平成 26 年 10 月
October 2014

国立国会図書館
National Diet Library

本書は、国立国会図書館インターネット資料収集保存事業（http://warp.da.ndl.go.jp/）で公開している「ウェブアーカイブのしくみ」の PDF 版です。PDF 版の作成にあたり、一部に修正を加えています。以下の html 版も合わせてご参照ください。
http://warp.da.ndl.go.jp/contents/reccommend/mechanism/index.html

＜目次＞

第 1 章 ウェブアーカイブのしくみ

第 2 章 ウェブアーカイブをささえる技術

第 1 章 ウェブアーカイブのしくみ

1. ウェブアーカイブとは

皆さんは、ウェブサイトの情報が昨日までとは違う内容になっていたり突然に無くなったりして、探している情報が見つからないという経験をしたことはないでしょうか？ウェブサイトに掲載されている情報は更新や削除が頻繁に行われるため、そのようなことが度々起こります。

例えば、首相官邸のウェブサイト（http://www.kantei.go.jp/）は、内閣総理大臣が交代すると内容が大きく更新されます。過去に掲載されていた情報も残されていますが、公開当時そのままの形では残っていません。また、2002 年に日本と韓国で行われた FIFA ワールドカップの日本組織委員会ウェブサイトは、大会が終了した後にインターネット上から消えてしまいました。

それでは、このように消えて行くウェブサイトの情報に再びアクセスするためには、どうすればよいでしょうか？その解決法の一つが、ウェブサイトの情報が消えて無くなる前に、それらを集めて保存しておくことです。このようにウェブサイトを収集して保存することをウェブアーカイブ（Web Archive）と言います。

ウェブアーカイブは、世界各国の国立図書館や公的機関が中心となって行っており、日本では国立国会図書館が 2002 年よりインターネット資料収集保存事業 （WARP）[1]を実施しています。上で紹介した首相官邸や 2002 年 FIFA ワールドカップ日本組織委員会のウェブサイトも WARP で保存していますので、インターネット上から消えてしまった情報も見ることができます。

図 1：首相官邸ウェブサイト（小泉内閣）
保存日:2004 年 11 月 19 日
http://warp.da.ndl.go.jp/info:ndljp/pid/234460/www.kantei.go.jp/

[1] http://warp.da.ndl.go.jp/

図 2：2002 年 FIFA ワールドカップ日本組織委員会
保存日：2002 年 10 月 28 日
http://warp.da.ndl.go.jp/info:ndljp/pid/236044/www.jawoc.or.jp/

私たちは書籍や文書などに残された情報から過去の出来事を知ることができます。ところが、インターネットとデジタル技術の発展に伴い、これまでは紙に残されてきた情報が大量かつ急速にウェブサイトなどの電子情報に置き換わっています。後の世代の人々が過去を振り返ろうとした時、ウェブサイトの情報が残されていなければ、歴史の一部が大きく欠けることになるでしょう。そのようなことが無いよう、ウェブサイトにある情報をしっかりと保存して後の世代に伝えて行く必要があります。ウェブアーカイブは短中期的にウェブ情報へのアクセスを保障するだけではなく、歴史資料を未来に残していくという長期的な意義を持った事業でもあるのです。

ウェブアーカイブを行うためには専門的な知識や技術が必要となります。本書は、皆さんにウェブアーカイブについての理解を深めていただくために、ウェブアーカイブのしくみや、ウェブアーカイブをささえる技術などに焦点をあてて、わかりやすく解説します。

2. ウェブアーカイブのライフサイクル

ウェブアーカイブのライフサイクルは、「選定」、「収集」、「組織化」、「保存」、「公開」の 5 つの部分からなります（図 3）。ウェブサイトに掲載されている情報は時間の経過とともに変化していきます。ウェブアーカイブでは、このサイクルを定期的に繰り返しながらウェブサイトの変化を記録していきます。

