WWWコンテンツ統計調査報告書

～企業等のＷｅｂサイトにおけるコンテンツ情報発信量～

平成１９年３月

総務省　情報通信政策研究所

はじめに

Web サイトは、今や企業、国、大学といった組織から個人に至るまで、自ら情報を発信し、伝達する手段として広く利活用されている。これらの Web サイトはテキスト、画像、動画等の複数のファイルから構成されているが、我が国の Web サイトにおいてどの程度のファイルが存在しているのかという疑問があってもそれに回答する定量的なデータはこれまでほとんど公表されてこなかった。

総務省情報通信政策研究所では平成 10 年からインターネット関連の統計データ整備の一環として Web コンテンツ量（Web サイトにおけるコンテンツのファイル数やデータ量等）の調査研究を行ってきた。調査研究を開始した当初は我が国の Web サイトのドメインは JP ドメインが大半であり、JP ドメインの URL を持つ Web サイト上のファイル数をもとに我が国の Web コンテンツ量を推計してきたところである。その成果は多くの研究機関や研究者等に引用されており、継続的なデータ整備が重要であると考えてきたが、インターネット上で提供されるサービスの発展や関連する技術の発展によって Web サイトの利活用の構造が大きく変化し、従来の調査手法のままでは Web コンテンツ量の推計を行うことが困難になってきたことも事実である。

そこで調査手法を抜本的に見直すこととし、我が国全体ではなく特定組織の Web サイトに焦点を当て、各組織の Web サイトにおけるコンテンツ（ファイル）量を集計することとした。具体的な調査内容としては、各組織の Web コンテンツに焦点を当てて企業（東証一部上場）、国（各省庁等）、都道府県、市町村、大学及び ISP 系ポータルサイトについて組織単位で Web コンテンツ量を集計・分析することとした。これにより組織ごとの Web サイトのコンテンツの状況を把握することが可能になり、その傾向を分析した結果をとりまとめたのが本報告書の内容である。

本調査研究を進めるにあたっては、荒木万寿夫青山学院大学経営学部助教授及び山本渉電気通信大学電気通信学部講師に貴重なアドバイスを頂いた。また、三菱総合研究所の方々に格別のご協力を頂いた。心から御礼を申し上げたい。

本研究成果を Web に関する基礎データとして様々な機会において有効活用して頂ければ幸いである。

平成１９年３月

総務省情報通信政策研究所　調査研究部

目次

1　調査目的 3
1-1　背景と目的 3
2　調査概要 4
2-1　本調査におけるクローリング手法の概要 4
2-1-1　本調査におけるクローリングの仕組み 4
2-1-2　本調査におけるクローリングの具体例 5
2-1-3　タイムアウト 8
2-2　調査実施状況 9
2-3　調査対象 10
2-3-1　調査対象組織 10
2-3-2　調査対象ファイル 11
3　クローリング調査結果 14
3-1　調査対象組織全体の結果 14
3-2　組織別の分析 22
3-2-1　企業（東証一部上場） 22
3-2-2　国（各省庁等） 26
3-2-3　都道府県 27
3-2-4　市町村 28
3-2-5　大学 30
4　まとめ 32
参考　Webサイトのコンテンツの概況 33

1　調査目的

1-1　背景と目的

Webサイトは、今や一般企業、国、大学といった組織体から個人に至るまで自ら情報を発信する手段として利活用されている。このWebサイト上に存在するWebページ[1]は、検索エンジンによる検索の対象とされることはあっても、そもそも我が国のWebコンテンツがどのくらいあるのか、という定量的なデータはほとんど公表されていない[2]。

情報通信政策研究所では、Webサイトの利用が本格化し始めた平成 10 年から独自に開発したクローラー[3]（LOKI：ロキ）を使って、日本のカントリーコードであるJPドメイン[4]を有するWebページのURLリンクを辿ってファイルを収集し、Webサイトのコンテンツ量調査を継続的に実施してきた[5]。この調査結果は、我が国のWebコンテンツ量に関する唯一のデータとして「情報通信白書」や「e-Japan戦略ITベンチマーク集」等に掲載され、多くの研究機関や研究者等に引用されてきたが、Webの発展スピードとデータ規模の拡大は調査開始当初に想定していた以上に速く、LOKIのような小規模クローラーによる調査手法の限界が顕著となったため、平成 16 年 2 月調査を最後としていた。

今回、調査手法と対象を抜本的に見直し、我が国全体ではなく特定組織の Web サイトに焦点を当て、各組織の Web サイトにおけるコンテンツ（ファイル）量を把握し、各組織の属性による傾向等を分析することとした。具体的な調査内容としては、各組織の Web コンテンツに焦点を当てて企業（東証一部上場）、国（各省庁等）、都道府県、市町村、大学及び ISP 系ポータルサイトについて組織単位で Web コンテンツ量を集計・分析している。

---

[1] 「Web ページ」は HTML ファイルの他、ページを構成する要素として画像ファイルや動画ファイル等も含まれるが、この報告書では「ページ数」を HTML ファイルの数と同義とすることにした。、最近は HTML ファイルでなく Flash ファイルのみのページもあり、その構造は複雑になりつつなる。それに対し、Web サイトは Web ページの集まりと考えることができる。

[2] Web コンテンツ量を推計した過去の調査事例として、S. Lawrence and C.L.Giles「Searching the World Wide Web」（Science Vol280、pp.98-100、1998）や山名早人ら「分散型 WWW ロボットによる WWW 情報収集」（データ工学ワークショップ DEWS98、No.24（1998.3.5-7））等が挙げられる。

[3] クローラーとは Web 上の情報を取得し、自動的にデータベース化する Web 検索プログラム。クローラーを用いてインターネット上の Web ページのリンクをたどりながら情報を収集する行為をクローリングといい、検索エンジンによる Web サイトの検索結果（URL リスト）は、このクローリングによって収集されることが多い。

[4] ドメインは URL の一部であり、「～.co.jp」や「～.go.jp」、ﾊﾞ .com」、ﾊﾞ .net」等を指す。例えば、http://www.soumu.go.jp/hoge/なら「soumu.go.jp」がドメインとなる。

[5] 従来の調査手法は、既知 URL リンク発見率と新規発見 URL 数をグラフにプロットし、相当程度のファイルを収集した段階で線形近似法により全体のファイル数を推計するという方法。詳細については情報通信政策研究所調査研究報告書「WWW コンテンツ統計調査報告書～我が国の Web 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 10 月）を参照のこと。

2　調査概要

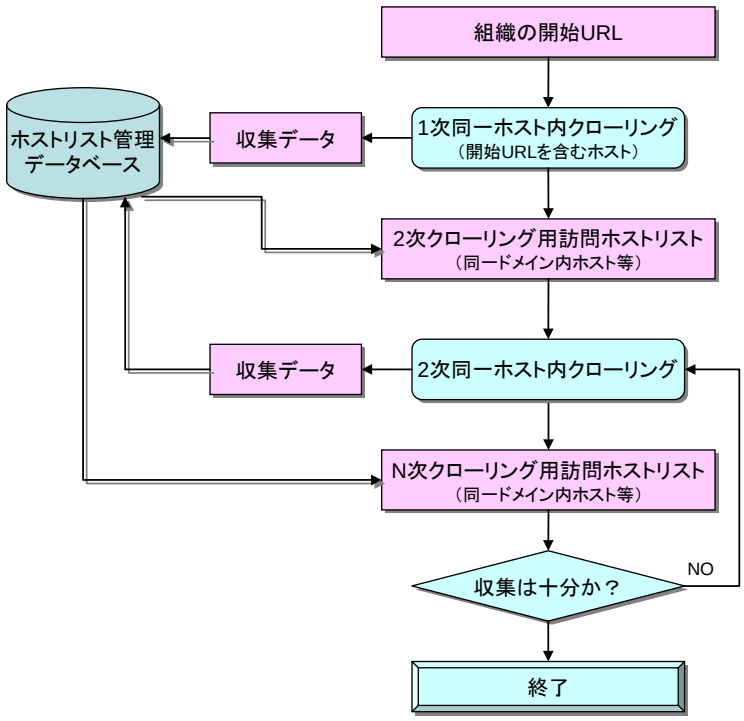
2-1　本調査におけるクローリング手法の概要

本調査では、URL ベースで指定した Web サーバ内のファイルについてクローリングを実施し、企業（東証一部上場）、国（各省庁等）、都道府県、市町村、大学及び ISP 系ポータルサイトについて組織単位で Web コンテンツ量を集計した。

2-1-1　本調査におけるクローリングの仕組み

組織毎に 1 つのトップページを特定し、それを当該組織の開始URLとする。まず、開始URLが含まれるホスト（1 次ホスト）内のページについてクローリングを実施する（1 次クローリング）。次に、1 次クローリングで収集されたリンク情報をもとに、1 次ホスト以外で同一組織に属すると思われるホスト（2 次ホスト）を抽出し、それらのURLリストを作成する。作成したURLリストに基づいて、ホスト内クローリング（2 次クローリング）を実施し、さらに 2 次クローリングで収集されたリンク情報をもとに、1 次、2 次ホスト以外で同一組織に属すると思われるホスト（3 次ホスト）を抽出し、それらのURLリストを作成する。このようにクローリングとURLリストの作成を繰り返し、収集が充分と判断できる段階で終了とする[6]（図表 2-1　クローリングの仕組み）。

図表 2-1　クローリングの仕組み

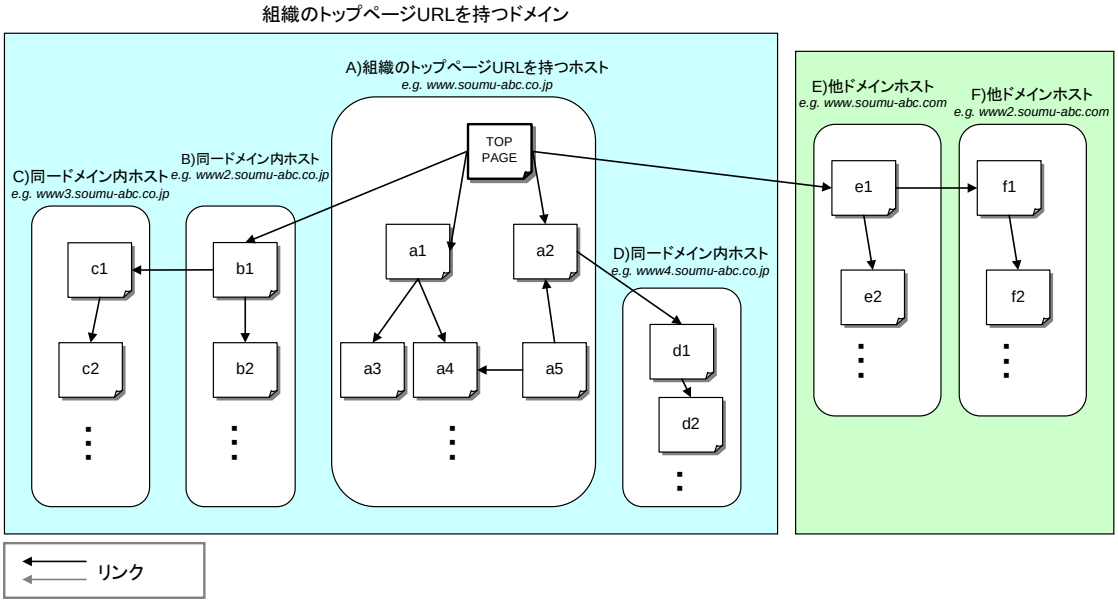


[6] 今回の調査においては、2 次クローリングを終了した時点で、調査対象組織のデータ収集を終了とした。本来、調査の終了条件は、調査対象組織について新しいホストが出現しなくなることであるが、これを厳密に満たすための時間は膨大なものとなると予想されるため、本調査では 2 次クローリングを終了した時点で調査終了としている。

