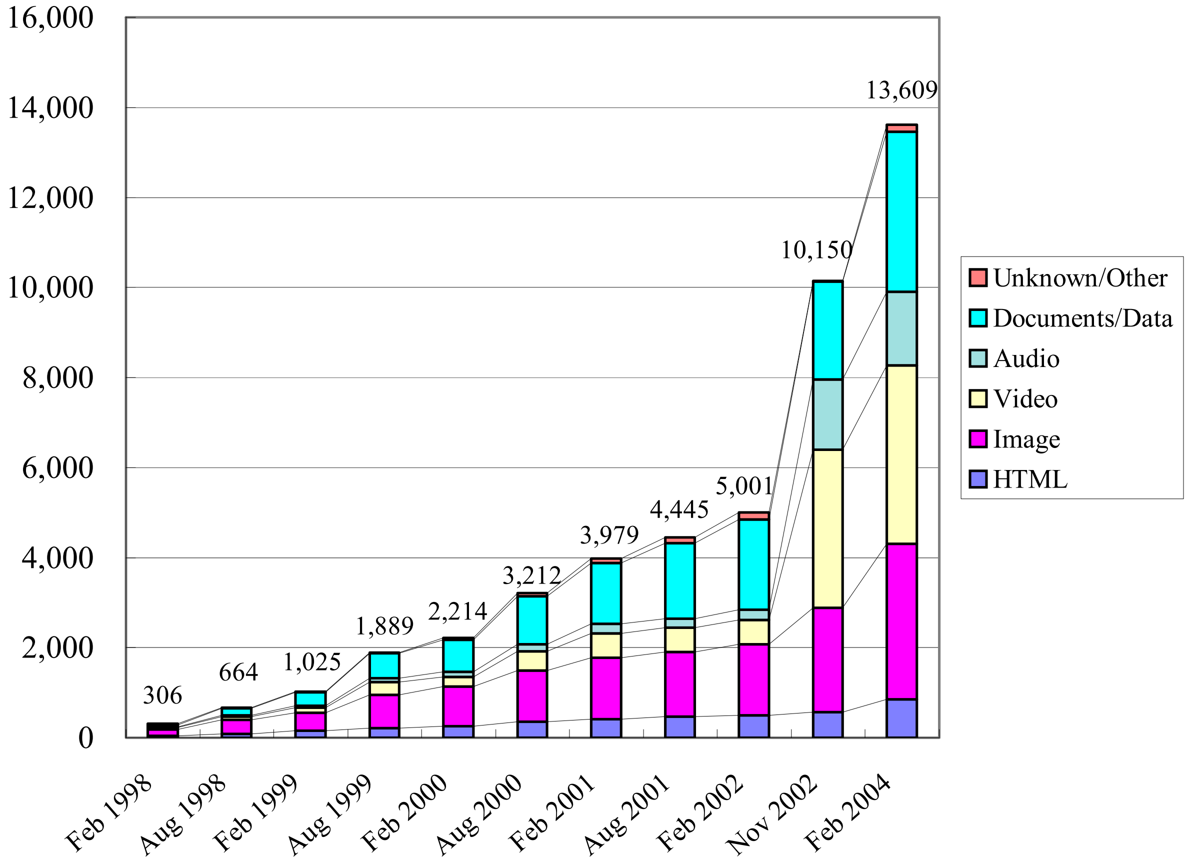
Sustained increase in step with growth of multimedia data

The total volume of data tends to increase at a higher rate than the total numbers of files or servers. The result from the November 2002 survey was double that from the previous survey (February 2002). The latest survey also shows a large increase of 34% compared with the result from the previous survey. There has been substantial growth in the amount of data in video and audio files, and the expansion of these types of multimedia data is causing the overall volume of data to increase.

Figure 13: Trends in Total Volume of Data in .jp Domains

	1998		1999		2000		2001		2002	2002	2004
	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Nov	Feb
Total volume of data (GB)	305	664	1,024	1,889	2,214	3,212	3,980	4,445	5,002	10,150	13,609
Increase (%)	–	117.70	54.22	84.47	17.20	45.08	23.91	11.71	12.51	102.92	34.08

