WWWコンテンツ統計調査報告書

～我が国の Web 上のコンテンツ情報量から見たインターネットの発展～

平成１６年　７月

総務省　情報通信政策研究所

主任研究官　佐伯　千種
研究官　島田　博也
研究官　田畑　伸哉

目　次

図表目次

はじめに－Ｗｅｂの役割とＷｅｂコンテンツ情報量調査の意義

【Ｗｅｂの役割】

1990 年代以降、インターネット利用は世界的に急拡大が続いている。その最大の要因は World Wide Web[1]（以下「Web」と言う。）利用者とその用途の急拡大にある。インターネットの恩恵を「情報流通の時間・距離・費用の克服」と捉えるなら、その用途の拡大に Web ほど貢献したツールはないと言える。

Web は文書だけでなく、画像や音声、動画等を含めたマルチメディア情報を提供することができ、ブロードバンドインターネット需要拡大の礎を築いた。Web は膨大なコンテンツを手軽に利用できる形で蓄積するという機能を有するだけでなく、蓄積された情報がオンラインで利用されることによって、ビジネス、生活、社会活動のあり方そのものの変革も促している点で、その経済・社会に与える影響は極めて大きい。

最近のインターネットにおいては、P2P[2]やマルチキャスト[3]などの新たなツールが生まれてきているが、利用者・用途の広範さと言う観点で見れば、Web は将来も相当な期間インターネットの中核を占めるツールであり続けるものと思われる。

【Ｗｅｂコンテンツ情報量調査の意義】

こうした Web の及ぼす経済・社会的影響に鑑み、勃興期から Web の発展の推移を概観的に記録しておくことは、インターネットの社会的役割の変化、Web が原因となる社会活動の変化を裏付ける上で大変重要な作業である。本調査は、1998 年から毎年定期的に国内のＪＰドメインの Web コンテンツ情報量を推計するもので、今年で７年目を迎える。長期にわたり、定期的に、統計的手法を用いて Web のコンテンツ情報量を測定したという点で、世界的にも例を見ない唯一の調査である。

本稿では、Webの勃興期からブロードバンド化が進んだ今日までのWebコンテンツの発展について、2003年度（第11回）調査の最新調査結果とともにその推移を概観し、最近の調査結果から見られたWeb上の新たな現象について考察する。

---

[1] インターネットやイントラネットで標準的に用いられるハイパーテキスト・システム。ハイパーテキストとは、文章内にあるテキスト文字列が、さらに別のテキストやファイルにリンクしている文書システムのこと。欧州核物理学研究所(CERN)の Tim Berners-Lee 氏が所内の論文閲覧システムとして 1989 年に考案したものを基礎としている。広く一般に公開されたのは 1991 年のこと。インターネット標準のハイパーテキスト・システムとして 1990 年代中頃から爆発的に普及し、現在では世界規模での巨大な WWW 網が築かれている。インターネットで最も多く利用されるアプリケーションである。

[2] サーバを介さずにコンピューター同士を直接接続してデータをやり取りする通信形態。

[3] 単一のパケットを同時に特定の複数のコンピューターに送信する技術。

図表 １　インターネット用途の拡大

	１９７０年代	１９８０年代	１９９０年代
主たる役割	米政府のツール コンピュータ間の データ交換手段	学者・研究者などの コミュニケーション ツール	経済活動・社会活動・生活 の重要なツール 社会インフラ化
中核のツール	TCP/IP・FTP	電子メール	Web
新たな ネット利用者	政　　府	大学・研究関係者など	一般団体・個人
新たな用途	データ通信	メール ニュース	マルチメディア情報検索 （仕事・学業・趣味・商品） 商取引・掲示板・報道・放送 教育・公的手続きなど

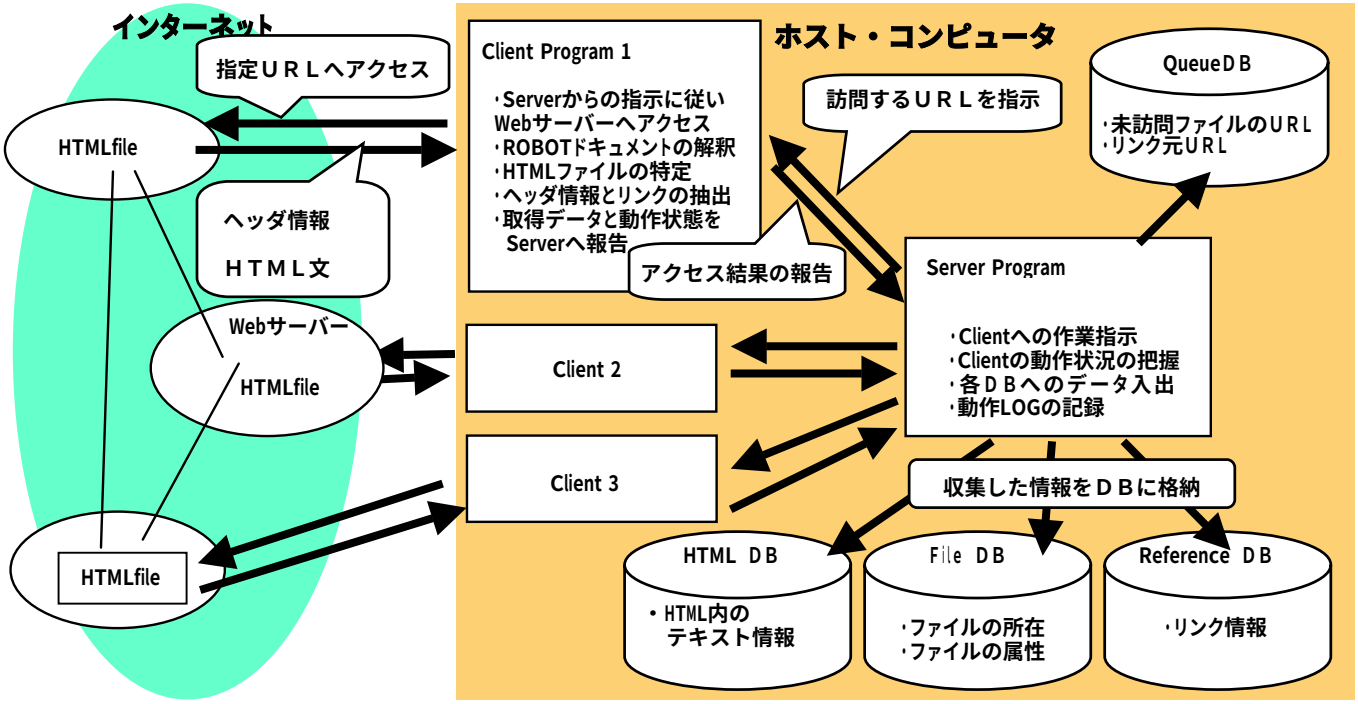
１．調査の概要

情報通信政策研究所では、旧郵政研究所調査研究部時代の 1998 年 2 月から、アライド・ブレインズ株式会社と共同開発したサーチロボットの走査結果をもとに推計する独自の手法により、国内 Web コンテンツ量の統計調査を実施している[4]。

Web の膨大な規模や成長速度を考えると、どれほど高性能なサーチロボットを用意しても、公開されている Web ページすべてにアクセスすることは現実には不可能である。そこで、Web 上に存在する「ある程度の」ページをサーチロボットで調査し、そこで得られる URL(Uniform Resource Location)[5]等のデータをもとに統計的推測を加えて、Web のサーバ数、ファイル[6]数、ページ[7]数、及びデータ量のそれぞれの総数[8]を求めている。

1-(1)　調査の仕組み・原理

図表 ２　サーチロボット「Ｌｏｋｉ」の仕組み



[4]「インターネットコンテンツ統計に関する調査研究　郵政研究所月報 2002 年 9 月」
（http://www.soumu.go.jp/iicp/seika/data/research/monthly/2002/168-h14_09/168-asearch2.pdf ）
「メディアとしてのＷｅｂの成長を測る～サーチロボットを使ったＷｅｂコンテンツ統計調査の試み
マス・コミュニケーション研究 No.62 2003 年 3 月」
（http://www.a-brain.com/HP/rep/rep15/index.html）を参照。

[5] インターネット上に存在する HTML(HyperText Markup Language)、文書/データ、画像などのファイルの場所を指し示す記述方式。情報の種類やサーバ名、ポート番号、フォルダ名、ファイル名などで構成されている。