2-1-2 本調査におけるクローリングの具体例

本調査におけるクローリングの範囲を示すため、図表 2-2 に示す一般的なウェブサイト例を用いる。一般に、組織は複数のホストで構成されたウェブサイトを持っている（A～F）。組織とドメインは一定の対応関係があるが、複数のドメインを利用している組織もある[7]。図表 2-2 は、組織のトップページと同一ドメインのホスト（A-D）と、他ドメインのホスト（E,F）を利用している組織の例である。

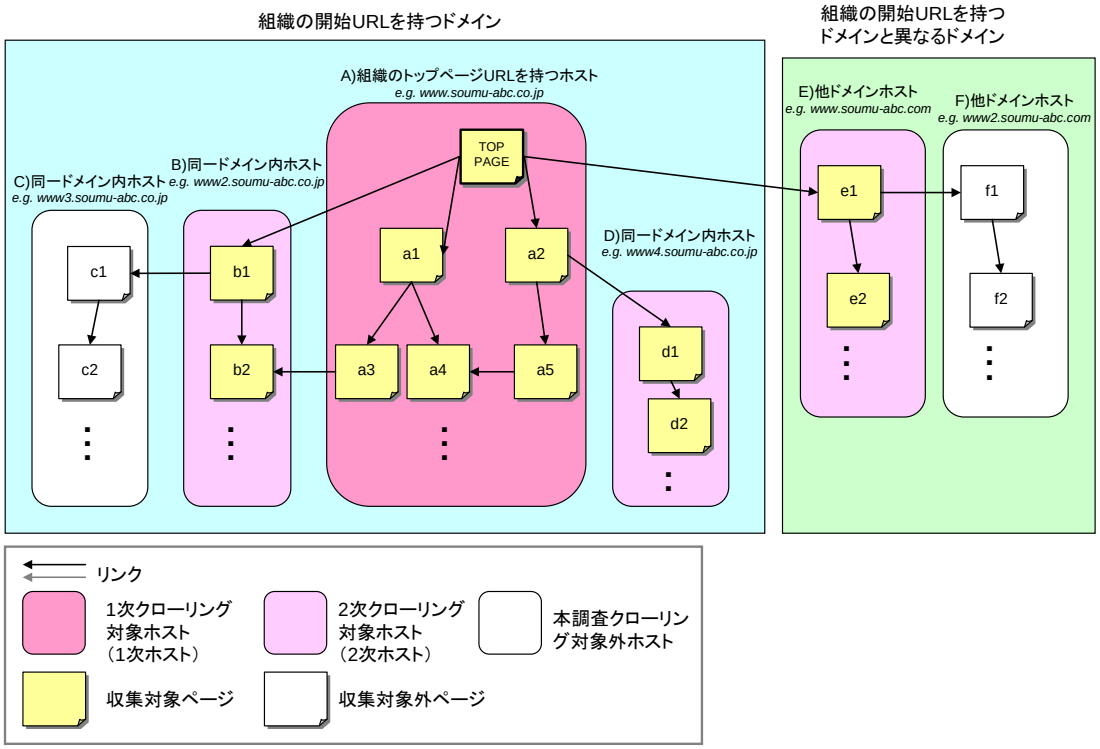
図表 2-2 ウェブサイト例



[7] 複数組織で同一ドメインを利用している場合もある。

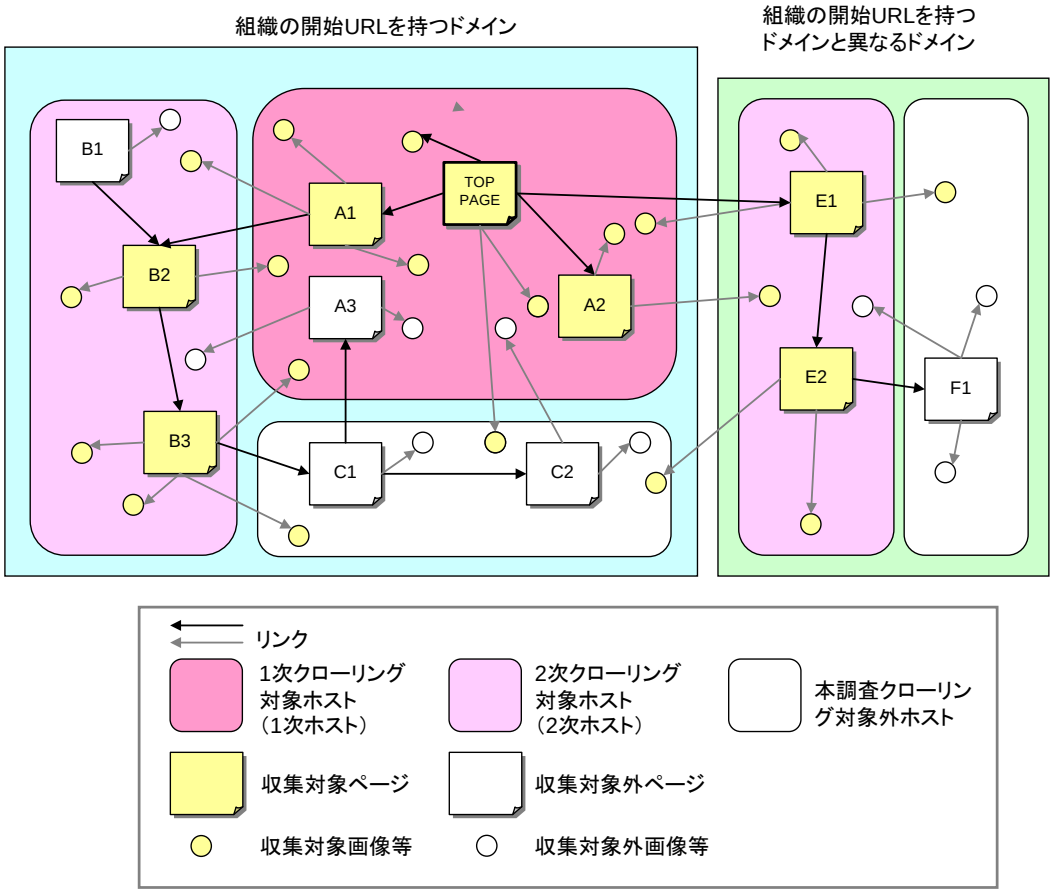
図表 2-3 に本調査での調査範囲を示す。本調査では、組織の開始 URL(TOP PAGE)を起点として、当該組織のクローリングを行う。従って、2 次クローリングの範囲内で TOP PAGE からリンクでたどれるページ以外のページの収集は出来ない。また、たどれるページについても、本調査においては、2 次ホストまでの範囲のみ収集している（C,F は収集しない）。

図表 2-3 本調査での調査範囲



図表 2-4 に本調査でのファイルレベルでの調査範囲を示す。本調査ではホスト内クローリングを繰り返すことによって組織のクローリングを実施している。ここで、ホスト内か、ホスト外かの判断を行うのはページ（HTML ファイル等）についてのみである。ホスト内と判断されたページの構成要素（画像等）は、ホスト内外に関わらずすべて収集する。従って、A1、A2 等の 1 次ホスト内のページ、B2、E1 等の 2 次ホスト内のページについては、それらに含まれる画像等のファイルはすべて収集する。

図表 2-4 ファイルレベルで見る本調査での調査範囲

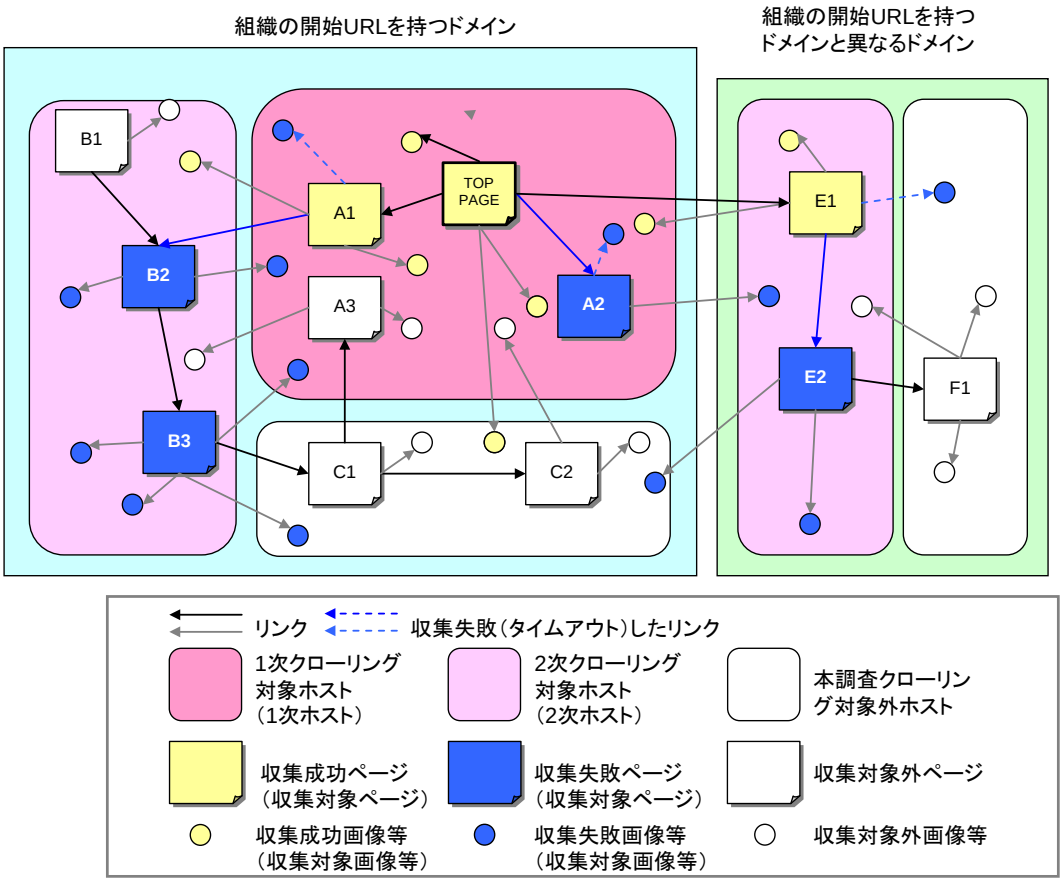


2-1-3　タイムアウト

タイムアウトとは、データ転送などを行なう際、一定の時間が経過して処理が終わらなかったり相手から返事が無いことであり、その場合には、処理を打ち切って終了させるタイムアウト処理を行う。