（1）選定

対象となるウェブサイトを選定します。特定の主題にターゲットを絞ったものから、一国全体のウェブサイトを対象とするもの、世界中のウェブサイトを包括的に集めるものまで、その目的や実施機関の種類、規模によって様々です。大きく分けると選択収集とバルク収集の 2 種類があり、両者を組み合わせて行っているウェブアーカイブもあります。

(i) 選択収集

特定のウェブサイトにターゲットを絞って収集することを「選択収集（Selective Harvesting)」といいます。サイト単位やページ単位などの収集単位も指定します。小～中規模のウェブアーカイブの場合や、以下に紹介する「バルク収集」のための法律制度が無い場合などに採用される収集方法です。ウェブサイトにも著作権があるため、法律により著作権が制限されていない場合には、事前に発信者の許諾を得てから行う必要があります。

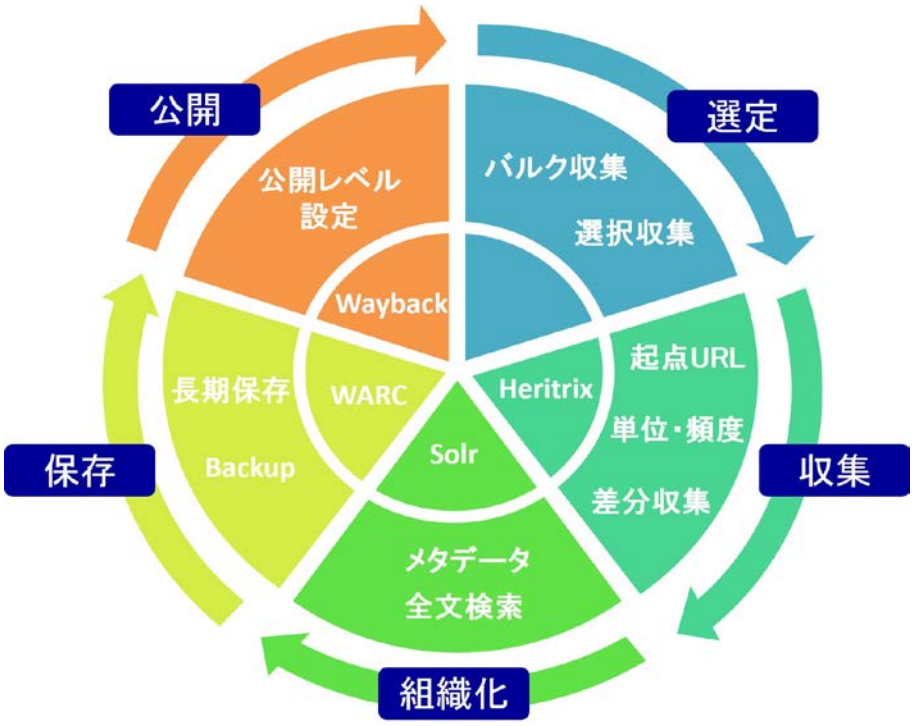


図 3：ウェブアーカイブのライフサイクル

(ii) バルク収集

「バルク収集（Bulk Harvesting)」とは、「.fr」や「.de」などの国別ドメイン全体を対象にウェブサイトを大規模に収集することです。なかには世界全体のウェブサイトを収集対象とするインターネットアーカイブ[2]のような機関もあります。

一国全体を対象とするバルク収集の多くは、国立図書館などの公的機関が法律制度に基づいて行っています。法律によってウェブサイトの著作権を制限しているため、事前に発信者の許諾を得る必要はありません。国立国会図書館も 2010 年 4 月に施行された改正国立国会図書館法に基づいて、公的機関のウェブサイトを発信者の許諾を得ること無く収集を行っています。このように法律制度に基づいて行う収集は「制度収集」とも呼ばれます。