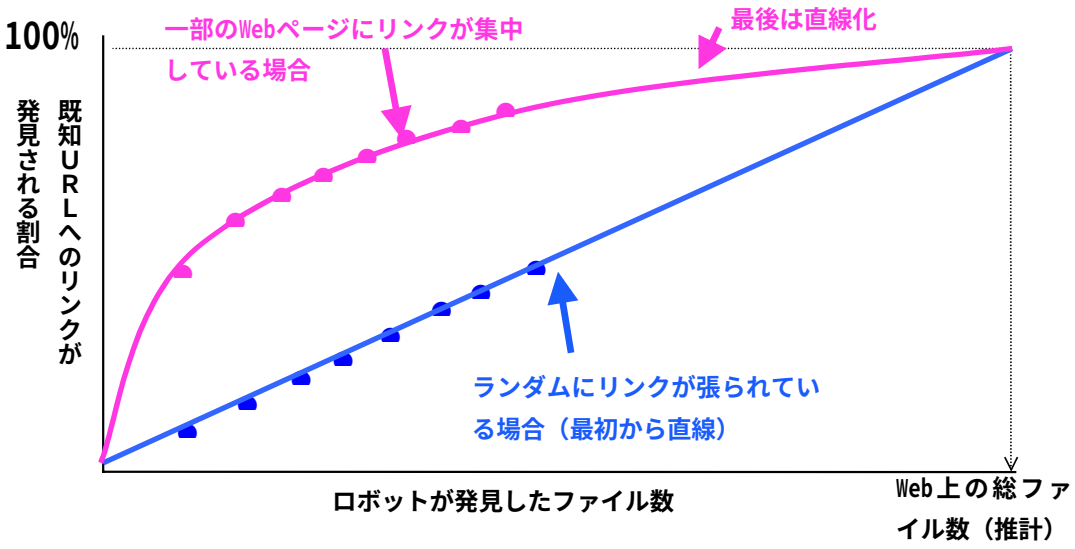
[6] 本調査では、独立した URL をもつ Web サイト上のコンテンツである HTML、文書/データ、画像、動画、音声などのファイルをいう。

[7] 本調査では、HTML ファイルと同義とする。HTML は文書の論理構造や見栄えなどを記述するために使用され、文書の中に画像や音声、動画、他の文書へのリンク情報などを埋め込むこともできることから、HTML ファイルを「ページ」と呼び、他のファイルと区別している。

[8] 総数の他、ファイル種類別のデータ量、ファイル数を推計している。

本調査のために開発したサーチロボット「Ｌｏｋｉ」は、HTMLのハイパーリンクを自動的に辿ってクローリング（走査）することにより、インターネット上の広範なWebページのURL情報を自動収集するプログラムである。情報収集のスピードアップのため複数の情報収集プログラム（クライアント）を同時に動かして並列でクローリングを行う。サーバープログラムは、これらのクライアントプログラムの動作管理や収集した情報の整理を行う。

図表　３　既知ＵＲＬ発見率の変化



サーチロボットは新しい Web ページにアクセスすると、その HTML 文に記述されているリンク情報を取得するが、それらの中にはサーチロボットにとって未知のＵＲＬへのリンクもあれば過去既に発見済みのＵＲＬへのリンクもある。サーチロボットの走査が進むにつれて、サーチロボットが既に知っているＵＲＬの数（以下「既知ＵＲＬ」という。）は増えていくので、新たに発見したＵＲＬの比率はどんどん低下し、一方で既知ＵＲＬが重複してみつかる比率はどんどん上昇する。サーチロボットがアクセス可能なすべての Web ページのＵＲＬを取得し終わったときには、その中のどの Web ページにアクセスしても、そこに記述されているリンク先はすべて既知ＵＲＬとなるはずである。したがって、サーチロボットがある程度の数のＵＲＬを取得した段階で、新たに取得したリンク情報の中の既知ＵＲＬの比率を調べれば、それまでに取得したＵＲＬが Web 上のファイル全体の何パーセントにあたるか推定できることになり、すべての Web ページにアクセスしなくても Web の総ファイル数を推計できることになる。

この手法で Web の総ファイル数等を推計する場合、Web ページ間のリンクが Web 全体にランダムに、かつ偏りなく張られていれば、サーチロボットが発見したファイル数と既知ＵＲＬが発見される割合は比例し、サーチロボットの調査が進むにつれて、既知ＵＲＬ発見率は図表６のように直線的に上昇するはずである。しかし、実際の Web ではリンクの分布に大きな偏りがあり、一部の「有名」Web ページに多数のリンクが集中している。このような有名 Web ページは早期にサーチロボットに発見される確率が高く、またいったん発見されると既知ＵＲＬ発見率の値を一気に高めることになる。つまり、実際の調査では、既知ＵＲＬ発見率の値は直線的に上昇するの

ではなく、初期の段階に急激に上昇し、その後漸増し、最後にリンク構造が均一化すると直線的に漸増すると考えられる。このような前提に立ち、グラフが直線になった時点で、線形近似法[9]という手法を用いてWebの総ファイル数の推計を行っている。

１-(2)　調査対象

本調査で対象としているWebの範囲は、一般に公開されているＪＰドメイン[10]のWebページである。最近では、国内企業等が公開している「.com」などの汎用ドメイン[11]をもつものが増えているが、本調査ではＪＰドメインのみを対象としている[12]。

なお、サーチロボットは必ずしもすべてのWebページにアクセスできるわけではなく、次のようなWebページやファイルは収集ができないため、調査対象から外れている。

- 外部からのリンクを持たないもの
- 会員限定等のアクセス制限のあるWebページ
- 「robots.txt」でサーチロボットのアクセスを禁じているWebページ
- CGI[13]等のスクリプトで自動生成されるもの
- 検索エンジンで自動生成されるもの
- 再生、停止ボタンが埋め込まれている動画ファイル
- ストリーミングの動画、音声ファイル[14]

---

[9] 情報通信政策研究所（旧郵政研究所）とアライド・ブレインズ株式会社の共同で特許出願中。
[10] 国ごとに割り当てられる国別ドメイン名（ccTLD）で、日本のccTLDである「.jp」のこと。セカンドレベルが組織種別をあらわす属性型ＪＰドメイン名やセカンドレベルに取得者の希望する名前を登録する汎用ＪＰドメイン名などがある。
[11] 国の区別なく世界中で自由に取得可能な分野別ドメイン名（gTLD）。「.com」、「.net」、「.org」等
[12] ＪＰドメインを対象としている理由については、後述３-(1)-1)-1を参照。なお、2002年11月調査は単年で汎用ドメインも調査対象に含めている。
[13] Common Gateway Interfaceの略。Webサーバが、Webブラウザからの要求に応じて、プログラムを起動するための仕組み。従来、Webサーバは蓄積してある文書をただ送出するだけであったが、CGIを使うことによって、プログラムの処理結果に基づいて動的に文書を生成し、送出することができるようにするもの。
[14] Windows Media PlayerやReal Playerのファイルがこれに該当する

２．調査結果について

２-(1)　これまでの調査実施状況

サーチロボットを使った調査は 1998 年２月の初回調査から 2001 年８月の第８回調査までは半年ごとに実施していたが（２月と８月)、2002 年以降は年１回のペースで実施している。

図表　４　クローリング(走査)実施実績

	サーチロボット走査期間
第1回	1998年　2月10日～　2月26日
第2回	1998年　8月　3日～　9月　7日
第3回	1999年　2月16日～　3月11日
第4回	1999年　8月　4日～　9月26日
第5回	2000年　1月17日～　3月　7日
第6回	2000年　8月30日～　9月27日
第7回	2001年　2月10日～　3月19日
第8回	2001年　7月20日～10月30日
第9回	2002年　2月　3日～　4月23日
第10回[15]	2002年　10月22日～11月　6日
第11回	2004年　1月　5日～　2月26日

[15] 第 10 回調査ではサーチロボットに株式会社エヌ・ティ・ティ・エックス（現エヌ・ティ・ティレゾナント株式会社）のサーチエンジンを使用。

２-(2)　調査結果

２-(2)-1)　2004 年２月（第 11 回）調査結果

２-(2)-1)-1．ＪＰドメイン総コンテンツ量

【総サーバ数・総データ量の伸び大きく、総ファイル数の伸び低調】

総サーバ数、総ファイル数、総ページ数、総データ量とも前回（2002年11月）調査に比べ増加しており、今回の調査では、総データ量は13,609ギガバイト（ＧＢ）[16]と推計された。