個々のファイル取得に際して、タイムアウトが生じた場合、当該ファイルを収集失敗として、再取得は行わず、次のファイルの取得を試みる。あるホストのクローリングにおいて、一定回数連続してファイル取得がタイムアウトした場合、そのホスト自体の収集を中止する。本調査においては、収集を中止したホストについては、再度 TOP PAGE から取得を試みた。

図表 2-5　タイムアウトが生じた場合のファイルの取得



2-2　調査実施状況

URL ベースで指定した Web サーバ内のファイルについて、平成 18 年 8～10 月にクローリングを実施した。クローリング調査の実施状況は図表 2-6 のとおり。

図表 2-6　クローリング調査の実施期間

クローリング概要	対象	実施期間
試験クローリング[8]で検証された結果及び明らかになった課題を元に、各組織に対してクローリング調査を実施し、分析可能なデータを取得する。	企業 国 都道府県 市町村 大学 ISP 系ポータルサイト	平成 18 年 8～10 月

[8] 試験クローリングは、独自に開発したクローラーが正しく動作するかどうかの予備調査として平成 18 年 2 月～3 月に実施した。各組織における限定数の Web サイトに対してクローリングを実施することにより、新規開発したクローラーをテストすると共に、組織毎の Web サイトのコンテンツ内容を概観した。

2-3　調査対象

2-3-1 調査対象組織

今回、調査対象として選定した組織は、図表 2-7 のとおりである。調査対象組織別の Web サイトの URL リストを作成し、各 Web サイトに対してクローリング調査を実施した。

図表 2-7 調査対象組織

組織種	調査対象組織 （URL リストの作成方法）	対象組織数※1
企業	東京証券取引所第一部上場企業 （東京証券取引所、東洋経済新聞社、株式新聞社の企業 URL データ情報を基に抽出）	1,698
国	各省庁及び関連組織 （首相官邸及び電子政府の総合窓口（e-Gov）の Web サイトのリンク集より抽出）	72※2
都道府県	都道府県 （(財)地方自治情報センターの Web サイトのリンク集より抽出）	47
市町村	市町村 （総務省資料）	1,820
大学	国公私立 4 年制大学 （文部科学省の Web サイトのリンク集より抽出）	709
ISP 系ポータルサイト	インターネットサービスプロバイダ（ISP）のうち、加入率上位のもの （該当 ISP の Web サイトから直接作成）	5
	合計	4,351

※1　対象組織数は、クローリングによるファイルの取得が可能であったもの。
※2　国の対象数は、各省庁の関連組織として可能性がある組織の Web サイトを含めた。

2-3-2 調査対象ファイル

今回の調査で対象とするファイルは、HTMLファイルの他、原則ページ内に存在している画像ファイルやPDFファイル[9]等を取得することとしたが、その種類は多数存在する。

ファイルの種類の判断基準としては、拡張子やコンテントタイプ（Content-Type）がある。今回はコンテントタイプを主な判断基準とし、図表 2-8 の判別ルールでファイルを収集・集計した[10]。ただし、PDFファイルとFlashファイル[11]については拡張子が「.pdf」又は「.swf」のもので判断されるので、その 2 種類のファイルのみ拡張子で判断した。つまり、収集されたデータの集計を行う際の手順としては、最初に拡張子が「.pdf」又は「.swf」に完全一致するコンテンツを抽出し、「.pdf」として抽出されたものをPDFファイル、「.swf」として抽出されたものをFlashファイルとした。その上で残りのファイルについては、コンテントタイプの前方一致で分類した。

図表 2-8 ファイルの種類の判別ルール

区分	説明	判別ルール Content-Type	拡張子
FLASH	Flash ファイル	-	swf
PDF	PDF ファイル	-	pdf
HTML	HTML/XHTML	text/html	-
	XML	text/xml application/xml application/rdf+xml	-
テキスト	プレーンテキスト	text/plain	-
	その他のテキストファイル	他に挙げられていない text/*	-
スクリプト	CSS	text/css	-
	JavaScript[12]	application/x-javascript text/javascript	-
	VBScript	text/vbscript	-
オフィス	MS-Word ファイル	application/msword	-
	MS-Excel ファイル	application/vnd.ms-excel	-
	MS-PowerPoint ファイル	application/vnd.ms-powerpoint	-
	一太郎ファイル	application/x-js-taro	-
	RTF	application/rtf	-

[9] PDF ファイルとは、Adobe 社が開発した、特定の OS やブラウザに依存せずに文書を表示できるドキュメントフォーマットである。

[10] 一般に拡張子やコンテントタイプが Web サイトに置かれているファイルの種類の判断基準とされているが、ファイルによっては両者ともに必ずしも正確なファイルの種類を記しているとは限らない。試験クローリングの結果、後者の方がファイルの種類をより正確に記していることが多かったので、今回の調査ではコンテントタイプを採用した。ただし、PDF ファイルや Flash ファイルのように、使用するブラウザやソフトウェアによってコンテントタイプではなく拡張子のみで判断されてしまうファイルの種類もあるため、それらの特性を考慮した集計手順にした。

[11] Flash ファイルとは、Web 用アニメーションを制作できるオリジナルフォーマットのファイルである。

[12] JavaScript とは、Web ブラウザの利用に適したスクリプト言語（簡易プログラミング言語）、またはその言語で記述されたプログラムのことである。

アーカイブ	ZIP	application/zip application/x-zip-compressed	−
	stuffit	application/x-stuffit	−
	binhex	application/mac-binhex40	−
	LHA	application/x-lha-compressed	−
	tar	application/x-tar	−
	gzip	application/gzip application/x-gzip	−
	JAR	application/java-archive	−
アプリケーション	その他のアプリケーション	他に挙げられていない application/*	−
画像	GIF	image/gif	−
	JPEG	image/jpeg	−
	TIFF	image/tiff	−
	PNG	image/png image/x-png	−
	BMP	image/bmp image/x-bmp image/x-ms-bmp	−
	その他	他に挙げられていない image/*	−
音声	MPEG/MP3	audio/mpeg audio/x-mpeg	−
	WAV	audio/wav audio/x-wav	−
	AIFF	audio/aiff audio/x-aiff	−
	au	audio/basic	−
	MIDI	audio/midi audio/x-midi	−
	Real Media（ストリーミング以外）	audio/x-realaudio audio/x-pn-realaudio audio/x-pn-realaudio-plugin	−
	Windows Media（ストリーミング以外）	audio/x-ms-wma	−
	その他の音声	他に挙げられていない audio/*	−
音声ストリーミング	ストリーミング[13]	audio/x-ms-wax	−
動画	MPEG	video/mpeg video/x-mpeg	−
	Quick Time	video/quicktime	−
	Real Media（ストリーミング以外）	application/vnd.rn-realmedia video/vnd.rn-realvideo	−
	AVI	video/x-msvideo	
	Windows Media（ストリーミング以外）	video/x-ms-asf video/x-ms-wmv	−
	その他（ストリーミング以外）	他に挙げられていない video/*	−
動画ストリーミング	ストリーミング	video/x-ms-wvx video/x-ms-asx	−
バイナリ	バイナリファイル	application/octet-stream	−
その他	その他	x-*/*、model、message、multipart、memo など	−

[13] ストリーミング形式とは、動画や音声等のマルチメディアデータを視聴する際にデータを受信しながら同時に再生を行なう配信形式のことである。

また、今回の調査では、一般に公開されている Web サイト及びそのサイト内のファイルを調査対象としている。

なお、専用クローラーを使用するため、専用クローラーで取得できない、あるいはアクセスできない以下のようなファイルは、調査対象外となる。

・組織のトップページからリンクで辿ることが出来ないファイル
・アクセスに認証を必要とするファイル、外部に公開されていないネットワークに存在するファイル
・「http:」や「https:」以外から始まる URL を持つファイル（「mailto:」や「ftp:」等）
・携帯電話機等の専用ブラウザや特定言語を指定した際に返されるファイル
・タイムアウトしたWebサイト及びファイル[14]
・CGI等のプログラムで、引数を用いて動的生成されるページ[15]
・FLASHファイルからリンクされたファイル[16]
・ストリーミングの動画・音声データ[17]

[14] メンテナンスや混雑の影響があるため、一定回数連続してタイムアウトが発生した Web サイトは Web サイトごとタイムアウトと判定している。しかしながら、一度タイムアウトした Web サイトでも、何度か再クローリングを試み、出来る限りファイルの収集に努めた。

[15] Web ブラウザから引数を入力して、要求を受けた Web サーバがそれに対応するプログラムを起動し、実行結果を Web ブラウザに返すもの。動的生成される代表的な Web サイトの例として、ブログ、掲示板、チャット、フォームメール、ショッピングサイト、検索エンジン等がある。

[16] Flash は音声とアニメーションを組み合わせて Web コンテンツを作成するソフトウェア。Flash で作成されたファイルのリンク先は、HTML ファイルと異なり、クローラーでは解析できない。

[17] ストリーミングは、動画や音楽ファイルをダウンロードしながら再生する技術で、動画・音声データ自体の所在を記述した定義ファイルが置かれている。クローラーは定義ファイルのみを収集している。

3　クローリング調査結果

3-1　調査対象組織全体の結果

クローリング調査における各対象組織（企業、国、都道府県、市町村、大学、ISP 系ポータルサイト）及び全体の集計結果を以下に示す。

調査対象組織全体におけるコンテンツ情報発信量（図表 3-1）を見てると、調査対象組織全体で収集したページ総数（HTMLファイル数）は約 1,360 万ページ、ファイル総数は約 3,890 万ファイル、データ総量[18]は約 2.7TBである。全体のファイル数をファイル種別に見てみると、画像ファイルが最も多く、続いてHTMLファイル、PDFファイルの順となっており、全体のデータ量をファイル種別に見みると、PDFファイルが最も多く、続いて画像ファイル、動画ファイル、HTMLファイルの順となっている。これらは、ファイル種別の構成比（図表 3-4）からも確認することが出来る。このようにファイル数では、画像ファイルとHTMLファイルがPDFファイルよりも多いにもかかわらず、データ量ではPDFファイルが最も大きくなっているが、その要因として、画像ファイルやHTMLファイルは 1 ファイルあたりのデータ量がPDFファイルに比べ小さく、PDFファイルは 1 ファイルあたりのデータ量が比較的大きいことが考えられ、1 ファイルあたりの平均データ量（図表 3-7）により確認することができる。