(2) 収集

対象となるウェブサイトを実際に収集します。収集ロボット（クローラ）と呼ばれる自動収集プログラムを用いて収集します。収集する頻度や収集する深さなども指定します。

▶ 第 1 章 3～6、10（p.8～21、p.33）
▶ 第 2 章 1（p.36）

(3) 組織化

集めたウェブサイトに対してタイトルや公開者などの情報を付与します。これらの情報はメタデータと呼ばれます。また、全文検索サービスを提供する場合にはインデクス処理を行います。

▶ 第 1 章 7（p.22）
▶ 第 2 章 2（p.39）

(4) 保存

収集したウェブサイトを電子書庫（ストレージ）に保存します。長期にわたって利用を保障できるように、ウェブアーカイブに適したファイルフォーマットで保存します。多くの機関でウェブアーカイブの保存用ファイルフォーマットである WARC（Web ARChive）が採用されています。

▶ 第 1 章 8（p.26）
▶ 第 2 章 3（p.43）

(5) 公開

ウェブアーカイブの目的や事情に応じて公開の範囲は様々です。収集するだけで非公開（ダークアーカイブ)、学術研究など限られた目的に対してのみ公開や施設内でのみ公開（グレイアーカイブ)、インターネット上で公開（ホワイトアーカイブ）など色々な公開レベルがあります。

▶ 第 1 章 9（p.30）
▶ 第 2 章 4（p.46）

[2] https://archive.org/

3. ウェブを収集するしくみ

ウェブアーカイブでは収集ロボット（クローラ）と呼ばれる自動プログラムを用いて、ウェブサイトを収集します。

（1）収集ロボットによる収集

収集ロボットは、最初にスタート地点となるウェブページ（起点 URL）にアクセスをします（図 4）。そして、そのページの html ファイルを収集すると同時に、html ファイル内のソースを解析して文書、画像、音声、動画、スタイルシートなどのファイルを収集します。さらにそこからリンクしているページに移動して、同様の処理を繰り返し行います。

このようにリンクをたどりながらページを移動し続け、新たなリンク先がなくなるまで処理を続けます（「第 2 章 1. 収集ロボット Heritrix」（p.36））。

（2）ウェブページのつくり

ウェブページは一見するとページが 1 枚だけあるように見えますが、実際には html ファイルや画像ファイル、スタイルシート、スクリプトファイルなど、多数のファイルが組み合わさって構成されています。

例えば、2014 年 9 月 17 日時点の国立国会図書館のトップページ（図 5）は表 1 のように、1 つの html ファイル、5 つの CSS ファイル、8 つの JavaScript ファイル、61 の画像ファイル、合計 75 のファイルから構成されています。

