

日本の Web サイトの網羅的収集、蓄積及び保存
に関する調査報告

概要



1. はじめに

将来、日本における Web データを収集、蓄積及び保存するために、その実施可能性や方法に関して検討することを目的として、平成 16 年 10 月から平成 17 年 3 月まで日本における Web データの調査を行った。この結果、平成 17 年 3 月の時点で日本における Web データ総量は 18.4 TB、ファイル総数は 4 億 5000 万ファイルであると推定された。この結果と併せて Web アーカイブの要件について検討した結果の概要を述べる。

2. 調査概要

2.1. クロール調査

現在、日本における Web データがどの程度あり、そのうち保存管理に用いることのできる記述メタデータ（タイトル、著者名、キーワード等）がどの程度設定されているかを把握するためにクローラを用いた調査を行った。この調査の仕様と要件を表 1 に示す。

表 1 調査仕様・要件

項番	分類	項目	仕様・要件
1	調査仕様	調査日数	約 60 日
2		平均ファイル収集速度	600 万ファイル/日
3		調査対象 Web サーバ	JP ドメイン及び JPNIC 管轄下の IP アドレスをもつ Web サーバ
4		ロボット排除指定	robots.txt 及び ロボット排除用 META タグ 遵守
5		調査ファイル種別	全て
6		リンク抽出対象	HTML, JavaScript, PDF, カスケーディングスタイルシート等。 ただし、アクセスのための認証が必要な リンクや問合せ FORM は対象外。
7	クローラ仕様	コンテンツ管理データベース	Oracle10g、PC サーバ 4 台のクラスタ 3.5 億レコード
8		インターネット接続回線	60Mbps (最大 100Mbps)
9		クローラ動作方式	複数スレッド/プロセスによる分散処理
10		クローラ台数	PC サーバ 13 台

2.2. クロール調査結果

2.2.1. 調査データ量と推定データ総量

調査に関する問合せ対応等に要した期間を除いた実質クロール調査日数は 23 日であり、この期間での調査 Web データ量は 4.9TB、調査ファイル数は 1 億 2000 万ファイルとなった。既読 URL 発見率の推移から日本における Web データ総量とファイル総数を推定するとそれぞれ 18.4TB、4 億 5000 万ファイルとなった。これを表 2 に示す。

表 2 調査 Web データ量と推定 Web データ総量

項目	量
調査 Web データ量	4.9TB
調査ファイル数	1 億 2000 万ファイル
推定 Web データ総量	18.4 TB
推定ファイル総数	4 億 5000 万ファイル

2.2.2. ファイルタイプ別調査 Web データ量とファイル数

本調査で得られた Web データ量とファイル数をファイルタイプ別に分けて表 3に示す。

表 3 ファイルタイプ別調査データ量

種別	データ量 (GB)	ファイル数	平均ファイルサイズ (KB)
静的 HTML	332.3	29,898,744	11.1
動的 HTML	182.5	13,812,351	13.2
画像	1,105.9	55,128,641	20.1
文書	981.5	3,186,376	308.0
動画	888.4	333,751	2,661.9
音声	114.1	178,399	639.6
データ	859.0	590,268	1,455.2
ストリーミングメタファイル	1.5	176,254	8.8
その他	438.2	18,698,988	23.4
合計	4,903.5	122,003,772	40.2

2.2.3. ドメイン別ホスト数

本調査において対象とした URL のホスト名部分からドメイン別にユニークホスト数を集計した結果を図 1に示す。jp ドメイン全体では約 18 万ホストで、このうち co.jp ドメインの Web ホストが約 3 分の 1 を占める。jp ドメイン以外では約 13 万ホストで、このうち.com ドメインの Web ホストが半分以上を占める。