1 組織あたりの平均コンテンツ情報発信量（図表 3-2）を見ると、調査対象組織全体における 1 組織あたりの平均値は 3,192 ページ、9,132 ファイル、654MB[19]となっている。ファイル数について、総数を組織種別に比較すると、ISP系ポータルサイトが最も多く、続いて都道府県、国の順となっており、ファイル種別で見てみると、どの組織においても画像ファイルが占める割合が最も大きく、続いてHTMLファイルが大きい。また、国、都道府県及び市町村においては、他の組織に比べ、PDFファイル、オフィスファイルの占める割合が大きい。国、都道府県及び市町村においてPDFファイル、オフィスファイルの占める割合が大きい要因として、国、都道府県及び市町村が公開する行政文書がPDFファイル、オフィスファイルの形式で公開されることが多いということが考えられる。ISP系ポータルサイトにおいては、他の組織に比べ動画ファイルの占める割合が大きい。また、データ量について、総量を組織種別に比較すると、都道府県が最も大きく、続いて国、ISP系ポータルサイトの順に大きい。

1 ホスト[20]あたりの平均平均コンテンツ情報発信量（図表 3-3）を見ると、1 組織あたりの平均（図表 3-2）とは異なり、調査対象組織全体における 1 対象ホストあたりの平均値は 938 ペ

---

[18] コンピュータが扱う情報の最小単位を「bit (ビット)」といい、省略形として「b」と表す。その上の単位として「Byte (バイト)」が存在し、省略形として「 B」と表す。 8b=1B であり、おおよそ一文字が 1～2B に相当す.る。一般的に、大きいデータ量を表す際は、K (Kilo/キロ)， M (Mega/メガ)， G (Giga/ギガ)， T (Tera/テラ) などの倍量単位が用いられる。本報告書に用いる単位としては、1024 B = 1 KB ， 1024 KB = 1 MB, 1024 MB =1 GB, 1024 GB =1 TB である。

[19] 図表 3-2 においては、KB 表示であるが、本文では 1024 KB = 1 MB として、MB に換算しているため、図表 3-2 と本文で直接数値が一致していないことに留意。。以下、図表と本文において、単位を変換しているものについては同様の留意が必要。

[20] ここでのホストは、クローリングを試みた対象ホストの数であり、ホスト名や IP アドレスの組み合わせが異なるホストは別のホストとして数えている。クローラーは対象ホスト内のページ（HTML ファイル）を取得するが、それらのページに含まれる画像等は他のホストのものであっても取得する。

ージ、2,684 ファイル、192MBとなっている。ファイル数について、総数を組織種別に比較すると、国が最も多く、続いて都道府県、ISPとなっている。データ量について、総量を組織種別に比較すると、国が最も大きく、続いて都道府県、市町村の順に大きい。つまり、国や都道府県等の政府系のWebサイトでは、1 ホストあたりのコンテンツ量が比較的大きいということが伺える。他方、企業や大学では 1 ホストあたりのコンテンツ量が小さく、企業や大学では複数のドメイン名を取得したり、多くのホストを利用することにより、コンテンツを複数のホストに分散させていることが伺える。

発信されるコンテンツ情報のファイル数におけるファイル種別の構成比（図表 3-4、3-5）によると、どの組織種においても画像ファイルと HTML ファイルが占める割合が大きいが、企業、大学は画像、ＩＳＰはＨＴＭＬの割合が比較的高い。また、国、都道府県及び市町村はＰＤＦの割合が他の組織に比べてやや高い。一方、発信されるコンテンツ情報のデータ量におけるファイル種別の構成比（図表 3-4、3-6）によると、組織種ごとにファイル種の割合を見ると、ISP 以外の組織では、いずれもＰＤＦの割合が最も大きいが、その中で大学や企業におけるＰＤＦの割合は比較的低くなっている。また、ISP では、画像ファイルの割合が最も大きく、次いでＨＴＭＬが大きくなっており、他の組織と異なっている。さらに大学、市町村、ＩＳＰでは、動画の割合がＰＤＦや画像に次いで大きくなっている。

発信されるコンテンツ情報の 1 ファイルあたりの平均データ量（図表 3-7）によると、調査対象組織全体では動画ファイル（MPEG ファイル等）が 4.4MB と最も大きく、次いでアーカイブファイル（ZIP ファイル等）の順となっている。

図表 3-1　調査対象組織全体におけるコンテンツ情報発信量

		企業	国	都道府県	市町村	大学	ISP	全体
概要	対象組織数	1,691	71	47	1,752	694	5	4,260
	対象ホスト数	4,031	470	555	3,178	6,037	220	14,491
ファイル数	総数	10,297,754	3,646,648	4,191,148	9,970,974	10,063,993	730,569	38,901,086
	FLASH	24,809	2,220	3,280	5,573	12,911	720	49,513
	PDF	599,421	561,374	650,428	1,191,592	485,497	663	3,488,975
	HTML（ページ数）	2,969,352	1,321,414	1,311,029	3,722,535	3,926,488	348,932	13,599,750
	テキスト	17,005	85,481	24,254	42,218	219,314	2,675	390,947
	スクリプト	38,683	7,976	16,451	30,754	45,783	1,670	141,317
	オフィス	4,148	116,186	212,798	717,240	43,846	13	1,094,231
	アーカイブ	11,414	933	390	873	38,566	222	52,398
	アプリケーション	4,182	8,370	6,800	3,337	38,377	215	61,281
	画像	6,581,975	1,530,266	1,943,942	4,219,735	5,187,893	361,527	19,825,338
	音声	18,067	2,312	12,881	14,125	18,571	1,631	67,587
	音声ストリーミング	3	15	0	20	2	0	40
	動画	11,052	1,754	3,083	6,297	14,374	12,125	48,685
	動画ストリーミング	305	30	125	1,130	55	7	1,652
	バイナリ	17,225	8,304	4,671	15,146	29,534	168	75,048
	その他	113	13	1,016	399	2,782	1	4,324
データ量（KB）	総量	494,999,210	346,160,172	312,390,195	736,428,973	949,276,154	15,262,114	2,854,516,818
	FLASH	6,414,325	615,762	634,828	1,396,544	4,586,896	130,642	13,778,997
	PDF	234,660,778	206,428,522	203,255,141	443,445,684	295,465,883	556,128	1,383,812,136
	HTML	60,067,938	9,911,401	14,614,980	35,477,908	45,013,231	4,339,649	169,425,107
	テキスト	17,249,199	12,776,849	6,571,384	22,723,653	85,879,121	552,683	145,752,889
	スクリプト	85,909	20,850	45,358	122,633	201,375	4,217	480,342
	オフィス	1,377,947	23,126,479	19,863,236	24,987,718	28,111,980	587	97,467,947
	アーカイブ	24,231,023	16,927,511	632,530	794,227	37,358,386	203,784	80,147,461
	アプリケーション	12,804,764	2,034,864	1,120,307	786,045	40,053,523	157,866	56,957,369
	画像	84,891,668	63,683,909	54,354,517	110,298,036	268,110,661	7,627,789	588,966,580
	音声	4,201,382	992,886	2,496,220	12,201,459	14,817,118	467,765	35,176,830
	音声ストリーミング	0	0	0	512	0	0	512
	動画	15,117,163	4,896,755	6,140,026	76,849,299	116,734,838	1,005,439	220,743,520
	動画ストリーミング	57	5	17	340	7	0	426
	バイナリ	33,882,884	4,744,071	2,566,055	6,577,122	12,488,159	215,560	60,473,851
	その他	14,111	292	95,574	767,739	454,867	0	1,332,583

図表 3-2　1 組織あたりの平均コンテンツ情報発信量

		企業	国	都道府県	市町村	大学	ISP	全体
概要	対象組織数	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	対象ホスト数	2.4	6.6	11.8	1.8	8.7	44.0	3.4
ファイル数	総数	6,090	51,361	89,173	5,691	14,501	146,114	9,132
	FLASH	15	31	70	3	19	144	12
	PDF	354	7,907	13,839	680	700	133	819
	HTML（ページ数）	1,756	18,611	27,894	2,125	5,658	69,786	3,192
	テキスト	10	1,204	516	24	316	535	92
	スクリプト	23	112	350	18	66	334	33
	オフィス	2	1,636	4,528	409	63	3	257
	アーカイブ	7	13	8	0	56	44	12
	アプリケーション	2	118	145	2	55	43	14
	画像	3,892	21,553	41,360	2,409	7,475	72,305	4,654
	音声	11	33	274	8	27	326	16
	音声ストリーミング	0	0	0	0	0	0	0
	動画	7	25	66	4	21	2,425	11
	動画ストリーミング	0	0	3	1	0	1	0
	バイナリ	10	117	99	9	43	34	18
	その他	0	0	22	0	4	0	1
データ量(KB)	総量	292,726	4,875,495	6,646,600	420,336	1,367,833	3,052,423	670,074
	FLASH	3,793	8,673	13,507	797	6,609	26,128	3,235
	PDF	138,770	2,907,444	4,324,577	253,108	425,743	111,226	324,839
	HTML	35,522	139,597	310,957	20,250	64,861	867,930	39,771
	テキスト	10,201	179,956	139,817	12,970	123,745	110,537	34,214
	スクリプト	51	294	965	70	290	843	113
	オフィス	815	325,725	422,622	14,262	40,507	117	22,880
	アーカイブ	14,329	238,416	13,458	453	53,831	40,757	18,814
	アプリケーション	7,572	28,660	23,836	449	57,714	31,573	13,370
	画像	50,202	896,956	1,156,479	62,956	386,327	1,525,558	138,255
	音声	2,485	13,984	53,111	6,964	21,350	93,553	8,257
	音声ストリーミング	0	0	0	0	0	0	0
	動画	8,940	68,968	130,639	43,864	168,206	201,088	51,818
	動画ストリーミング	0	0	0	0	0	0	0
	バイナリ	20,037	66,818	54,597	3,754	17,994	43,112	14,196
	その他	8	4	2,033	438	655	0	313