収集ロボットがこれら全てのファイルを漏れなく収集することで、ウェブページをオリジナルと同じように再現することができるのです。

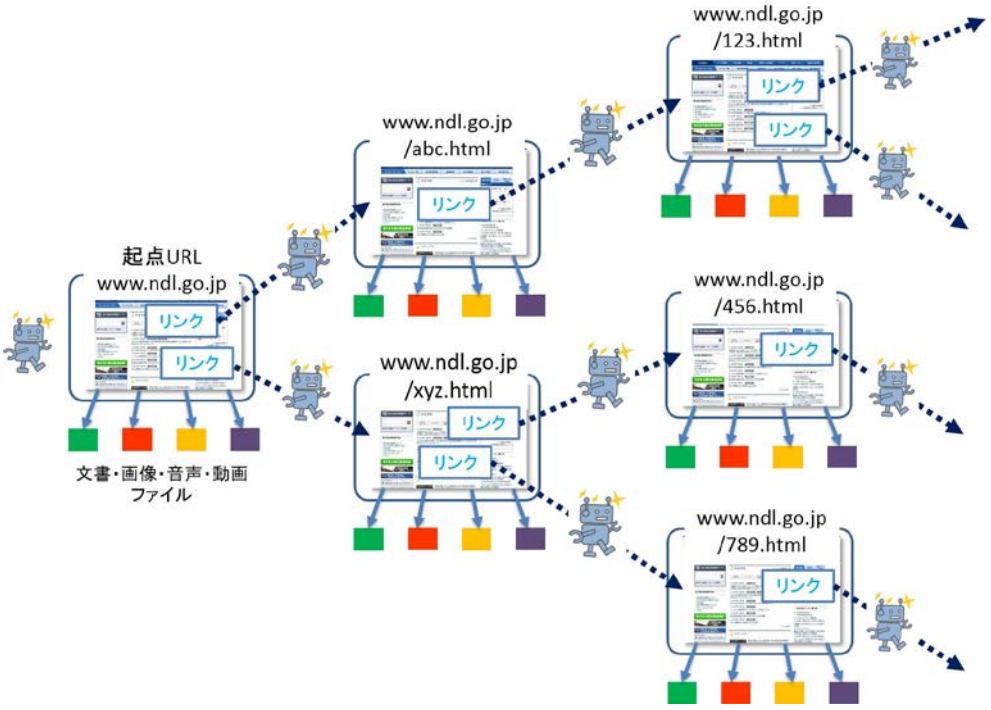


図 4：収集ロボットによるウェブサイト収集のイメージ

図 5：国立国会図書館のトップページ（2014 年 9 月 17 日時点）

表 1：国立国会図書館のトップページを構成するファイル

ファイルの URL	ファイルの種類
http://www.ndl.go.jp/	html ファイル
http://www.ndl.go.jp/common/css/common.css	CSS ファイル
http://www.ndl.go.jp/common/css/noscript.css	
http://www.ndl.go.jp/common/css/re_print.css	
http://www.ndl.go.jp/common/css/re_top.css	
http://www.ndl.go.jp/common/css/search.css	
http://www.ndl.go.jp/common/js/calendar.js	JavaScript ファイル
http://www.ndl.go.jp/common/js/calendar_holiday.js	
http://www.ndl.go.jp/common/js/calendar_summary.js	
http://www.ndl.go.jp/common/js/func.js	
http://www.ndl.go.jp/common/js/ndlsearch-inside.js	
http://www.ndl.go.jp/common/js/search.js	
http://www.ndl.go.jp/common/js/tab_change.js	
http://www.ndl.go.jp/urchin.js	

http://www.ndl.go.jp/common/images/banner_archive.png	画像ファイル
http://www.ndl.go.jp/common/images/banner_ca.gif	
http://www.ndl.go.jp/common/images/banner_digi.gif	
http://www.ndl.go.jp/common/images/banner_hourei.gif	
http://www.ndl.go.jp/common/images/banner_jikocho.gif	
http://www.ndl.go.jp/common/images/banner_kaigiroku.gif	
http://www.ndl.go.jp/common/images/banner_kindai.gif	
http://www.ndl.go.jp/common/images/banner_ndl_opac.gif	
http://www.ndl.go.jp/common/images/banner_reference.gif	
http://www.ndl.go.jp/common/images/banner_research.gif	
http://www.ndl.go.jp/common/images/bg_header.jpg	
http://www.ndl.go.jp/common/images/bg_header_l.jpg	
http://www.ndl.go.jp/common/images/bg_r_search.gif	
http://www.ndl.go.jp/common/images/bg_title_top.gif	
http://www.ndl.go.jp/common/images/bnr_kids.jpg	
http://www.ndl.go.jp/common/images/bt_event_off.gif	
http://www.ndl.go.jp/common/images/bt_issue_off.gif	
http://www.ndl.go.jp/common/images/bt_news_off.gif	
http://www.ndl.go.jp/common/images/bt_press_off.gif	
http://www.ndl.go.jp/common/images/bt_pulldown.gif	
http://www.ndl.go.jp/common/images/bt_recruit_off.gif	
http://www.ndl.go.jp/common/images/bt_search-b_off.gif	
http://www.ndl.go.jp/common/images/bt_search-b2_off.gif	
http://www.ndl.go.jp/common/images/copyright_bg.gif	
http://www.ndl.go.jp/common/images/footer_bg.gif	
http://www.ndl.go.jp/common/images/ico_arrow.gif	
http://www.ndl.go.jp/common/images/ico_arrow_c.gif	
http://www.ndl.go.jp/common/images/ico_arrow_down.gif	
http://www.ndl.go.jp/common/images/ico_etc.gif	
http://www.ndl.go.jp/common/images/ico_event.gif	
http://www.ndl.go.jp/common/images/ico_home.gif	
http://www.ndl.go.jp/common/images/ico_issue.gif	
http://www.ndl.go.jp/common/images/ico_news.gif	
http://www.ndl.go.jp/common/images/ico_point.gif	
http://www.ndl.go.jp/common/images/ico_recruit.gif	
http://www.ndl.go.jp/common/images/ico_rss.gif	
http://www.ndl.go.jp/common/images/ico_stop.gif	