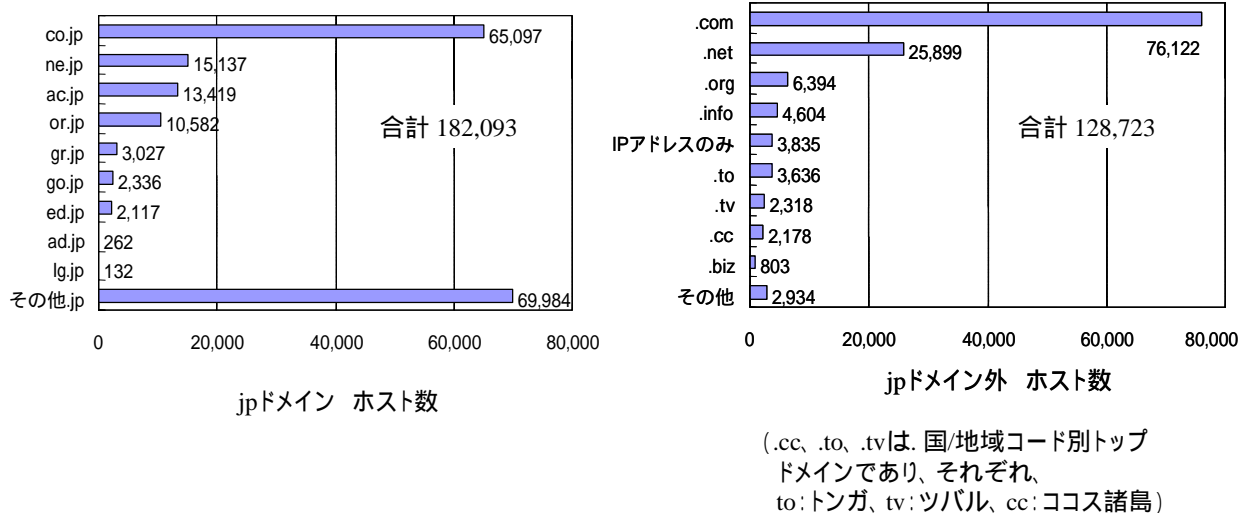


図 1 ドメイン別 Web ホスト数

2.2.4. ホスト内 Web データ量とファイル数

本調査において各ホスト内の Web データ量とファイル数についても集計を行った。その結果を表 4と表 5に示す。これらの表から 96%の Web ホストはファイルを 2,000 以下しか保有していないことが分かる。また、少数であるが膨大な量の Web データを持つ Web サイト(最高で 102GB)と膨大な数のファイルをもつ Web ホスト(最高で 72,773 ファイル)が存在することが確認できた。

表 4 ホスト内 Web データ量

Web データ量(MB)	ホスト数	頻度	累積頻度
20	282,298	0.9082	0.9082
40	11,195	0.0360	0.9443
60	4,804	0.0155	0.9597
80	2,754	0.0089	0.9686
100	1,811	0.0058	0.9744
120	1,294	0.0042	0.9786
140	945	0.0030	0.9816
160	735	0.0024	0.9840
180	609	0.0020	0.9859
200	460	0.0015	0.9874
220 以上	3911	0.0126	1.0000
合計	310,816	1.0000	---

表 5 ホスト内ファイル数

ファイル数	ホスト数	頻度	累積頻度
1,000	284,028	0.9138	0.9138
2,000	14,556	0.0468	0.9606
3,000	4,880	0.0157	0.9763
4,000	2,274	0.0073	0.9837
5,000	1,250	0.0040	0.9877
6,000	875	0.0028	0.9905
7,000	582	0.0019	0.9924
8,000	423	0.0014	0.9937
9,000	291	0.0009	0.9947
10,000	232	0.0007	0.9954
11,000 以上	1425	0.0046	1.0000
合計	310,816	1.0000	---

2.2.5. メタデータ設定数とその割合

タイトル等メタデータについて集計を行った結果を表 6に示す。タイトルについては 95%の HTML ファイルで設定されていたものの、著者については 5%、概要については 14%、キーワードについては 14%の割合でしか設定されていないことが分かった。さらに、ダブリンコアの設定 (<meta_dc.title>、<link_dc.creator>の様な形式で設定されているもの) については非常に少数 (0.3%未満) であった。このように記述メタデータの設定割合は低く、こういったメタデータだけで分類を行うのは不十分であることが分かる。

表 6 メタデータの設定数とその割合 (一部抜粋)

No.	分類	項目	設定 ファイル数 (千)	設定 割合	設定長 合計 (MB)	最大長 (Byte)	平均長 (Byte)
1	タイトル	<TITLE>タグ, あるいは<META name="title">で設定	47,717	95.03%	805	5,073	17.7
2	著者	<META name="author">で設定	2,523	5.02%	33	592	13.9
3	概要	<META name="description"> で設定	7,058	14.06%	389	4,958	57.8
4	キーワード	<META name="keywords">で設定	7,064	14.07%	531	5,537	78.9
5	使用言語	<HTML lang="XXX">で設定	8,991	17.91%	19	95	2.2
6	文字コード	<META>タグ, あるいは HTTP ヘッダで文字コード設定	39,308	78.28%	---	---	---
7	サイト	RDF ファイル数	13	0.03%	---	---	---
8	記述	RSS ファイル数	14	0.03%	---	---	---
9	クッキー	Set-Cookie ヘッダ	11,480	22.86%	762	3,860	69.6