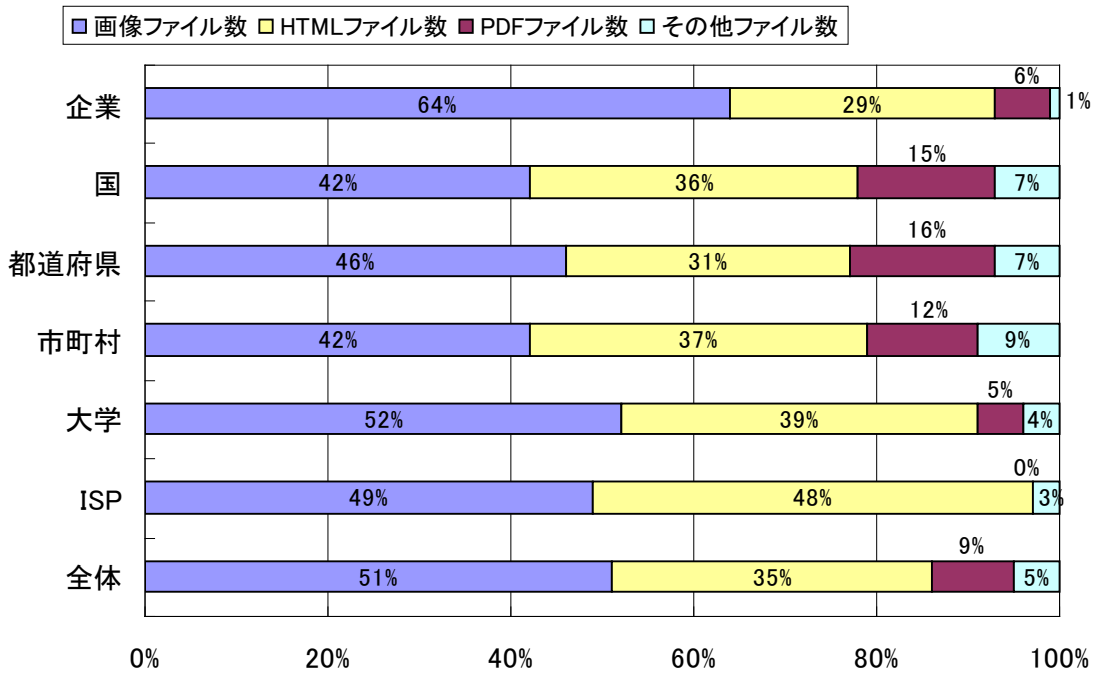
図表 3-3　1 ホストあたりの平均コンテンツ情報発信量

		企業	国	都道府県	市町村	大学	ISP	全体
概要	対象組織数	0.4	0.2	0.1	0.6	0.1	0.0	0.3
	対象ホスト数	1.0	1.0	1.0	1.0	1.0	1.0	1.0
ファイル数	総数	2,555	7,759	7,552	3,137	1,667	3,321	2,684
	FLASH	6	5	6	2	2	3	3
	PDF	149	1,194	1,172	375	80	3	241
	HTML（ページ数）	737	2,812	2,362	1,171	650	1,586	938
	テキスト	4	182	44	13	36	12	27
	スクリプト	10	17	30	10	8	8	10
	オフィス	1	247	383	226	7	0	76
	アーカイブ	3	2	1	0	6	1	4
	アプリケーション	1	18	12	1	6	1	4
	画像	1,633	3,256	3,503	1,328	859	1,643	1,368
	音声	4	5	23	4	3	7	5
	音声ストリーミング	0	0	0	0	0	0	0
	動画	3	4	6	2	2	55	3
	動画ストリーミング	0	0	0	0	0	0	0
	バイナリ	4	18	8	5	5	1	5
	その他	0	0	2	0	0	0	0
データ量(KB)	総量	122,798	736,511	562,865	231,727	157,243	69,373	196,985
	FLASH	1,591	1,310	1,144	439	760	594	951
	PDF	58,214	439,210	366,225	139,536	48,943	2,528	95,495
	HTML	14,901	21,088	26,333	11,164	7,456	19,726	11,692
	テキスト	4,279	27,185	11,840	7,150	14,225	2,512	10,058
	スクリプト	21	44	82	39	33	19	33
	オフィス	342	49,205	35,790	7,863	4,657	3	6,726
	アーカイブ	6,011	36,016	1,140	250	6,188	926	5,531
	アプリケーション	3,177	4,329	2,019	247	6,635	718	3,931
	画像	21,060	135,498	97,936	34,707	44,411	34,672	40,644
	音声	1,042	2,113	4,498	3,839	2,454	2,126	2,427
	音声ストリーミング	0	0	0	0	0	0	0
	動画	3,750	10,419	11,063	24,182	19,337	4,570	15,233
	動画ストリーミング	0	0	0	0	0	0	0
	バイナリ	8,406	10,094	4,624	2,070	2,069	980	4,173
	その他	4	1	172	242	75	0	92

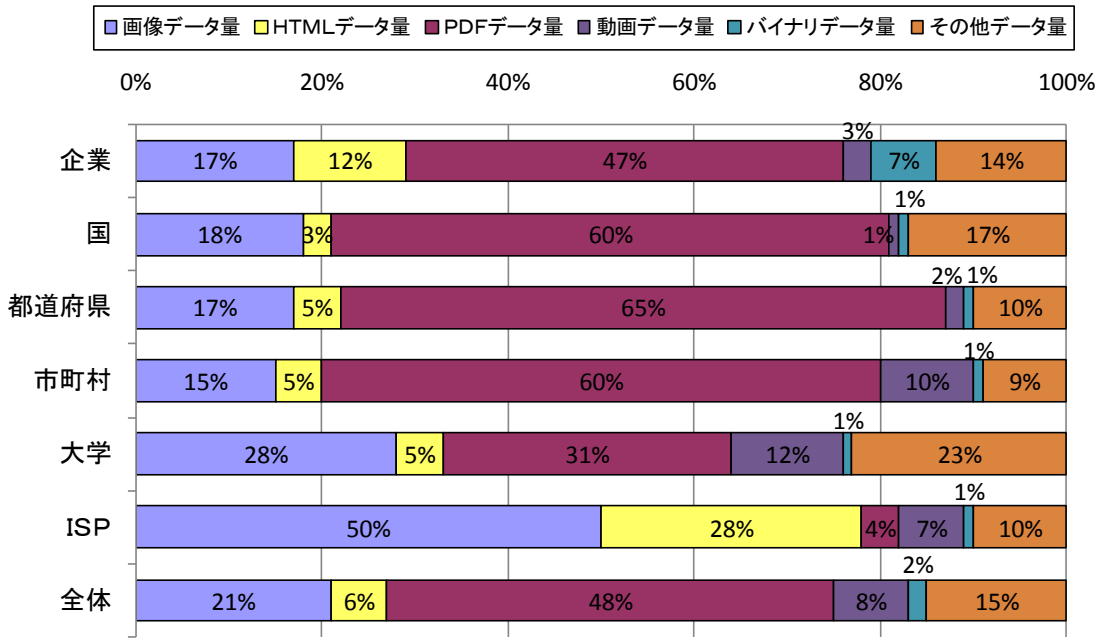
図表 3-4 発信されるコンテンツ情報のファイル種別の構成比

		企業	国	都道府県	市町村	大学	ISP	全体
ファイル数	総数	100%	100%	100%	100%	100%	100%	100%
	FLASH	0%	0%	0%	0%	0%	0%	0%
	PDF	6%	15%	16%	12%	5%	0%	9%
	HTML（ページ数）	29%	36%	31%	37%	39%	48%	35%
	テキスト	0%	2%	1%	0%	2%	0%	1%
	スクリプト	0%	0%	0%	0%	0%	0%	0%
	オフィス	0%	3%	5%	7%	0%	0%	3%
	アーカイブ	0%	0%	0%	0%	0%	0%	0%
	アプリケーション	0%	0%	0%	0%	0%	0%	0%
	画像	64%	42%	46%	42%	52%	49%	51%
	音声	0%	0%	0%	0%	0%	0%	0%
	音声ストリーミング	0%	0%	0%	0%	0%	0%	0%
	動画	0%	0%	0%	0%	0%	2%	0%
	動画ストリーミング	0%	0%	0%	0%	0%	0%	0%
	バイナリ	0%	0%	0%	0%	0%	0%	0%
	その他	0%	0%	0%	0%	0%	0%	0%
データ量(KB)	総量	100%	100%	100%	100%	100%	100%	100%
	FLASH	1%	0%	0%	0%	0%	1%	0%
	PDF	47%	60%	65%	60%	31%	4%	48%
	HTML	12%	3%	5%	5%	5%	28%	6%
	テキスト	3%	4%	2%	3%	9%	4%	5%
	スクリプト	0%	0%	0%	0%	0%	0%	0%
	オフィス	0%	7%	6%	3%	3%	0%	3%
	アーカイブ	5%	5%	0%	0%	4%	1%	3%
	アプリケーション	3%	1%	0%	0%	4%	1%	2%
	画像	17%	18%	17%	15%	28%	50%	21%
	音声	1%	0%	1%	2%	2%	3%	1%
	音声ストリーミング	0%	0%	0%	0%	0%	0%	0%
	動画	3%	1%	2%	10%	12%	7%	8%
	動画ストリーミング	0%	0%	0%	0%	0%	0%	0%
	バイナリ	7%	1%	1%	1%	1%	1%	2%
	その他	0%	0%	0%	0%	0%	0%	0%

図表 3-5　発信されるコンテンツ情報のファイル数におけるファイル種別の構成比



図表 3-6　発信されるコンテンツ情報のデータ量におけるファイル種別の構成比



図表 3-7　発信されるコンテンツ情報の 1 ファイルあたりの平均データ量（KB）

		企業	国	都道府県	市町村	大学	ISP	全体
データ量(KB)	総量	48	95	75	74	94	21	73
	FLASH	259	277	194	251	355	181	278
	PDF	391	368	312	372	609	839	397
	HTML	20	8	11	10	11	12	12
	テキスト	1,014	149	271	538	392	207	373
	スクリプト	2	3	3	4	4	3	3
	オフィス	332	199	93	35	641	45	89
	アーカイブ	2,123	18,143	1,622	910	969	918	1,530
	アプリケーション	3,062	243	165	236	1,044	734	929
	画像	13	42	28	26	52	21	30
	音声	233	429	194	864	798	287	520
	音声ストリーミング	0	0	0	26	0	0	13
	動画	1,368	2,792	1,992	12,204	8,121	83	4,534
	動画ストリーミング	0	0	0	0	0	0	0
	バイナリ	1,967	571	549	434	423	1,283	806
	その他	125	22	94	1,924	164	0	308

3-2 組織別の分析

3-2-1 企業（東証一部上場）

企業（東証一部上場）について、組織ごとに Web コンテンツのファイル総数が多い順にファイル総数、HTML ファイル数、データ総量（KB：キロバイト）、HTML データ量（KB）を図表 3-8 に示す。

ファイル総数が 10 万ファイルを超えている企業の組織は 12 組織（調査対象の企業全体の 0.7%）である。企業 1 組織でも巨大なコンテンツ量を持つ Web サイトが散見される。