Figure 14: Trends in Total Volume of Data in .jp Domains by File Type



	1998		1999		2000		2001		2002	2002	2004
	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Aug	Feb	Nov	Feb
HTML	46	86	150	211	256	354	411	468	498	564	846
Image	141	306	409	745	885	1,135	1,364	1,440	1,579	2,317	3,461
Video	40	78	113	280	206	434	540	530	543	3,507	3,962
Audio	11	29	39	88	119	155	210	211	216	1,575	1,642
Documents/ Data	53	151	300	546	709	1,057	1,356	1,677	2,009	2,174	3,540
Unknown/ Other	15	14	14	19	39	77	98	119	156	13	158
Total	306	664	1,025	1,889	2,214	3,212	3,979	4,445	5,001	10,150	13,609

(GB)

2.2.2.5 Comparison of Average Data Volume per 10,000 Pages and Average Number of Pages per Server in .jp Domains

Decline in average pages per server, increase in volume of data per page

At the time of the February 1998 survey, the average volume of data per 10,000 pages was 0.3 GB. The results from the February 2004 survey show that the average has increased approximately 5.3 times to 1.58 GB. Internet users browsing the Web today download over five times more information than they did six years ago.

The average number of pages per server rose 1.58 times, from 283.3 pages to 447.4, between the February 1998 and February 2000 surveys. However, the average then began to decline, and the February 2004 survey yielded a figure of 202.1 pages, which is below the average at the time of the February 1998 survey. Since the number of pages on a server is unlikely to change dramatically after the server has been established, this trend appears to indicate that there are many new servers that have small numbers of pages.

However, since the average amount of data linked to each page has increased 5.3 times, it appears that there has been a sustained increase in the amount of information per server despite the decline in the average number of pages.

Figure 15: Average Data Volume per 10,000 Pages and Average Number of Pages per Server

	Total number of pages (millions)	Average data volume per 10,000 pages (GB)	Total number of servers	Average number of pages per server
Feb 1998	10.20	0.3	36,000	283.3
Feb 1999	29.50	0.35	75,000	393.3
Feb 2000	42.50	0.52	95,000	447.4
Feb 2001	61.01	0.65	152,000	401.4
Feb 2002	65.55	0.76	197,000	332.7
Nov 2002	74.38	1.37	308,000	241.5
Feb 2004	85.90	1.58	425,000	202.1

2.2.2.6 Trends in Development of Total Content Amount in .jp Domains by Year

This analysis examines yearly increases in the total numbers of servers, pages, and files and the total volume of data, with estimates from the first survey in February 1998 represented as 100.

Total data volume 44.6 times higher than figure from initial survey

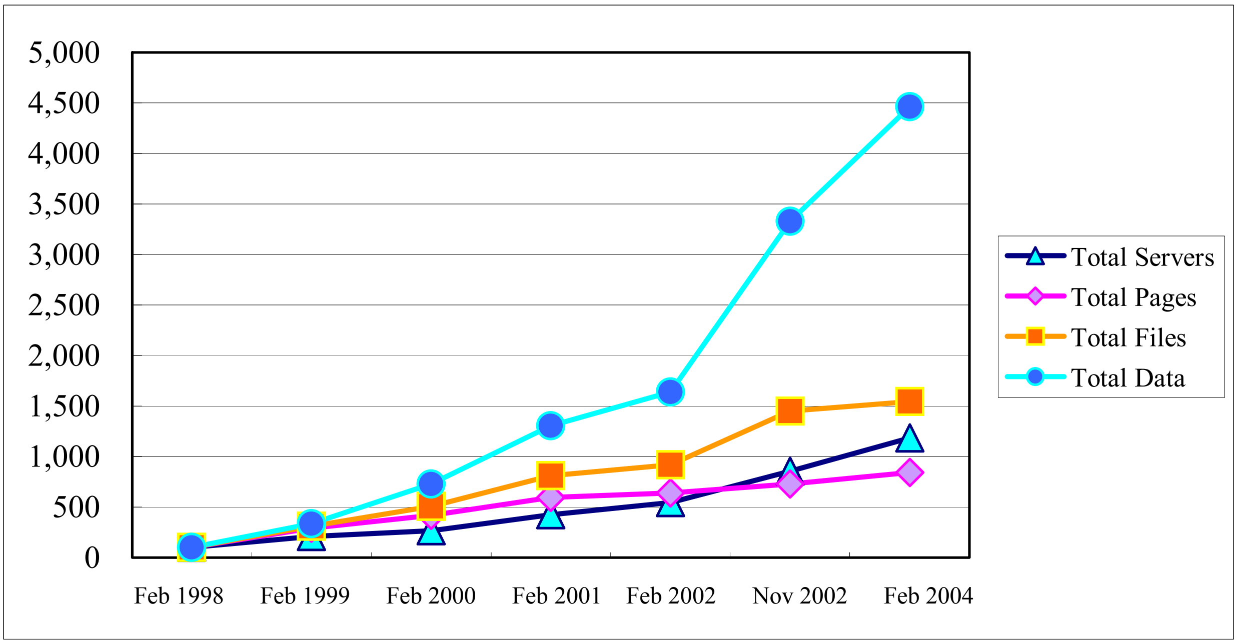
In the past the total number of pages rose faster than the total number of servers. However, the increase in the total number of pages has slowed in recent years, while the total number of servers is now increasing at a faster rate. As a result, the growth pattern has been reversed. While the total number of files is increasing only gradually, there is lively growth in the total volume of data, which in February 2004 was approximately 44.6 times higher than the total as of February 1998.