伸び率で見ると、前回調査との比較では、総ファイル数の伸び（６．３９％）に比べ、総サーバ数の伸び（３７．９９％）と総データ量の伸び（３４．０８％）が大きい。ファイル数の少ないファイル構造の簡単なサーバが多数出現していると考えられる。また、総ファイル数の伸びに比べて総データ量の伸びが大きいことから１ファイルあたりのデータ量は増えているものと考えられる。

図表　５　2004 年 2 月（第 11 回）調査結果（総サーバ数、総ファイル数、総ページ数、総データ量）

	２００２年　１１月（参考）	２００４年　２月
総サーバ数（台）	308,000	425,000
前回比伸び率（%）		37.99
総ページ数（万Ｐ）	7,438	8,590
前回比伸び率（%）		15.49
総ファイル数（万Ｆ）	27,421	29,173
前回比伸び率（%）		6.39
総データ量（ＧＢ）	10,150	13,609
前回比伸び率（%）		34.08

[16] バイトは情報量を表わす単位で 1 ビット（bit）を 8 つ集めた情報量。1 ビットで 0 か 1 かの 2 種類の情報を表わす。1,024 バイトを 1KB（キロバイト）、1,024KB を 1MB（メガバイト）、1,024MB を 1GB（ギガバイト）、1,024GB を 1TB（テラバイト）と呼ぶ。

２-(2)-1)-2．ＪＰドメイン総ファイル数のファイル種類別シェア

【ファイル数では画像が多く、次いでＨＴＭＬ】

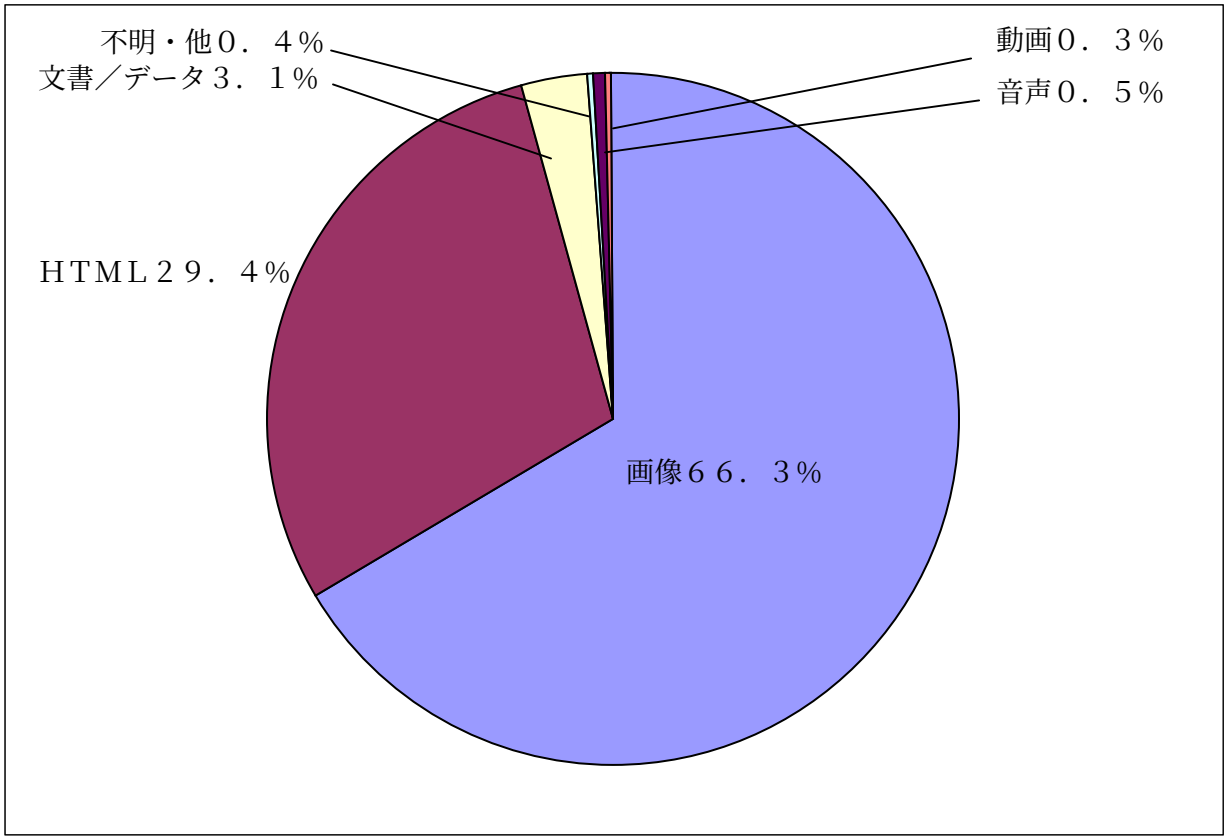
総ファイル数のファイル種類[17]ごとの構成比で見れば、多くは画像であり、全体の６６．３％を占める。次いでＨＴＭＬが２９．４％となっている。

動画、音声はそれぞれ０．３％、０．５％を占めるにすぎない。

図表 ６　総ファイル数のファイル種類別シェア（2004 年 2 月調査結果）

（単位：％）

画像	HTML	文書／データ	音声	動画	不明／他
66.3	29.4	3.1	0.5	0.3	0.4



[17] ファイルの種類は URL 最後の「拡張子」に基づいて分類している。各種類別の代表的な拡張子は次のようなものである。

HTML：「.htm」「.html」
画 像：「.jpg」「.gif」「.bmp」「.pict」「.tif」「.eps」「.png」
動 画：「.mpg」「.avi」「.mov」
音 声：「.au」「.ra」「.midi」「.mp3」「.rmi」「.wav」
文書/データ：「.pdf」「.txt」「.doc」「.jbw」「.lzh」「.tar」「.xls」「.exe」「.java」

２-(2)-1)-3．ＪＰドメイン総データ量のファイル種類別シェア

【データ量では動画ファイルが３割を占める】

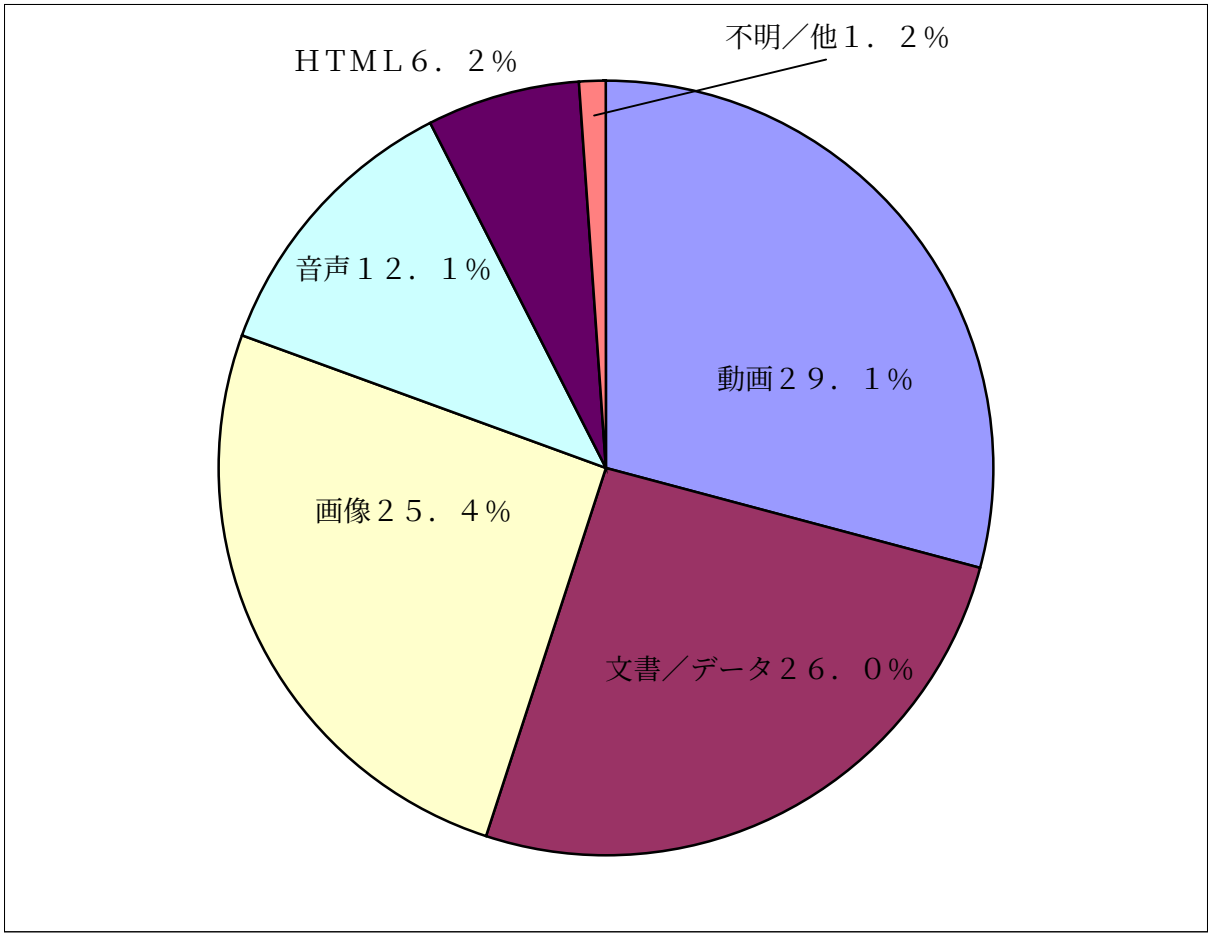
総データ量のファイル種類別構成比で見ると、動画・音声等のマルチメディアファイルや文書・データのデータ量が大きな比率を占め、動画が２９．１％と最も多く、次いで文書/データが２６．０％、画像が２５．４％、音声が１２．１％となっている。