3. Web アーカイブに向けた要件

本調査で明確になった Web アーカイブの要件について、「収集」、「蓄積・保存」、「閲覧」の3つの機能に分けて概要を示す。

3.1. Web データの収集

収集において、事前公告などの事前準備と、収集条件の設定や収集機能に関してポイントとなる事項を以下にまとめる。

3.1.1 事前公告と問合せ等の受付対応

収集方法や条件などを記載した説明ページの設置と、FAQ のメンテナンス

特に収集拒否（ロボット排除指定）の方法を明示することが重要である。また公開時の利用制限種別の指定方法も、それまでに決めて公告する必要がある。

優先収集する Web サイトの確定と協力依頼

優先収集する Web サイト（やドメイン）を確定し、当該 Web サイトのファイル数とデータ量を事前確認しておく。あらかじめ定めた収集期間内に収集完了できる条件を算出し、必要があれば、当該 Web サイト管理者に高頻度収集等について事前許諾（もしくはファイルの媒体提供協力）を得ることが必要である。

一般サイトの収集条件（リクエスト間隔や最大転送速度など）の決定

被収集サイトに過剰な負荷を与えないように収集する必要がある。本調査中に寄せられたサイト管理者等からの問合せやコメントから判断すれば、リクエスト間隔 30 秒で転送速度最大 1Mbps までとしファイルサイズを最大 60MB までとした、本調査での収集条件を一応の目安とすることができる。

インターネットサービスプロバイダ等への協力依頼

人権侵害などの理由によりプロバイダの判断で編集・削除した場合に、Web アーカイブでも収集ファイルに対して同じ対応ができるように、編集・削除の内容を通知してもらう必要がある。

起点 URL の収集

収集前に、クロールの起点となる URL を十分に収集しておく。理想的には、収集対象となる URL を継続して集め、効率的に収集できるよう精査選別しておくことが望ましい。

3.1.2 収集範囲と収集データ量

収集範囲

本調査の範囲と同じとする。しかしながら、継続して、利用されるプロトコルやデータ形式の普及度合いに応じて随時見直していく。

推定収集データ量

平成 18 年度で約 5.4 億ファイル、22 TB となる。これまでの推移から、今後数年は年率 20～30%程度で増加するものと推定されるが、毎年見直しが必要である。

3.1.3 収集条件

リクエスト間隔と転送速度

事前許諾を得ていないサイトでは連続アクセス時には 30 秒を空けるものとし、また、最大の転送速度は 1Mbps までとする。この条件で収集期間内に収集し終わらないサイトは、途中までの収集となる。収集できるファイル数は、サイト内のファイルサイズ分布によるが、収集期間が 60 日であれば、1 サイトあたり平均して 6 万ファイルとすることが見込まれる。

サイト毎の収集条件設定

優先収集するサイトでの高頻度高速転送に対応するため、あるいは、収集日時を指定されるサイトに対応するために、サイト毎に収集条件を設定できる必要がある。

ロボット排除指定の遵守

WARP (ユーザエージェント名: ndl-japan-warp-0.1) 用など、国会図書館が用いている全てのクローラのユーザエージェント名をサイト管理者が認識できるようにする。また、google や Yahoo! など代表的な検索サービス事業者のクローラで独自拡張しているロボット排除指定方法にも対応しておく。

3.1.4 収集性能

一般のサイト

1 年間に 60 日程度の収集期間とすると、平成 18 年度は平均して約 1,000 万 URL / 日の収集性能が必要であり、その後は、少なくとも年率 20 ~ 30% 程度で性能向上が可能な拡張性を有していることが必要である。

優先サイト

ファイル数の大きな Web サイトでは、リクエスト間隔を短くする必要がある。例えば 30 万ファイルを有するサイトを 60 日程度で収集するためには、10 秒程度の間隔とする必要がある。さらに、リンカー貫性を高く保つ必要のあるサイトでは、1 秒に 1 回もしくはそれ以上でリクエストできるようにする。

3.1.5 収集機能

リンク抽出

特に、JavaScript のようにブラウザで動的に生成されるリンクの抽出精度がポイントとなるが、技術的に精度を高めるのが難しく今後も検討が必要である。また、この中には、ユーザが入力したデータによって URL が生成されるものがあり、これらに対しては機械的に URL を生成することはできない。

安全性の確保

ページを書き換えてしまうようなリンクや、共有カレンダーなどでほぼ無限に自動生成されるリンク、あるいは悪意をもったサイト攻撃スクリプトへのリンクなど、危険なリンクにアクセスしないようにする。このためには、リンク情報だけでなくコンテンツ内容の解析も合わせることで安全性を向上するなど、継続して技術開発を進める必要がある。