Figure 16: Trends in Total Content Amount

Estimates	Feb 1998	Feb 1999	Feb 2000	Feb 2001	Feb 2002	Nov 2002	Feb 2004
Total number of servers	36,000	75,000	95,000	152,000	197,000	308,000	425,000
Total number of pages (millions)	10.23	29.53	42.55	61.07	65.58	74.38	85.90
Total number of files (millions)	18.91	58.22	96.26	152.60	173.88	274.21	291.73
Total volume of data (GB)	305	1,024	2,214	3,980	5,002	10,150	13,609

Figure 17: Trends in Total Content Amount (February 1998 = 100)

Indices	Feb 1998	Feb 1999	Feb 2000	Feb 2001	Feb 2002	Nov 2002	Feb 2004
Total number of servers	100	208	264	422	547	856	1,181
Total number of pages	100	289	416	597	641	727	840
Total number of files	100	308	509	807	920	1,451	1,544
Total volume of data	100	336	726	1,305	1,640	3,328	4,462



3. New Phenomena Observed in Recent Survey Results

This survey has been carried out since 1998 as a means of estimating the amount of content on Web servers in Japan. Estimates based on the survey have been published regularly and are widely used as key indicators of the size of the Internet.[16] The results of the scans by search robots in recent years have identified various phenomena that appear to imply the evolution of the Web. The following analysis examines new Web-related phenomena that were not anticipated when the survey was first launched.

3.1 Changes Concerning Survey Design

3.1.1 Changes in Domestic Share of .jp Domains

This survey was initiated to record the increase in the amount of content on Web servers in Japan. Furthermore, the scope of the survey was limited to .jp domains that are allocated to individuals and organizations located in Japan.

One reason for this policy was the fact that domestic sites with generic domains, such as ".com", were unusual and limited in number when the survey began in 1998. Another reason was that there was no way of determining whether or not a generic domain server that had been deduced to contain Japanese language content was located in Japan, since information identifying geographical locations could not be obtained from the server.

However, the state of the Web in Japan has changed steadily over the period in which these surveys have been conducted, from 1998 to 2004. When the total amount of content on Japanese-language sites with generic domains was surveyed on a single-year basis in November 2002 (Figure 18), the estimated total number of Japanese-language servers with generic domains (465,000) was greater than the estimated total number of .jp domain servers (308,000). Even allowing for the fact that generic domains include a large number of Japanese-language sites on overseas servers, it is likely that there are a significant number of generic domain sites in Japan.

To ascertain the extent of Web content in Japan in this environment, it may be necessary to reconsider the policy of surveying only .jp domains.

Figure 18: Total Content Amount on Japanese-Language Sites with Generic Domains (Based on Results of November 2002 Survey)

	Total number of servers	Total number of pages (millions)	Total number of files (millions)	Total volume of data (GB)
Total content amount on generic domains in Japanese	465,000	25.02	96.42	5,444
Total content amount on .jp domains (for reference)	308,000	74.38	274.21	10,150

[16] One example is the e-Japan Priority Policy Program Benchmarks (December 18, 2003) http://www.kantei.go.jp/jp/singi/it2/dai22/22siryou6.pdf

3.1.2 Increase in Multimedia Data That Cannot Be Acquired by Robots

When the survey was first launched, there were numerous Web pages made up of simple HTML files, image files, video files, and the other files. The search robots are programmed to access such Web pages and to collect the data of the files, while, the data of the following types of Web pages and files cannot be collected.