動画・音声はは総データ量のファイル種類別構成比では約４割を占めるものの、総ファイル数のファイル種類別構成比ではわずか０．８％に過ぎない（図表６参照）。動画1ファイル当たりのデータ量が他のファイルに比べ圧倒的に大きいことが伺える。今後、動画をはじめとするマルチメディアファイルの割合が総ファイル数においても増加すれば、Web上の総データ量はますます増加していくものと考えられる。

図表 7　総データ量のファイル種類別シェア（2004 年 2 月調査結果）

（単位：%）

動画	文書／データ	画像	音声	HTML	不明／他
29.1	26.0	25.4	12.1	6.2	1.2



２-(2)-2)　過去の調査結果との比較

２-(2)-2)-1．ＪＰドメイン総サーバ数の推移

【総サーバ数は一貫して順調な伸び】

ＪＰドメインの総サーバ数は1998年初回調査から一貫して増加しており、特に最近２年間は連続して増加ペースが速まっている。

図表　８　ＪＰドメイン総サーバ数の推移

	1998年		1999年		2000年		2001年		2002年	2002年	2004年
	2月	8月	2月	8月	2月	8月	2月	8月	2月	11月	2月
総サーバ数(台)	36,000	54,000	75,000	85,000	95,000	120,000	152,000	177,000	197,000	308,000	425,000
前回比伸び率(%)	―	50.00	38.89	13.33	11.76	26.32	26.67	16.45	11.30	56.35	37.99

【構成比ではacドメインは縮小、coドメインが拡大】

ＪＰドメインサーバのセカンドレベルドメイン別の構成比を見ると、調査開始当初において最も構成比が大きかったacドメインは、その後構成比では縮小を続け、代わってcoドメインの数が増加し、この４年間構成比では全体の過半数を占めるようになっている。調査初期の頃は、学術機関における情報交換のツールであったWebが、徐々に企業活動のツールにシフトしたことが伺える。

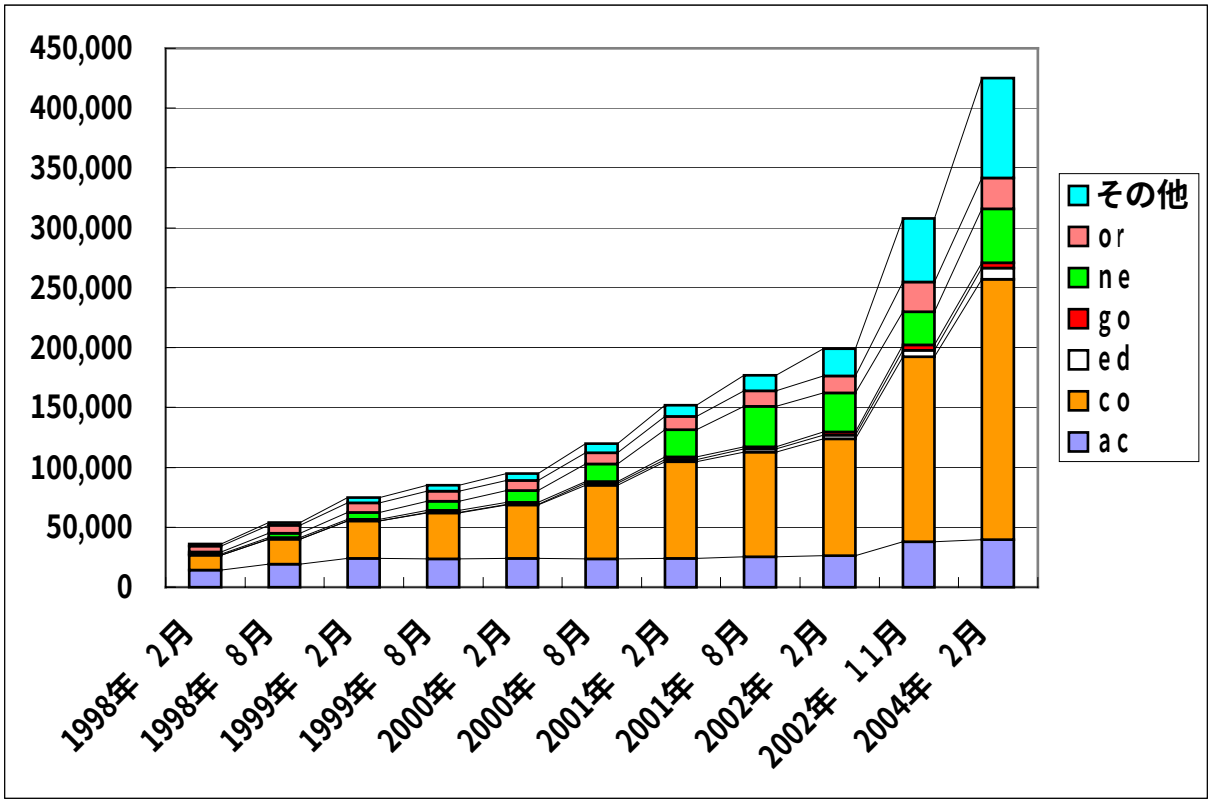
近年の顕著な動きとしては、「その他」ドメインの急速な増加があげられる。「その他」ドメインは2004年2月時点の構成比で約１９．６％を占めるところまで来ているが、これは2001年以降の汎用ＪＰドメイン[18]の導入・普及の影響が大きいものと思われる。