3.1.6 ユーザによる収集状況の確認手段

収集から公開まで期間が空く場合には、ユーザが URL を指定すると収集されたかどうかの確

認ができるような機能を持つことが必要である。

3.2. Web データの蓄積・保存

蓄積・保存に関して、保存すべき情報と形式、および消去申出対応についての概要を述べる。

3.2.1 保存情報

コンテンツに関する情報

クロウラの IP アドレスやリクエスト情報などリクエストした時の条件によって、収集されるファイルが変わる。このため、収集されるファイルだけでなく TCP/IP レイヤ以上の送受信データと収集日時を含めて保存しておく必要がある。

収集除外理由

ページ内に記述されている META タグで排除されている場合には、そのページそのものが残せないため、この META タグ情報が残せない。文書等で収集拒否を受け付けた場合も含め、これらの情報を別途保存しておく。

消去 / 編集履歴

収集後に消去あるいは編集した履歴情報も保存しておく。

利用制限情報

閲覧に関して、ファイルの著作権者等から利用制限の条件が指定された場合には、この情報を保存しておく。

3.2.2 保存形式

オープンな規格にもとづくことが望ましいため、ARC 形式をベースに拡張するのがよい。しかしながら、既存のツール類（オープンソースの閲覧ソフトなど）が使えなくなる可能性もあり、決定には、収集範囲等とあわせて十分な検討が必要である。

WARP や The Internet Archive など類似のシステムでの形式を調査し、これらを参考に詳細化を図る必要がある。

3.2.3 消去・編集の申出対応

受付方法

できる限り Web の Form 受付とするなど自動化を図るようにする。本人性の確認はロボット排除設定がなされているかどうかを確認することで行う。消去の申出をしたい人が、該当ファイルに対するロボット排除の設定ができる場合は、正当な申出であるとみなすことができる。したがって、Web の Form 受付で申出をしてもらった際に、このロボット排除設定をしてもらうように促し、この設定が確認できれば消去する、といった自動化を行う方法が考えられる。

消去・編集の単位

ARC 形式で保存する場合は、消去・編集はファイル単位で行うこととなる。しかしながら、文書ファイルなどで編集が不可能な設定がされている場合は、編集できない。

3.3. Web データの閲覧

閲覧について、Web アーカイブに固有の重要な閲覧機能等について概要を述べる。

3.3.1 閲覧機能

リンク書き換え機能

収集したファイルの中の全てのリンクを、Web アーカイブサイト内のアドレスを指すように書き換える必要がある。このためには、全てのリンクを抽出して書き換える。リンクを抽出する機能に関する問題点はクローラと同じである。リンクを書き換える機能に関しては、PDF など文書ファイルでは、文書ファイルのレイアウトを壊してしまったり、編集が出来ないような設定がされていたりする場合があるため、対象外となる。

アクセス制御・ファイル提供制限機能

館内のみ閲覧を許可する場合や、閲覧は一切行わず保存のみ行う場合など、利用制限をファイル単位で指定できるようにしておく。また、ファイル種別によって提供を制限することも想定されるため、ファイル種別とドメイン（ホスト）でも条件指定が可能となるようにしておく。

ナビゲーション機能

通常の Web サイトでのナビゲーションに関する要件に加えて、アーカイブでは、閲覧者がリンクを辿っている間に、勝手に別の時代のドキュメントに行かないようにすることが必要である。しかしながら、ブラウザで生成される JavaScript などでのリンクは、Web アーカイブシステムで制御することができないために閲覧時点の URL となってしまう。このため、気がつかない内に、アーカイブの外に出てしまうという問題が生じる。この解決のためには、さらに検討を要する。

3.3.2 検索機能

収集日時と URL を指定して閲覧に供するのが基本的な機能となる。タイトル、著者、キーワードなどのメタデータによる検索は、これらが設定されている HTML でのみ有効であるが、タイトル以外の設定率は 10%程度と低いいため、検索の網羅性は低くなる。このため、全文検索機能の提供や、高度なコンテンツ解析を適用した自動分類支援機能を使った分類情報の提供を検討していく必要がある。

3.3.3 利用制限の申出対応

「3.2.3 消去・編集の申出対応」と同じく、できるだけ自動化を検討する。

3.3.4 他システムとの連携インタフェース

ポータルサイトや他のデジタルアーカイブシステムとの連携や、ユーザのコンピュータからの自動アクセス手段を提供することで、様々な利用価値が生じる。Web サービスで用いられる標準的なインタフェースや、OAI-PMH のようなメタデータハーベストプロトコルによるインタフェースを整備していくことが望まれる。