- Pages generated automatically by scripts, such as CGI programs
- Video files with embedded “Play” and “Stop” buttons
- Streaming video and audio files

In recent years these types of pages and files have proliferated and become important Web tools, to the extent that they now appear to be the mainstream formats for video and audio. While it is apparent from this survey that there has been a rapid increase in video and audio content, the survey’s inability to monitor a significant portion of video and audio content has become a problem from the viewpoint of estimating information volumes.

3.1.3 Quantitative Web Growth and the Need for Enhanced Survey Infrastructure

The Web has continued to expand over the years, and Loki has been enhanced to allow it to access large numbers of Web pages in a short space of time. However, there are inherent limits to this approach, in terms of both technology and costs.

Various improvements have been made to the Loki program. For example, the program has been run in parallel, allowing multiple machines to be operated simultaneously with the load shared appropriately among them. In practice, however, there is a limit to this type of enhancement, since a variety of problems start to appear when the number of robots operating in parallel exceeds a certain level.

3.1.4 Hub Congestion Caused by Network Traffic Growth

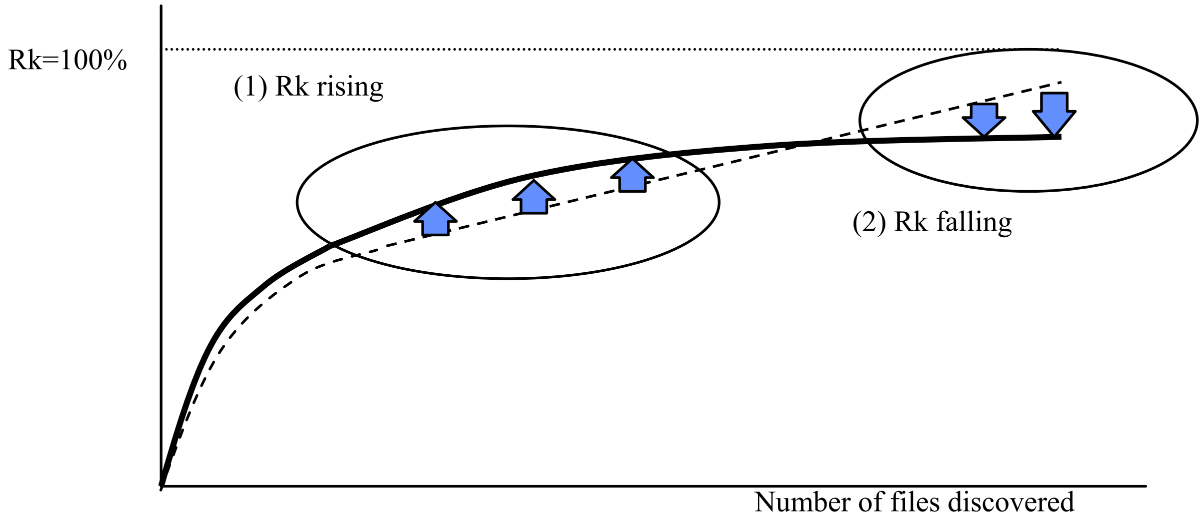
Loki scans are also affected by the network environment. In recent surveys, scanning times have increased because of hub congestion resulting from traffic growth. There is a need for improvements not only to the machines, but also to the installation environment, especially line capacity (including Internet access from a point as close as possible to the backbone).

3.2 Changes Concerning Structural Model of the Web

3.2.1 Phenomena Identified from Recent Survey Results

As described in Section 1.1 (Survey Structure and Principles), the total number of files is estimated by plotting the discovery rate for known URLs (Rk value) on a graph. Recent survey results show two phenomena that were not previously seen.

Figure 19: Two Phenomena That Appeared on Loki's Known URL Discovery Ratio (Rk) Graph



(1) Higher Rk Line with Increased Curvature

The Rk line in the graph initially rises steeply. After it reaches a certain level, however, the rate of climb slows, and the line turns into a gradual curve that leads into a straight line with a very shallow gradient. The initial steep rise in the Rk graph is caused by the large number of links to the files obtained. The curving of the line reflects a sustained decline in the average number of links to the files obtained. The shift to a straight line with a shallow gradient indicates that the number of links to the pages has fallen as far as it can, and that links are now uniformly sparse.

The survey was initially based on this model of the Web, and the early surveys produced Rk graphs that confirmed the validity of the model. However, while Rk graphs based on recent survey results still show the rapid initial rise in the Rk value, instead of flattening out into a straight line, the curve is lengthened. In the latest survey, it was difficult to obtain a clearly defined straight line.