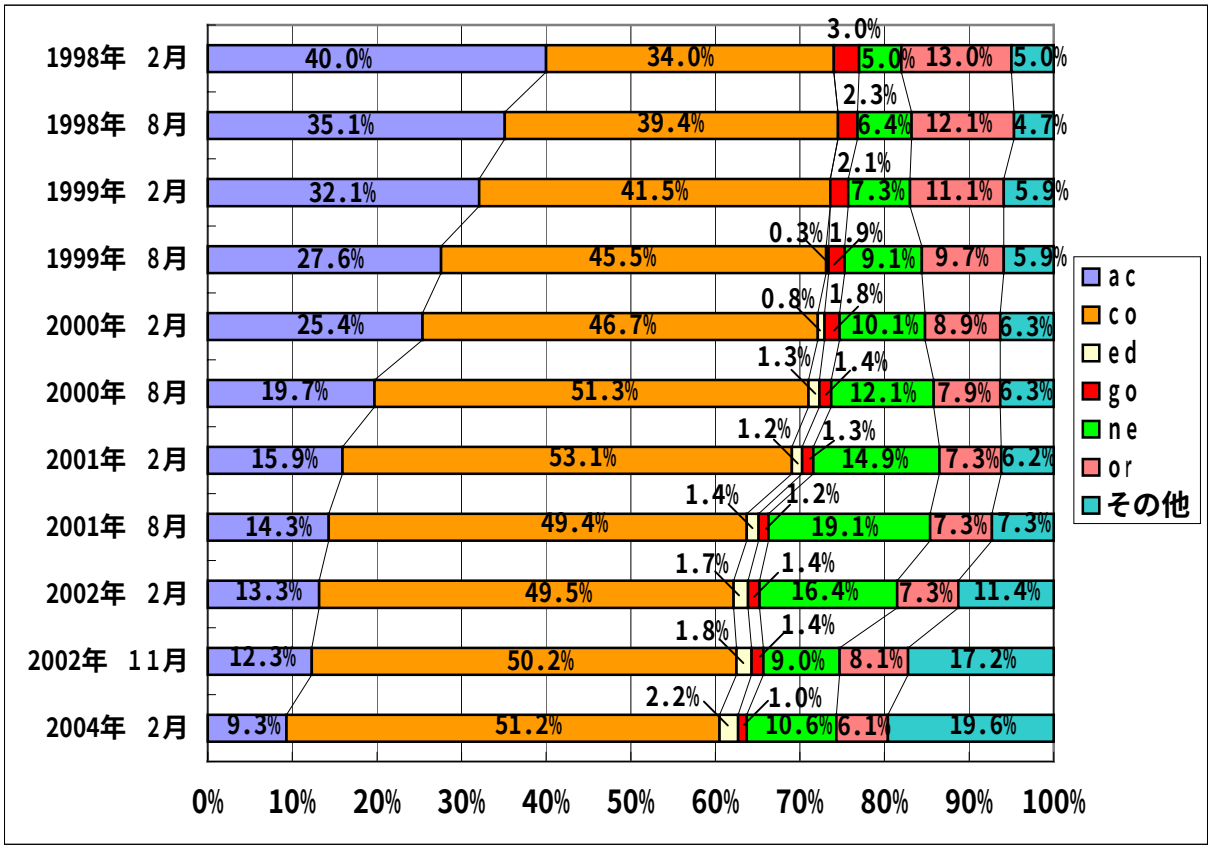
図表　９　主要なセカンドレベルドメイン

ac･･･大学系教育機関等	go･･･政府機関
co･･･一般企業等	ne･･･ネットワークサービス等
or･･･会社以外の団体等	ed･･･小、中、高等学校

[18] 属性型ＪＰドメイン名は１つの組織につき１つしか使用することができないなどの制約があるが、汎用ＪＰドメインは同一のものがない限りいくつでも自由に登録できる。

図表　１０　ＪＰドメイン総サーバ数のセカンドレベルドメイン別シェアの推移





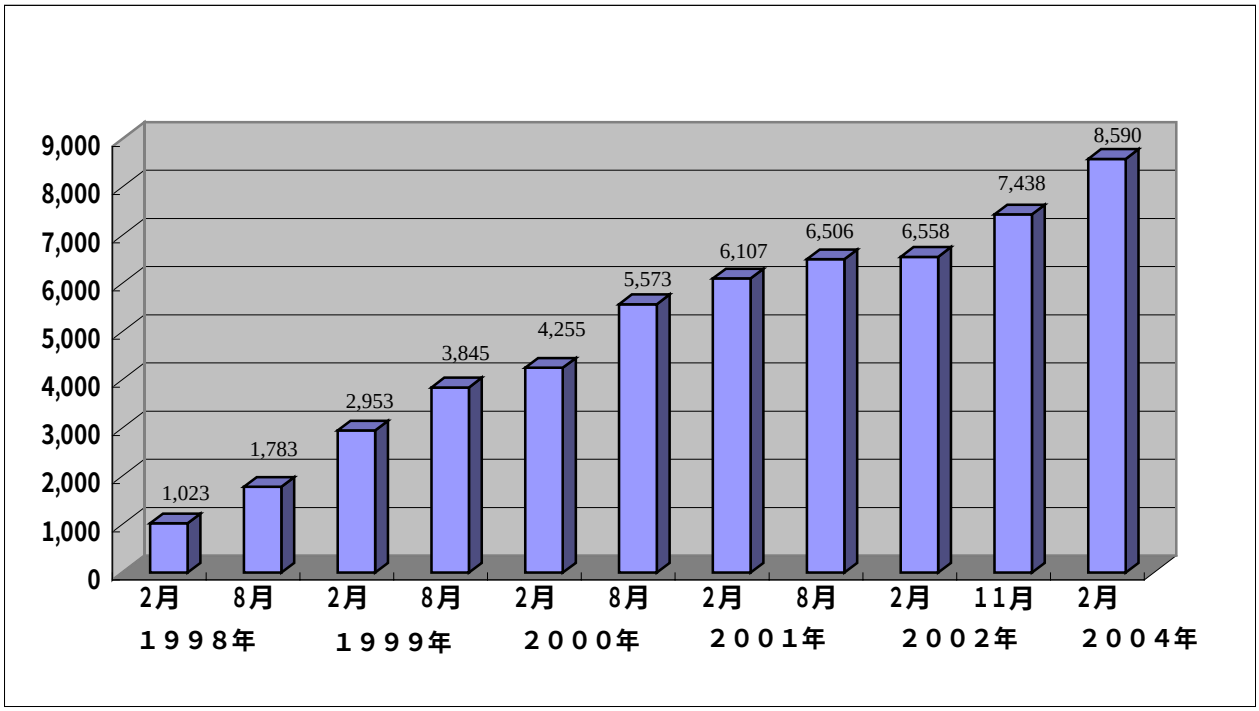
２-(2)-2)-2．ＪＰドメイン総ページ数の推移

【総ページ数は一時伸び率が鈍化したが近年復調】

ＪＰドメインの総ページ数は、2001年から一時伸び率は低下していたが、この２年間にまた増加ペースを早めている。

図表 １１　ＪＰドメイン総ページ数推移

	1998年		1999年		2000年		2001年		2002年	2002年	2004年
	2月	8月	2月	8月	2月	8月	2月	8月	2月	11月	2月
総ページ数(万P)	1,023	1,783	2,953	3,845	4,255	5,573	6,107	6,506	6,558	7,438	8,590
前回伸び率(%)	―	74.29	65.62	30.21	10.66	30.98	9.58	6.53	0.80	13.42	15.49



２-(2)-2)-3　ＪＰドメイン総ファイル数の推移

【総ファイル数は2002年後半に急速な伸び】

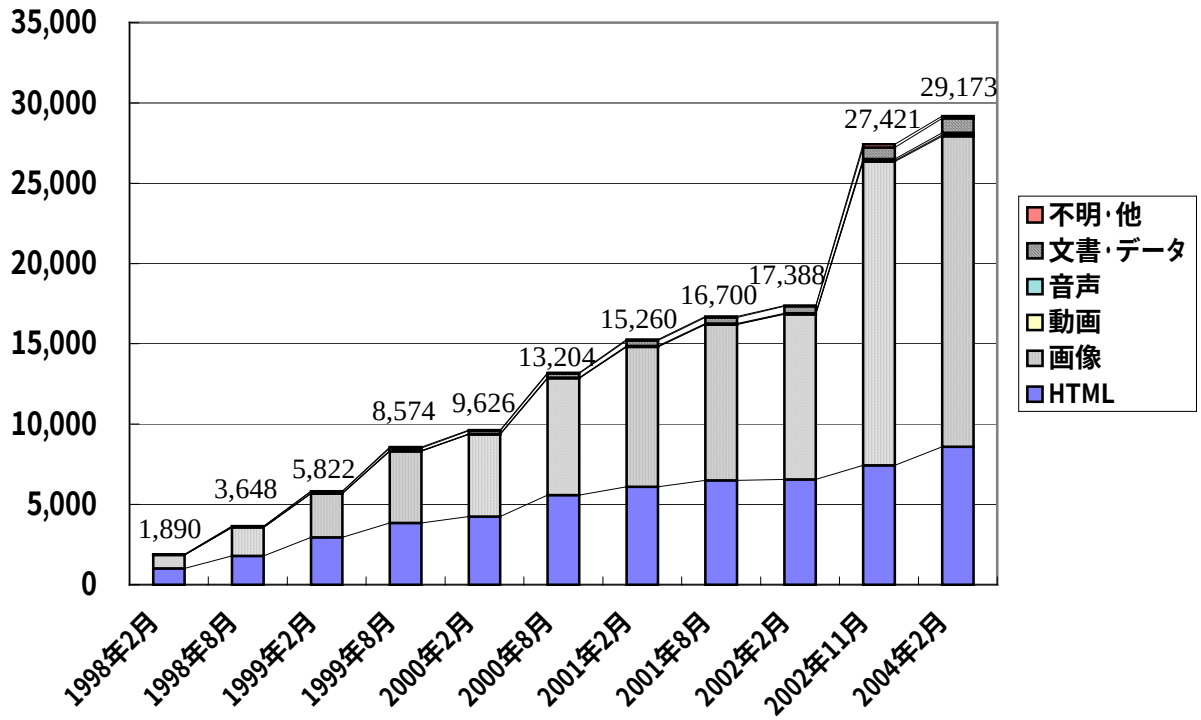
ＪＰドメインの総ファイル数も一貫して増加を続けているが、2001年頃にはその増加率は低下傾向にあった。しかし、2002年後半からは一転して大きな伸びが見られた。