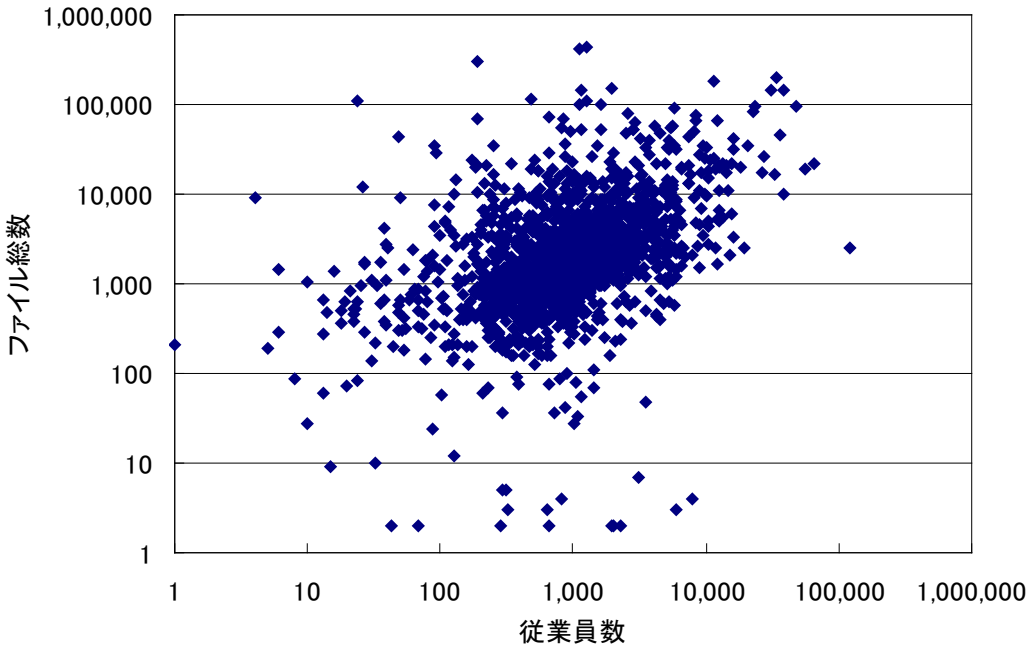
図表 3-8 ファイル総数が多い企業の組織上位 20

組織	ファイル総数	HTML ファイル数	データ総量（KB）	HTML データ量(KB)
企業1	411,625	121,199	5,182,050	589,330
企業2	302,814	151,124	5,439,806	2,707,842
企業3	203,557	93,343	30,739,777	1,272,530
企業4	179,290	55,292	14,164,083	1,082,834
企業5	153,197	84,141	5,939,551	4,418,894
企業6	142,784	36,170	7,318,699	515,728
企業7	141,695	50,784	16,396,481	989,995
企業8	141,494	29,267	1,875,693	284,635
企業9	113,992	38,201	4,099,405	2,310,170
企業10	107,919	11,628	2,898,431	127,236
企業11	107,907	28,285	14,479,837	423,631
企業12	101,822	62,743	3,071,328	1,491,682
企業13	96,957	17,533	7,789,225	404,738
企業14	96,556	26,575	10,109,070	286,860
企業15	90,520	17,100	7,749,137	240,105
企業16	81,547	13,239	4,532,808	156,341
企業17	81,140	25,376	2,161,649	456,151
企業18	76,321	54,614	2,001,508	1,560,093
企業19	71,981	17,511	921,539	125,233
企業20	69,903	8,394	3,742,267	165,152

企業のWebサイトにおけるファイル総数と従業員数との関係（散布図）を図表 3-9 に示す[21]。この散布図によるとデータの分布は全体的にまばらである。

また、ファイル総数と従業員数との相関係数を原数値で計算した結果、0.21 となった。この結果からも企業の Web サイトにおけるコンテンツ情報発信量と組織の規模との相関は弱いと考えられる。

図表 3-9 企業の Web サイトにおけるファイル総数と従業員数



企業の業種別[22]に分析を行うため、企業の業種別に平均コンテンツ量を計算した結果を図表 3-10 に示す。これによると、業種別のファイル総数では電気機器、情報・通信業、小売業、サービス業が多く、データ総量では、電気機器、情報・通信業、機械、卸売業の順となっている。1 組織当たりの平均で見ると、ファイル総数では空運業が最も多く、続いて情報・通信業、電気・ガス業、電気機器の順となっており、データ総量では、電気機器が最も多く、続いて情報・通信業、空運業、電気・ガス業の順となっている。

一方、平均ファイル総数では、鉱業が最も少なく、次いで倉庫・運輸関連業、金属製品、鉄鋼となっている。平均データ総量では、鉱業が最も少なく、次いで倉庫・運輸関連業、水産・農林業、海運業となっている。業種によって、コンテンツ情報発信量に相当の差があることが明らかになったが、多数の消費者、利用者に製品、サービスを提供している業種では、コンテンツ量が多い傾向にあることが伺える。

[21] ただし、このグラフでは従業員数のデータが得られなかったもの及びクローリングで取得できた HTML ファイル数が 1 又は 0 のものは除いている。また、従業員数（x 軸）、ファイル総数（y 軸）ともに常用対数（$\log_{10}$）でスケール化している。

[22] ここでの業種は「証券コード協議会」が定める業種別分類の中分類（33 業種）を用いた。

また、企業の業種を電子商取引（EC）が多い業種（以下「EC」とする）、電子商取引が少ない業種（以下「非EC」とする）に分類すると[23]、図表 3-11 のようになる。

1 組織当たりの平均ファイル総数を見ると、電子商取引が多い業種（銀行業、情報・通信業、証券・商品先物取引業、電気機器、倉庫・運輸関連業等）では、平均 6,695 ファイル、電子商取引が少ない業種（鉱業、医薬品、海運業、建設業、水産・農林業等）では、平均 4,700 ファイルとなっている。

電子商取引が多い業種はファイル数等のコンテンツ量が多く、Web サイトを活用した情報発信を積極的に行っている傾向が伺える。

[23] 電子商取引に関するデータについては、「平成 13 年事業所・企業統計調査」（総務省）において企業全体の電子商取引の導入率は約 10%であったため、電子商取引の導入率が 10%以上の業種を「電子商取引が多い業種」とし、10%未満の業種を「電子商取引が少ない業種」とした。

図表 3-10 企業の業種別による Web コンテンツ量

	組織数	対象ホスト数	合計				平均				標準偏差				平均順位			
			ファイル総数	HTMLファイル数	データ総量/KB	HTMLデータ量/KB	ファイル総数	HTMLファイル数	データ総量/KB	HTMLデータ量/KB	ファイル総数	HTMLファイル数	データ総量/KB	HTMLデータ量/KB	ファイル総数	HTMLファイル数	データ総量/KB	HTMLデータ量/KB
卸売業	141	294	440,476	149,535	20,739,722	5,339,468	3,124	1,061	147,090	37,869	12,940	7,057	574,294	370,462	22	15	19	8
建設業	103	195	387,023	73,657	12,965,478	897,886	3,758	715	125,878	8,717	6,732	1,348	219,629	17,509	18	21	22	23
小売業	139	293	612,676	198,592	16,565,166	4,216,658	4,408	1,429	119,174	30,336	11,672	5,342	369,920	199,708	14	8	23	9
不動産業	49	126	278,276	85,611	9,032,256	2,672,368	5,679	1,747	184,332	54,538	12,779	6,133	417,632	266,512	7	7	14	7
保険業	9	17	40,590	10,463	2,543,727	119,509	4,510	1,163	282,636	13,279	3,225	1,042	142,744	11,605	13	12	7	17
水産・農林業	6	11	14,792	2,539	381,085	40,159	2,465	423	63,514	6,693	1,452	291	23,993	5,929	27	28	31	27
鉱業	6	8	3,971	676	80,460	7,574	662	113	13,410	1,262	493	77	9,927	696	33	33	33	33
サービス業	81	226	534,045	213,175	18,992,660	6,692,334	6,593	2,632	234,477	82,621	14,947	8,562	491,197	303,707	6	5	9	4
機械	120	223	450,501	170,429	21,642,363	2,738,559	3,754	1,420	180,353	22,821	8,317	6,363	299,003	137,296	19	9	16	11
食料品	75	192	406,142	76,014	10,887,214	1,323,997	5,415	1,014	145,163	17,653	8,905	1,670	200,842	36,465	9	18	20	13
情報・通信業	88	372	1,577,861	527,785	63,887,086	8,497,851	17,930	5,998	725,990	96,566	56,187	20,759	1,523,241	355,909	2	2	2	2
繊維製品	47	86	149,334	29,406	4,297,049	378,840	3,177	626	91,427	8,060	4,516	1,054	142,519	18,542	21	25	28	25
ガラス・土石製品	29	46	95,027	19,210	3,151,814	222,217	3,277	662	108,683	7,663	4,076	686	132,198	8,137	20	23	25	26
輸送用機器	60	149	304,855	62,658	12,937,807	909,639	5,081	1,044	215,630	15,161	8,617	1,776	366,630	31,815	10	17	11	14
化学	120	248	519,988	106,277	20,506,430	1,266,184	4,333	886	170,887	10,552	9,383	2,048	421,535	22,211	15	20	18	21
金属製品	38	49	63,906	9,518	3,484,175	127,517	1,682	250	91,689	3,356	2,764	294	108,875	5,215	31	31	27	31
パルプ・紙	13	17	25,982	5,430	1,106,829	68,748	1,999	418	85,141	5,288	2,413	620	111,219	8,906	29	29	29	29
電気機器	163	589	2,418,840	731,653	181,230,565	14,894,770	14,840	4,489	1,111,844	91,379	45,715	14,754	3,418,463	363,161	4	3	1	3
医薬品	35	85	148,484	32,852	8,692,885	395,291	4,242	939	248,368	11,294	5,011	1,325	400,305	15,361	16	19	8	19
精密機器	24	57	116,676	25,267	8,874,874	344,623	4,862	1,053	369,786	14,359	6,886	1,632	613,384	21,531	11	16	6	15
石油・石炭製品	10	12	29,662	6,454	2,107,381	137,916	2,966	645	210,738	13,792	4,200	980	249,716	24,525	25	24	12	16
ゴム製品	12	22	36,673	6,591	1,345,343	112,619	3,056	549	112,112	9,385	5,448	952	156,190	21,084	23	27	24	22
鉄鋼	34	45	60,309	13,560	3,469,296	179,566	1,774	399	102,038	5,281	2,005	515	132,956	6,926	30	30	26	30
非鉄金属	23	47	88,604	25,227	4,964,702	284,187	3,852	1,097	215,857	12,356	5,146	1,918	337,304	18,298	17	14	10	18
その他製品	45	125	396,881	63,536	18,360,287	904,017	8,820	1,412	408,006	20,089	20,313	3,048	1,207,237	40,225	5	10	5	12
その他金融業	36	82	167,346	73,580	6,498,607	2,774,104	4,649	2,044	180,517	77,058	10,228	8,290	399,751	370,780	12	6	15	5
銀行業	84	119	255,638	56,615	11,976,099	724,142	3,043	674	142,573	8,621	2,982	669	181,712	8,606	24	22	21	24
証券、商品先物取引業	22	48	52,218	12,574	3,948,714	240,062	2,374	572	179,487	10,912	2,541	657	279,949	13,340	28	26	17	20
陸運業	33	85	185,968	39,812	6,327,691	753,512	5,635	1,206	191,748	22,834	5,724	1,394	314,564	38,968	8	11	13	10
倉庫・運輸関連業	15	19	15,526	3,435	295,915	46,902	1,035	229	19,728	3,127	710	202	20,523	3,628	32	32	32	32
海運業	10	10	26,837	11,528	691,601	59,819	2,684	1,153	69,160	5,982	5,364	2,597	107,046	11,450	26	13	30	28
空運業	4	68	92,103	16,484	2,499,119	244,817	23,026	4,121	624,780	61,204	25,524	4,197	555,890	55,965	1	4	3	6
電気・ガス業	17	66	300,544	109,209	10,514,810	2,452,083	17,679	6,424	618,518	144,240	19,264	12,509	563,623	360,321	3	1	4	1