(2) Flat Rk Values in Final Stage of Survey

Under the estimation model used with this survey, it is necessary to obtain an Rk graph line that ends in a straight line with a shallow gradient. The total number of files is estimated by extending this gradient until the known link ratio reaches 100%. Results from past surveys have supported this approach. As noted in (1), however, it has become increasingly difficult to identify a straight line based on recent results. Moreover, there is now a tendency for the line to level out. This means that the increase in the Rk value has halted, even though the actual number of files obtained is rising.

3.2.2 Identifying Possible Causes from Structural Changes in the Web

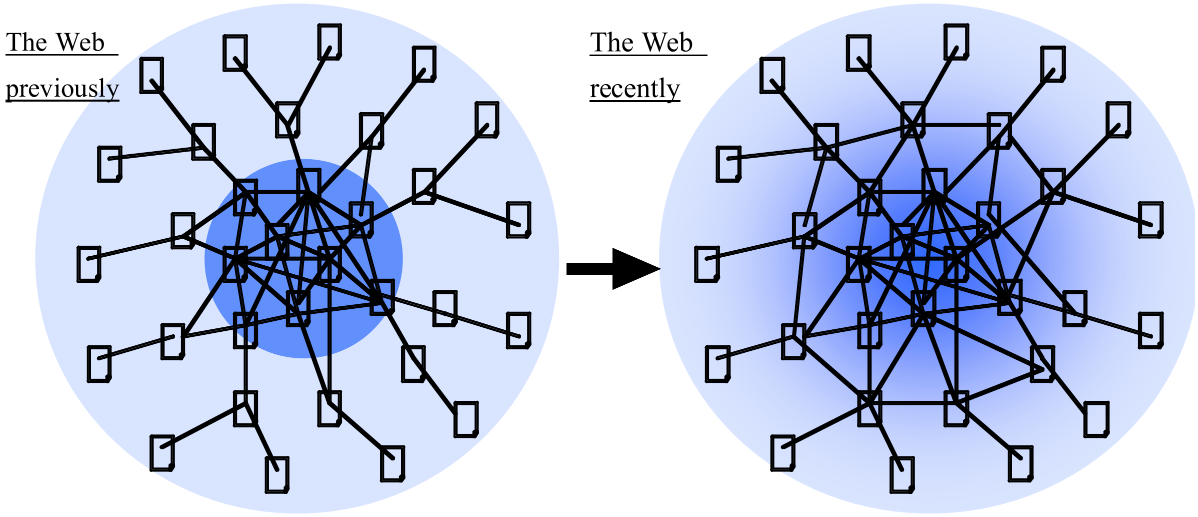
The purpose of the following analysis is to deduce the causes of the phenomena described in Section 3.2.1 in relation to recent survey results.

Increase in Web pages with medium-scale linkage

The Web model that was constructed when the surveys first began assumed a structure that was polarized between Web pages containing large numbers of links and Web pages that did not. Recent results have been characterized by the phenomenon of Rk values that fail to shift to a straight line after the initial sharp rise and curve phase, and which remain in the curve phase for a prolonged period. This suggests that this polarization no longer exists, and that there are now numerous Web pages with varying numbers of intermediate links to them, and that the demarcation line for uniform pages with sparse links is becoming blurred.

This change can be seen as an indication that the structure of the Web is maturing. As Web use becomes more common, its characteristic function as a system for linking information has become a feature of Web pages in general.

Figure 20: Increase in Web Pages with Medium-Scale Linkage



Increase in dead links

The likely cause of the of flattening static Rk values in the final stage of the survey is an increase in the number of dead links (links to pages that do not exist, or links that the robots cannot follow) on the periphery of the Web. Past estimates were based on the assumption that dead links made up only a small percentage of the total Web and would therefore have little effect on the Rk value. However, the incidence of dead links increases with proximity to the periphery of the Web. The existence of a high percentage of dead links will prevent the Rk value from approaching 100%.

Reasons for the existence of dead links include an increase in the number of old pages that have not been maintained for many years, the proliferation of pages with access limitations, and the growing number of automatically generated pages. Another potential factor is the increased number of pages that supply content according the type of browser used, such as sites designed to be accessed from mobile telephones.[17]

[17] There are servers on the Web that are geared for access from mobile telephones. These are known in Japan as *katte saito* (voluntary sites). Because Loki does not have a browser configured for mobile telephone use, it is refused access to these voluntary sites, which are consequently counted as dead links.