ファイル種別に見ると、画像ファイルが初回調査から高い割合を占めている。

図表 １２　ＪＰドメイン総ファイル数の推移

	1998年		1999年		2000年		2001年		2002年	2002年	2004年
	2月	8月	2月	8月	2月	8月	2月	8月	2月	11月	2月
ファイル数（万P）	1,890	3,648	5,822	8,574	9,626	13,204	15,260	16,700	17,388	27,421	29,173
前回比伸び率（%）	－	92.91	59.59	47.27	12.27	37.17	15.57	9.44	4.12	57.70	6.39

図表 １３　ＪＰドメイン総ファイル数の推移（ファイル種類別）



（単位：万Ｆ）

	1998年		1999年		2000年		2001年		2002年	2002年	2004年
	2月	8月	2月	8月	2月	8月	2月	8月	2月	11月	2月
HTML	1,023	1,784	2,953	3,846	4,255	5,573	6,107	6,506	6,558	7,438	8,589
画像	827	1,775	2,727	4,469	5,103	7,277	8,704	9,707	10,288	18,918	19,339
動画	2	4	5	7	8	10	12	12	14	61	81
音声	3	10	11	25	30	34	40	37	36	89	141
文書･データ	25	61	116	173	198	270	348	388	436	720	899
不明･他	11	14	11	54	32	40	49	51	55	195	125
合計	1,890	3,648	5,822	8,574	9,626	13,204	15,260	16,700	17,388	27,421	29,173

２-(2)-2)-4　ＪＰドメイン総データ量の推移

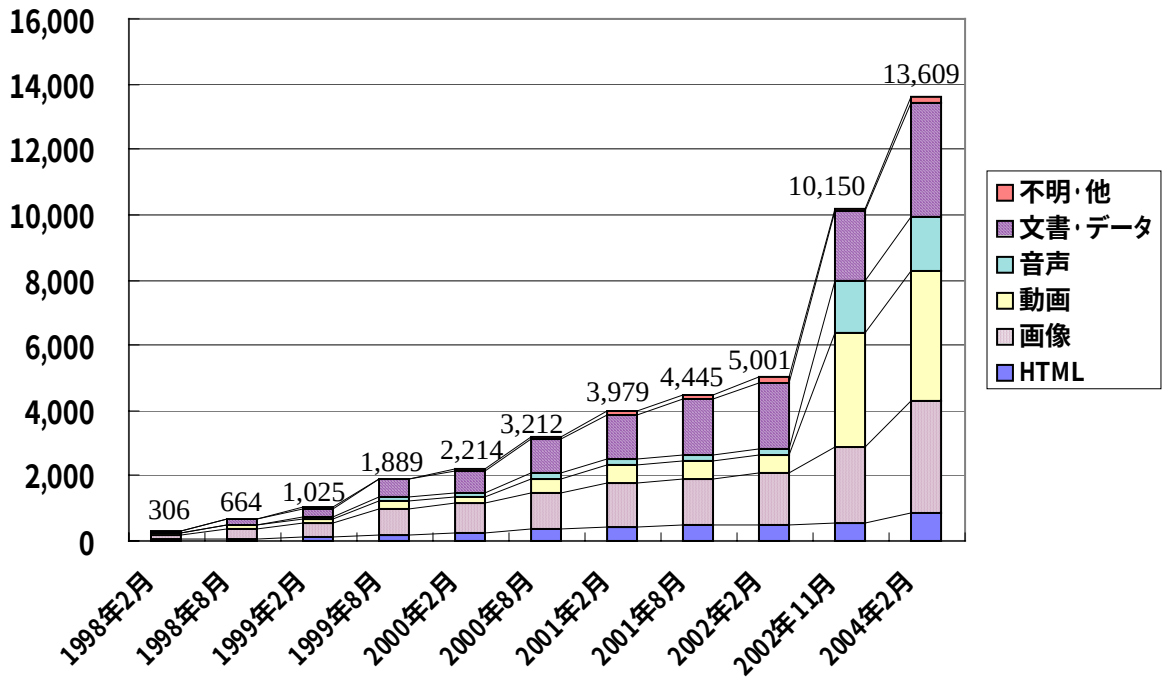
【マルチメディアデータの増大に合わせ順調な伸び】

データ量は、もともとファイルやサーバに比べて伸びが大きい傾向があり、2002年11月調査では前回(2002年２月)調査の２倍という大きな伸びがみられた。今回調査でも前回調査比で３４％と高い伸び率となっている。ファイル種別に見ると、動画・音声ファイルのデータの伸びが大きく、これらのマルチメディアデータの増加が全体のデータ量を押し上げている。

図表　１４　ＪＰドメイン総データ量の推移

	1998年		1999年		2000年		2001年		2002年	2002年	2004年
	2月	8月	2月	8月	2月	8月	2月	8月	2月	11月	2月
総データ量(GB)	305	664	1,024	1,889	2,214	3,212	3,980	4,445	5,002	10,150	13,609
前回比伸び率(%)	－	117.70	54.22	84.47	17.20	45.08	23.91	11.71	12.51	102.92	34.08

図表　１５　ＪＰドメイン総データ量の推移（ファイル種類別）



（単位：ＧＢ）

	1998年		1999年		2000年		2001年		2002年	2002年	2004年
	2月	8月	2月	8月	2月	8月	2月	8月	2月	11月	2月
HTML	46	86	150	211	256	354	411	468	498	564	846
画像	141	306	409	745	885	1,135	1,364	1,440	1,579	2,317	3,461
動画	40	78	113	280	206	434	540	530	543	3,507	3,962
音声	11	29	39	88	119	155	210	211	216	1,575	1,642
文書･データ	53	151	300	546	709	1,057	1,356	1,677	2,009	2,174	3,540
不明･他	15	14	14	19	39	77	98	119	156	13	158
合計	306	664	1,025	1,889	2,214	3,212	3,979	4,445	5,001	10,150	13,609

２-(2)-2)-5　ＪＰドメイン１万ページ当たり平均データ量とＪＰドメイン１サーバ当たり平均ページ数の比較

【サーバ１台当たりの平均ページ数は減少、ページ当たりのデータ量は増加】

Webページの閲覧者がダウンロードすることになるデータ量として、１万ページ当たりの平均データ量を比較すると、1998年２月調査では約０．３ギガバイト（ＧＢ）であるが、2004年２月調査では１．５８ギガバイトと約５．３倍に増加している。現在のWeb閲覧者は６年前より５倍以上のデータをダウンロードしていることになる。

なお、発信者側のサーバに蓄えられる情報量として、サーバ１台当たりの平均ページ数を見てみると1998年２月調査の２８３．３ページから2000年２月調査の４４７．４ページへと１．５８倍まで順調に拡大した後、一転して減少を続けており、2004年２月調査では２０２．１ページと、1998年２月調査の数字を下回っている。すでに立ち上がっているサーバのページ数が大きく変動するとは考えにくいので、新しく立ち上がるサーバの中にページ数の少ないサーバが多数存在し、それが平均ページ数を押し下げているものと思われる。

しかし、ページにリンクされる平均データ量が５．３倍になっていることを考えると、ページ数としては減少しても、データ量として見ればサーバの情報量は着実に増大していると言える。

図表　１６　１万ページ当り平均データ量と１サーバ当たり平均ページ

	総ページ数（万ページ）	1万ページ当り平均総データ量（ＧＢ）	総サーバ数（台）	1サーバ当たり平均ページ数（ページ）
1998 年 2 月	1,020	0.3	36,000	283.3
1999 年 2 月	2,950	0.35	75,000	393.3
2000 年 2 月	4,250	0.52	95,000	447.4
2001 年 2 月	6,101	0.65	152,000	401.4
2002 年 2 月	6,555	0.76	197,000	332.7
2002 年 11 月	7,438	1.37	308,000	241.5
2004 年 2 月	8,590	1.58	425,000	202.1