図表 3-11 EC 別コンテンツ量

	組織数	対象ホスト数	合計				平均				標準偏差			
			ファイル総数	HTMLファイル数	データ総量/KB	HTMLデータ量/KB	ファイル総数	HTMLファイル数	データ総量/KB	HTMLデータ量/KB	ファイル総数	HTMLファイル数	データ総量/KB	HTMLデータ量/KB
EC	1,178	2,927	7,886,557	2,256,695	406,674,429	42,387,570	6,695	1,916	345,224	35,983	25,160	8,942	1,440,100	228,112
非EC	513	1,104	2,411,197	712,657	88,324,781	17,680,368	4,700	1,389	172,173	34,465	9,795	5,250	345,279	192,045

3-2-2 国(各省庁等)

国（各省庁等）について、組織ごとに Web コンテンツのファイル総数が多い順にファイル総数、HTML ファイル数、データ総量（KB）、HTML データ量（KB）を図表 3-12 に示す。

ファイル総数が 10 万ファイルを超えている国の組織は 8 組織(調査対象の国全体の 11%)である。

図表 3-12 ファイル総数が多い国(各省庁等)の組織上位 20

組織	ファイル総数	HTML ファイル数	データ総量 (KB)	HTML データ量(KB)
国1	1,542,898	401,024	152,672,913	2,912,849
国2	340,935	105,108	24,339,646	1,613,156
国3	300,183	110,415	49,993,325	1,072,986
国4	249,090	169,906	17,976,999	697,802
国5	204,986	160,575	5,427,499	392,882
国6	146,268	54,336	13,732,432	432,819
国7	122,794	50,350	12,994,713	687,563
国8	104,871	27,844	3,391,805	182,822
国9	88,832	68,577	4,258,151	128,672
国10	74,748	37,665	2,927,086	217,789
国11	71,339	10,107	7,867,924	256,000
国12	65,036	8,814	15,210,740	69,707
国13	29,328	7,811	4,688,086	80,219
国14	27,688	6,526	981,475	48,705
国15	24,130	1,914	5,126,225	8,571
国16	23,810	9,740	1,974,510	127,816
国17	22,122	10,341	1,470,563	105,385
国18	18,984	4,768	2,143,673	58,741
国19	18,037	4,436	930,289	49,667
国20	14,545	10,803	498,458	99,785

3-2-3 都道府県

都道府県について､組織ごとに Web コンテンツのファイル総数が多い順にファイル総数、HTML ファイル数、データ総量（KB)、HTML データ量（KB）を図表 3-13 に示す。

ファイル総数が 10 万ファイルを超えている都道府県の組織は 19 組織（調査対象の都道府県全体の 40%）である。

図表 3-13 ファイル総数が多い都道府県の組織上位 20

組織	ファイル総数	HTML ファイル数	データ総量 (KB)	HTML データ量(KB)
都道府県1	300,136	103,853	20,513,308	1,346,374
都道府県2	289,065	73,667	23,273,322	1,033,506
都道府県3	247,622	106,052	15,956,622	903,457
都道府県4	209,803	66,040	13,833,198	734,823
都道府県5	206,179	55,908	7,031,609	555,305
都道府県6	168,107	53,213	13,041,734	500,839
都道府県7	166,430	46,218	14,917,493	484,230
都道府県8	158,744	49,441	14,839,368	648,458
都道府県9	140,358	46,027	12,631,754	477,070
都道府県10	139,163	32,439	11,309,322	558,186
都道府県11	135,486	52,499	9,670,060	395,723
都道府県12	132,905	29,262	12,106,382	283,975
都道府県13	127,646	40,153	9,318,554	430,385
都道府県14	126,370	30,211	11,835,001	501,816
都道府県15	116,445	31,900	6,453,552	348,040
都道府県16	110,992	43,024	5,698,086	322,328
都道府県17	107,097	31,266	7,959,409	337,901
都道府県18	106,624	30,789	7,193,951	442,139
都道府県19	102,907	30,642	4,999,785	217,223
都道府県20	93,968	27,405	10,622,897	328,479

3-2-4 市町村

市町村について、組織ごとに Web コンテンツのファイル総数が多い順にファイル総数、HTML ファイル数、データ総量（KB）、HTML データ量（KB）を図表 3-14 に示す。

ファイル総数が 10 万ファイルを超えている市町村の組織は 5 組織（調査対象の市町村全体の 0.3%）である。

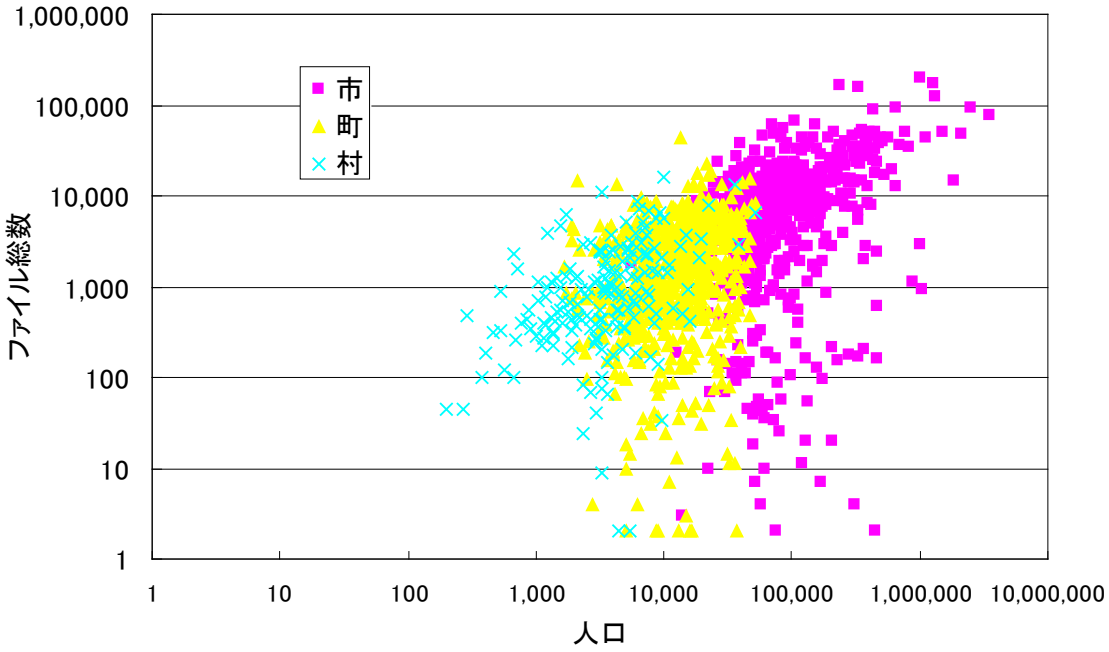
図表 3-14 ファイル総数が多い市町村の組織上位 20

組織	ファイル総数	HTML ファイル数	データ総量（KB）	HTML データ量(KB)
市町村1	195,008	81,121	28,386,082	1,106,126
市町村2	172,985	46,307	9,308,461	468,266
市町村3	166,984	113,325	4,904,559	660,975
市町村4	157,655	39,288	9,087,377	199,003
市町村5	124,732	106,873	1,733,396	925,470
市町村6	94,986	24,085	5,933,216	194,365
市町村7	93,664	31,018	5,825,672	366,565
市町村8	86,726	35,574	2,053,878	153,387
市町村9	78,841	33,984	7,254,798	355,539
市町村10	67,501	15,238	4,788,514	104,315
市町村11	61,040	19,305	2,190,761	191,426
市町村12	60,784	50,131	2,259,659	1,048,093
市町村13	54,465	12,860	3,550,208	103,043
市町村14	53,381	16,207	3,266,952	104,261
市町村15	51,241	15,679	2,855,912	137,973
市町村16	50,823	18,627	6,046,184	252,231
市町村17	49,436	29,724	3,393,922	234,108
市町村18	49,432	15,044	2,872,804	154,057
市町村19	49,181	8,654	2,078,894	56,638
市町村20	48,394	13,038	3,050,866	37,157

市町村のWebサイトにおけるコンテンツのファイル総数と人口との関係（散布図）を図表 3-15 に示す[24]。この図表によると、全体的なデータの分布として多少右上がりの傾向がみられる。市、町、村の 3 つの区分に分けてみると、市のデータはある程度の右上がりの傾向があり、町と村のデータはまばらに存在していることが伺える。

ファイル総数と人口との相関係数を原数値で計算した結果、0.56 となり、弱い正の相関があると考えられる。

図表 3-15　市町村の Web サイトにおけるファイル総数と人口



[24] ただし、このグラフでは人口のデータが得られなかったもの及びクローリングで取得できた HTML ファイル数が 1 又は 0 のものは除いている。また、従業員数（x 軸）、ファイル総数（y 軸）ともに常用対数（$\log_{10}$）でスケール化している。

3-2-5 大学

大学について、組織ごとに Web コンテンツのファイル総数が多い順にファイル総数、HTML ファイル数、データ総量（KB）、HTML データ量（KB）を図表 3-16 に示す。

ファイル総数が 10 万ファイルを超えている大学の組織は 23 組織（調査対象の大学全体の 3%）である。大学 1 組織でも巨大なコンテンツ量を持つ Web サイトが散見される。