２-(2)-2)-6　ＪＰドメイン総コンテンツ量発展推移（年別）

▼初回の1998年２月調査の推計値を１００とし、総サーバ数、総ページ数、総ファイル数、総データ量の各年別増加状況を見たものである。

【総データ量は初回調査の４４．６倍】

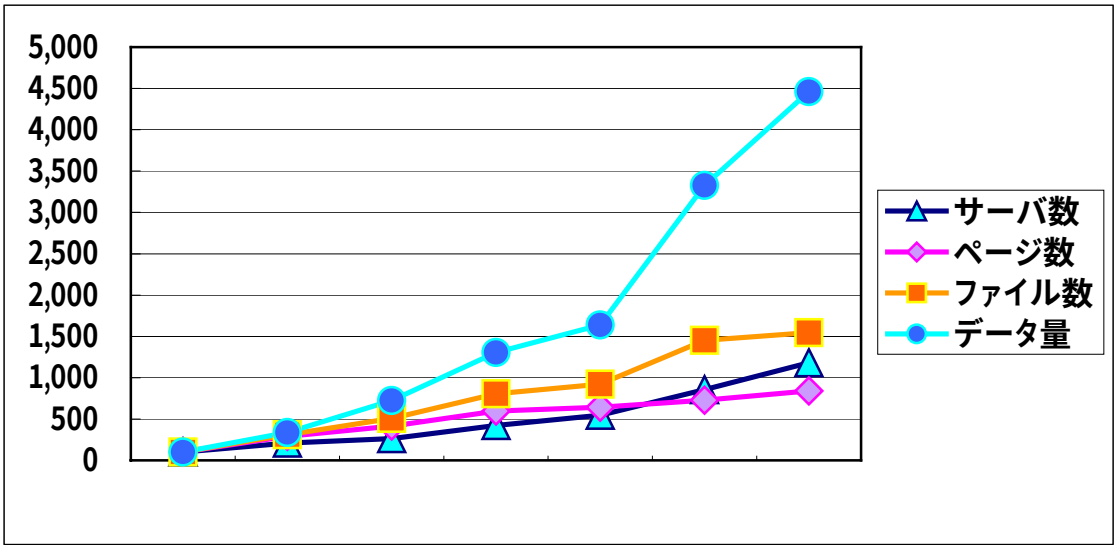
増え方は従来は総サーバ数より総ページ数の方が上回っていたが、近年は総ページ数の伸びが緩やかになり総サーバ数の伸びがあがってきたため、増え方が逆転している。総ファイル数は緩やかな上昇傾向にあるが、総データ量は活発な伸びを示しており、2004年２月は1998年２月に比べ、約４４．６倍にまで増加した。

図表　１７　総コンテンツ量の推移

推計値	1998 年2月	1999 年2月	2000 年2月	2001 年2月	2002 年2月	2002 年11月	2004 年2月
総サーバ数（台）	36,000	75,000	95,000	152,000	197,000	308,000	425,000
総ページ数（万 P）	1,023	2,953	4,255	6,107	6,558	7,438	8,590
総ファイル数（万 F）	1,891	5,822	9,626	15,260	17,388	27,421	29,173
総データ量（ＧＢ）	305	1,024	2,214	3,980	5,002	10,150	13,609

図表　１８　1998 年２月を 100 とする指数で見た総コンテンツ量の推移

指数	1998 年2月	1999 年2月	2000 年2月	2001 年2月	2002 年2月	2002 年11月	2004 年2月
総サーバ数	100	208	264	422	547	856	1,181
総ページ数	100	289	416	597	641	727	840
総ファイル数	100	308	509	807	920	1,451	1,544
総データ量	100	336	726	1,305	1,640	3,328	4,462



３．最近の調査結果から見られた新たな現象

1998 年から国内の Web サーバのコンテンツ量を推計し、定期的に公表している本調査の推計値は、インターネットの規模を表す代表的な指標として広く利用されているところであるが[19]、最近の調査結果から Web の規模の拡大や構造の変化が原因と思われる様々な現象が見られた。調査開始から６年が経過し、調査開始当初には想定していなかった新たな Web 上の現象について考察する。

３-(1)調査設計に関する変化

３-(1)-1)　ＪＰドメインの国内シェアの変化

本調査は、「勃興期にあった日本国内の Web の発展の軌跡をきちんと記録に留めておくこと」を目的に、国内にある Web サーバのコンテンツ量を推計するものとして始められたため、調査対象をＪＰドメインに限定している。

1998 年当時としては、汎用ドメインの国内サイトは特殊で限られたものであったことや、汎用ドメインを有する Web サーバについては、地理的所在を示す情報がとれないので、日本語かどうかは推測できても、国内にあるのかどうか分からない、といった問題があったためである。

しかしながら調査も６年を経て国内の Web の状況も確実に変わりつつあり、2002 年 11 月に単年で調査した汎用ドメインを有する日本語サイトの総コンテンツ量（2002 年 11 月）（図表 19 参照）を見ると、汎用ドメインを有する日本語サーバの推定総数は、465,000 台とＪＰドメインサーバの推定総数 308,000 台を上回っており、汎用ドメインに海外設置サーバの日本語サイトが多数含まれていることを考慮しても、国内に設置された汎用ドメインサイトも相当数に上ると考えられる。

このような状況の中で我が国の Web コンテンツ量の動きを捉えるには、ＪＰドメインのみを調査対象としていることについても見直しをする必要があるか、その場合、仮に対象を広げるにしても、汎用ドメインの日本語サイト中、国内のサーバに設置されたものであるかをどのように確認するのか、というようなことについて検討していくことが必要である。

図表　１９　汎用ドメインにおける日本語サイトの総コンテンツ量（2002 年 11 月調査）

	サーバ数（台）	総ページ数（万Ｐ）	総ファイル数（万Ｆ）	総データ量（ＧＢ）
汎用ドメイン中日本語によるサイトの総コンテンツ量	465,000	2,502	9,642	5,444
ＪＰドメインの総コンテンツ量	308,000	7,438	27,421	10,150

[19]「ＩＴ戦略会議ベンチマーク集　平成 15 年 12 月 18 日」http://www.kantei.go.jp/jp/singi/it2/dai22/22siryou6.pdf

３-(1)-2) ロボットで取得できないマルチメディアデータの増加

サーチロボットはすべての Web ページにアクセスできるようにはなっておらず、次のような Web ページやファイルのデータは収集できない。

・CGI 等のスクリプトで自動生成されるもの
・再生、停止ボタンが埋め込まれている動画ファイル
・ストリーミングの動画、音声ファイル

こうしたページやファイルは近年 Web 上で主要なツールとして普及してきており、動画や音声についてはむしろこの様な形態のものが主流となっているようである。従って本調査では動画・音声の急激な増加が観測されてはいるものの、情報量推計という観点からは動画・音声の主要な部分が把握できていないという問題がある。

３-(1)-3)　Web 規模の増大と調査インフラ増強の問題

Web の規模は年々拡大を続けており、短期間に多くの Web ページにアクセスできるよう Loki も年々改良しているが、そのような対応についても技術的にもコスト的にもおのずから限界がある。

これまで、Loki のプログラムを並列化することにより、複数のマシンを同時に稼動させて各マシンに適切に不可分散を行うなどの改良を行ってきたが、現実には一定数以上の並列処理になると様々な不具合が生じているところであり、この手法による増強にも限界がある。

３-(1)-4)　ネットワークのトラフィック増大によるハブの渋滞

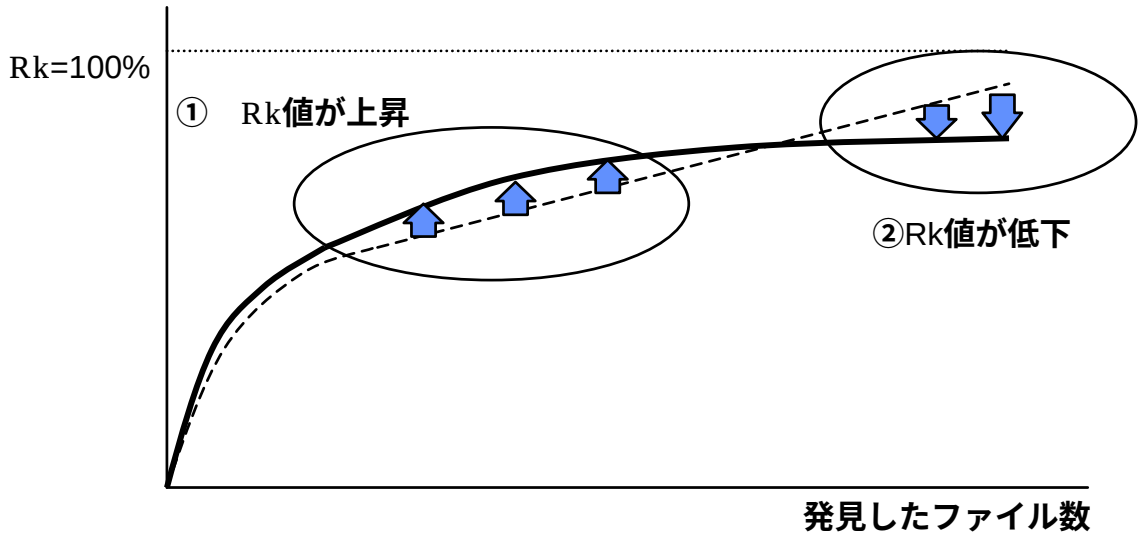
Loki の走査は、ネットワーク環境に左右されるところ、近年調査では、トラフィック増大によるハブの渋滞が走査期間を長引かせる原因となっている。マシンのみならず、回線能力を主とする設置環境の改善（なるべくバックボーンに近いところからインターネットに入るなど）も重要な課題となっている。

３-(2)　Ｗｅｂ構造の変化

３-(2)-1)　最近の調査結果から見られた現象

1-(1)　調査の仕組み・原理のところで述べたように、本調査は既知 URL の発見率（Rk 値）のグラフをもとにファイル総数を推計しているが、そのグラフに関して、最近の調査結果では従来見られなかった２つの現象が生じている。

図表　２０　Ｌｏｋｉ　既知ＵＲＬ発見率（RK）グラフの２つの現象



①Rk 値の上昇、曲線化

Rk 値グラフは当初急速に立ち上がり、ある程度までくると序々に傾きを緩やかにしつつ曲線を描き、最後には傾きが緩やかな直線になってくる。Rk 値グラフの傾きが当初急であることは、取得しているファイルの被リンク数が多いことを、傾きが曲線を描きつつ緩やかになっていくのは、１日の取得ページの平均被リンク数が絶えず減少しつづけていることを、最後に傾きが緩やかな直線になるのは、被リンク数が落ちるところまで落ち、疎な状態で均質化していることを示す。

これが調査開始当初に描いていた Web 構造のモデルであり、調査開始当初からこれを裏付ける Rk 値のグラフを得ていたところであるが、最近の調査結果では Rk 値のグラフが急速に上昇し、その後曲線が緩やかになってもなかなか直線化せず、長い期間曲線を描きつづける現象が生じている。今回の調査でもはっきりした直線が得にくく、直線の判断が極めて難しい状況であった。

②調査終盤の Rk 値の横這い化

本調査の推計モデルでは、Rk 値のグラフは最後に緩やかな傾きを持った直線となることが重要であり、この傾きに沿って既知リンク率１００％まで延長線を引くことで推定総数が求められる。これまでの結果では、これを裏付ける結果があらわれていたが、直近の結果では、上記①のようにはっきりとした直線が出にくくなった上、直線化した後にさらに横這いを続けるという現象が起きている。これは実績としての取得ファイル数は増えているにもかかわらず、Rk 値の上昇が止

まっていることを意味する。

３-(2)-2)　原因となる Web 構造の変化の推論

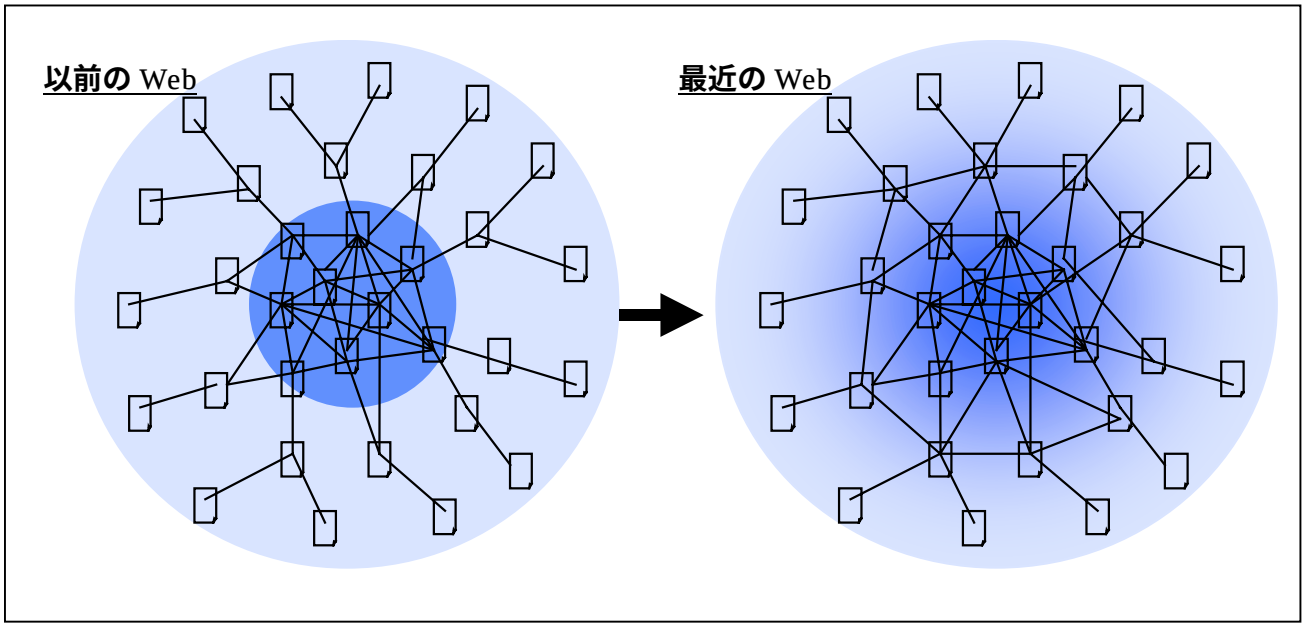
３-(2)-1)で挙げた、最近の調査結果で見られた現象について、その原因を推論する。

【中間的リンク規模のＷｅｂページの増加】

調査開始当初想定していた Web 構造のモデルは、リンクの多く貼られている Web ページと、そうでない Web ページとが２極化しているという前提に立ったものであるが、Rk 値が急速に上昇したあと曲線が緩やかになってもなかなか直線化せず、長い期間曲線を描き続ける現象が見られたことから、最近はこのような２極化をしておらず、中間的な、様々な被リンク数の Web ページが多数出現して、疎で均質なページとの境界があいまいになってきているものと思われる。

これは、Web の利用が一般にまで広く普及するとともに、その本来的な特徴ともいえる情報を「つなぐ」という機能が一般の Web ページにも広く浸透し、Web 構造の成熟化の進んでいることを示していると考えられる。

図表　２１　中間的リンク規模の Web ページの増加



【デッドリンクの増加】

調査終盤の Rk 値の横這い化現象については、Web 周辺部でデッドリンク（リンク先が存在しない、あるいはロボットが辿れないリンク）が増加したためではないかと考えられる。これまでの推計の前提は、デッドリンクは Web 全体でせいぜい数%程度であり、Rk 値の推移に大きな影響は与えないというものであった。しかし周辺部へ行くほどデッドリンクが増加し、かなりの割合でデッドリンクが存在するなら、Rk 値のグラフは１００%に向かわなくなる。

デッドリンクを生む原因としては、長年メンテナンスされていない古いページの増加、アクセス制限をかけているページの増加、自動生成ページの増加が考えられる。

また、携帯電話向けサイトなど、ブラウザ種別に応じてコンテンツを提供するページの増加も原因と考えられる[20]。

[20] Web 空間内にある「勝手サイト」と呼ばれる携帯向けサーバについては、Loki は携帯用のブラウザを有しておらず、勝手サイトからはアクセスを拒否され、デッドリンクとしてカウントしてしまう。