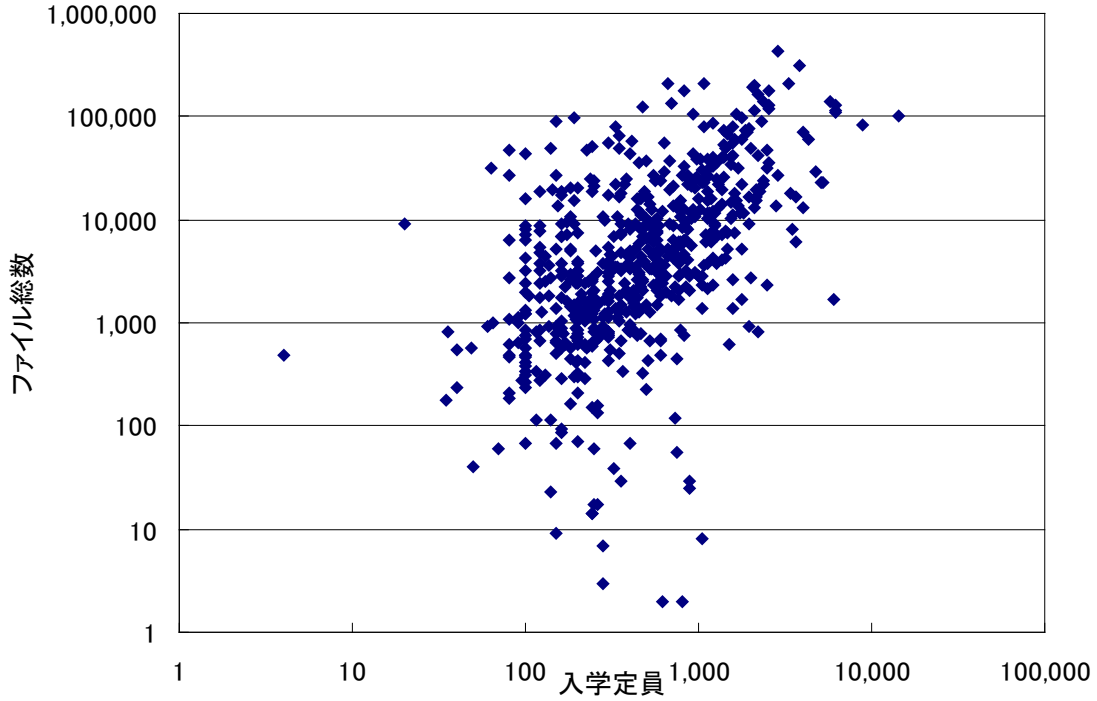
図表 3-16 ファイル総数が多い大学の組織上位 20

組織	ファイル総数	HTML ファイル数	データ総量（KB）	HTML データ量(KB)
大学1	438,724	110,252	62,950,580	974,758
大学2	306,820	114,596	60,115,726	953,351
大学3	210,678	71,064	11,317,037	251,114
大学4	205,074	201,937	2,092,222	1,729,365
大学5	204,979	63,177	24,834,421	708,695
大学6	200,521	77,133	20,259,600	675,514
大学7	195,992	64,570	21,527,517	390,928
大学8	178,376	49,844	13,209,353	479,532
大学9	176,350	88,166	10,984,579	293,230
大学10	166,788	119,590	6,390,597	812,557
大学11	141,817	44,560	16,040,806	357,212
大学12	139,793	49,109	17,975,228	318,549
大学13	136,724	76,074	9,784,020	285,636
大学14	130,935	110,573	4,429,177	2,641,945
大学15	130,514	42,634	41,041,141	325,579
大学16	123,801	59,417	6,391,254	89,906
大学17	121,218	39,662	13,406,273	270,425
大学18	116,547	77,471	4,455,691	928,593
大学19	114,379	32,423	9,150,279	255,942
大学20	110,098	41,453	11,911,619	433,032

大学のWebサイトにおけるコンテンツのファイル総数と入学定員との関係（散布図）を図表 3-17 に示す[25]。この散布図によると、全体的に右上がりのデータ分布となっている。大学においても、Webサイトにおけるファイル総数が 1000 ファイルに満たない組織が相当数あることが分かる。

ファイル総数と入学定員との相関係数を原数値で計算した結果、0.45 となり、弱い正の相関があると考えられる。

図表 3-17 大学の Web サイトにおけるファイル総数と入学定員



[25] ただし、このグラフでは入学定員のデータが得られなかったもの及びクローリングで取得できた HTML ファイル数が 1 又は 0 のものは除いている。また、従業員数（x 軸）、ファイル総数（y 軸）ともに常用対数（$\log_{10}$）でスケール化している。

4 まとめ

(1) クローリング調査（平成 18 年 8 月～10 月）における調査対象組織全体（企業、国、都道府県、市町村、大学、ISP 系ポータルサイト）のデータ総量は約 2.7TB。

(2) 調査対象組織全体で発信されるコンテンツ情報の構成は、全体のファイル数をファイル種別に見みると、画像ファイルが最も多く、続いて HTML ファイル、PDF ファイルの順となっており、全体のデータ量をファイル種別に見みると、PDF ファイルが最も多く、続いて画像ファイル、動画ファイル、HTML ファイルの順となっている。

(3) 1 組織あたりの平均コンテンツ情報発信量は ISP 系ポータルサイトが最も多く、続いて都道府県、国の順となっており、構成ファイルを見てみると、どの組織においても画像ファイルが占める割合が最も大きく、続いて HTML ファイルが大きい。国、都道府県及び市町村においては、他の種の組織に比べ、PDF ファイル、オフィスファイルの占める割合が大きく、ISP 系ポータルサイトにおいては、他の種の組織に比べ動画ファイルの占める割合が大きい。データ量について、総量を組織種別に比較すると、都道府県が最も大きく、続いて国、ISP 系ポータルサイトの順に大きい。

(4) 組織毎の Web サイトにおけるコンテンツ量は、ファイル数で見ると 10 万ファイルを超える組織がある一方、1000 ファイル以下の組織も相当数ある。極端にファイル数が小さい組織については FLASH の利用等によってクローリング調査で収集ができないファイルが存在することも考えられるが、全体的には組織毎にコンテンツ量に相当の差があることが伺える。

(5) 組織毎のコンテンツ量については、市町村や大学では組織の規模指標（人口、大学の入学定員）とファイル数との間に一定の相関関係が認められ、組織の規模が大きくなればなる程、組織毎のコンテンツ量が大きくなるようなゆるやかな傾向が見られる。

(6) 企業の場合は組織の規模指標（従業員数）とファイル数の強い相関関係は認められなかったが、業種別に比較分析したところ、業種によってコンテンツ量に相当の差があることが分かった。また、業種を電子商取引が多い業種と電子商取引が少ない業種に分類すると、電子商取引の多い業種の方が一組織当たりの平均コンテンツ量が大きいことが分かった。

参考 Webサイトのコンテンツの概況

インターネットの普及が始まった当初は、Web サイトのコンテンツは HTML ファイルを基本として 1 つの Web サイトに 1 つのドメイン、1 つのホスト（Web サーバ）という単純なコンテンツの構造であったが、平成 10 年からのコンテンツ量調査を実施してきた結果、Web サイトは以下の(1)～(9)のような変化がうかがえる。

(1) 規模の拡大

平成 10 年の調査を開始した時点から平成 16 年まで、サイト数、ページ数、データ量は年々増加しており、インターネット空間におけるコンテンツ量は日々絶え間なく増加していることが伺える。

(2) ファイル形式の多様化

インターネットの普及が始まった当初、Web コンテンツは HTML ファイルが基本であり、それに画像ファイルを加えて、Web ページが構成されるのが一般的であったが、年々、そのファイル形式は多様化が進んできた。

具体的には、まず文書として PDF ファイルが広く利用されるようになってきている。PDF ファイルは、文書であるという点で HTML ファイルと似ているが、HTML ファイルが持つリンク情報を持っていない。

また、FLASH ファイルが Web サイトに高い表現力を持たせる手法として普及してきている。

その他、音声ファイルや動画ファイルは、一般的なパソコンで音声や動画の再生が容易になってきており、Web コンテンツとして広く利用されているようである。また、これらのマルチメディアコンテンツは、ダウンロード形式の保存形式でなくストリーミング形式で提供されることも多く、専用のソフトウェアが必要となることも多い。

最近利用が増えているRSS[26]のような、人間がブラウザでWebページとして見ることを想定していないファイルも増加している。

(3) コンテンツ量の単位の変化

Webサイトのコンテンツの単位として、ページ（HTMLファイル）が基本であることは変わりないが、古くから存在しているフレーム[27]に加えて、最近では電子掲示板（BBS）やブログのように、ページではなく記事の単位でのコンテンツも増加している。

[26] RSS とは、Rich Site Summary あるいは RDF Site Summery の略で、Web サイトの見出しや要約等のメタデータ（文書に記載されている内容そのものではなく、その構造や属性に関する情報）を構造化して記述する XML ベースのフォーマットである。

[27] フレームとは、Web ブラウザの一つのウィンドウをいくつかに区切ってそれぞれに別々の内容を表示させる Web ページの表現技法の一つ。

(4)　端末の多様化

Web の発展とともにクライアントプログラムや端末の多様化も進んでいる。従来はパソコンによる Web サイトの利用がほとんどであったが、近年は携帯電話による利用も多くなっている。パソコンと携帯電話では、その性能や画面サイズ、利用シーンの違いから異なるコンテンツが用意されていることが多い。

また、個々の端末の違いに対応するために、端末の種類によって異なるサイトを返すサーバがある。さらに、クライアント側のブラウザから送信した言語指定によって異なるサイトを返すサーバも存在する。

(5)　動的生成によるコンテンツの増加

情報提供の多様化、管理負荷の低減等により、動的に生成されるコンテンツが増加している。動的生成による Web サイトの例としては、電子掲示板（BBS）、ブログ、通販サイト、ニュースサイト、検索サイト等がある。

JavaScript のようなクライアント側で実行されるプログラムだけでなく、サーバ側で実行されるものもある。サーバ側で動的生成されているものは、クライアント側で識別できない場合も多い。

動的生成の技術は、従来から広く利用されているCGI[28]だけでなく、最近ではPHP[29]等も広く利用されている。また、XOOPS[30]のように動的生成サイトを用意に構築できる環境が整ってきている。

(6)　有料コンテンツ・制限コンテンツの増加

Web サイトのビジネス利用やコンテンツビジネスが増加するに伴い、認証が必要なコンテンツが増加している。このコンテンツは、ダウンロードしたファイルの利用に認証が必要なものと、そもそもファイルへのアクセスに認証が必要なものがある。

(7)　ドメインの多様化

かつては我が国のドメインの利用は属性JPドメインが中心であり、1 つの組織が 1 つまでしか利用できなかった。しかし、汎用JPドメイン等の普及が進むにつれ、ドメインの利用が容易かつ安価になってきている。そのため、主に企業では製品毎のドメインの取得や短期的なドメインの利用が増えてきている。また、ドメインの変更も容易になってきている。他方で、多国籍企業やグループ企業では 1 つのドメインに集約する例もある[31]。

---

[28] CGI（Common Gateway Interface）とは、Web ブラウザ（クライアント）から要求を受けた Web サーバが、それに対応するプログラムを起動し、実行結果を Web ブラウザに返すための仕様（環境変数の名前や値等）を定めるものである。

[29] PHP（PHP:Hypertext Preprocessor）とは、動的な Web ページを作成しやすいように開発されたプログラミング言語、またはその言語で記述されたプログラムのことである。

[30] XOOPS（eXtensible Object Oriented Portal System：ズープス）とは、PHP 言語で書かれた Web サイト用 CMS（コンテントマネジメントシステム）ソフトウェア。

[31] 例えば、abc.jp から abc.com/jp へと変更する企業もある。

(8)　ホストと IP アドレスの関係の複雑化

ドメイン利用の多様化や動的生成の増加等が原因で、ホストと IP アドレスの関係が複雑化している。従来は 1 対 1 の関係で対応していたが、負荷分散のため 1 つの IP アドレスに対して 1 つのホストではなく複数のホストを対応させている例がある。逆に、多用途の利用やホスティングのため、1 つのホストに複数の IP アドレスが対応していることもある。

(9)　外部のファイルの増加

1 つの Web ページを構成するファイルは、以前は 1 つのホストから返されるのが一般的であったが、最近では負荷分散等の目的から画像等の特定ファイルは別ホストから提供している例が多い。また、広告や認証ロゴ等を他のサイトからリンクさせ、Web サイトの一部としている利用が多い。