http://www.ndl.go.jp/common/images/logo.jpg	画像ファイル
http://www.ndl.go.jp/common/images/menu1-01_off.jpg	
http://www.ndl.go.jp/common/images/menu1-02_off.jpg	
http://www.ndl.go.jp/common/images/menu1-03_off.jpg	
http://www.ndl.go.jp/common/images/menu1-04_off.jpg	
http://www.ndl.go.jp/common/images/menu1-05_off.jpg	
http://www.ndl.go.jp/common/images/menu1-06_off.jpg	
http://www.ndl.go.jp/common/images/menu1-07_off.jpg	
http://www.ndl.go.jp/common/images/menu1-title.jpg	
http://www.ndl.go.jp/common/images/menu2-01_off.jpg	
http://www.ndl.go.jp/common/images/menu2-02_off.jpg	
http://www.ndl.go.jp/common/images/menu2-03_off.jpg	
http://www.ndl.go.jp/common/images/menu2-04_off.jpg	
http://www.ndl.go.jp/common/images/menu2-05_off.jpg	
http://www.ndl.go.jp/common/images/menu2-06_off.jpg	
http://www.ndl.go.jp/common/images/menu2-title.jpg	
http://www.ndl.go.jp/common/images/tab_top_off.gif	
http://www.ndl.go.jp/common/images/tab_top_on.gif	
http://www.ndl.go.jp/common/images/title_ndl-opac.jpg	
http://www.ndl.go.jp/common/images/top_title1.gif	
http://www.ndl.go.jp/common/images/top_title2.gif	
http://www.ndl.go.jp/jp/spot/__icsFiles/thumbnail/2014/06/16/spot_survey.gif	
http://www.ndl.go.jp/jp/spot/__icsFiles/thumbnail/2014/07/03/spot_space.jpg	
http://www.ndl.go.jp/jp/spot/__icsFiles/thumbnail/2014/07/14/spot_forum.jpg	

4. ウェブを収集する単位

ウェブアーカイブでは、何らかのまとまり（単位）でウェブサイトを管理します。主なものとしてターゲット単位とページ単位があります。単位を決定するにあたっては、収集の規模、保存したサイトの見せ方、収集の効率性などがポイントとなります。

(1) ターゲット単位

(i) 小規模な収集

選択収集の多くは小規模なまとまりで収集するため、機関単位あるいはウェブサイト単位で収集する対象（ターゲット）を設定します。機関単位であれば「国立国会図書館（http://warp.da.ndl.go.jp/waid/280）」や「総務省（http://warp.da.ndl.go.jp/waid/1607）」、ウェブサイト単位であれば「平城遷都 1300 年祭（http://warp.da.ndl.go.jp/waid/11878）」や「さっぽろ雪まつり（http://warp.da.ndl.go.jp/waid/5223）」といった具合です。各単位で検索用のメタデータを作成し、その下に保存日ごとに分けてウェブサイトを管理します（図 6）。ターゲット単位で管理することで、収集や公開に関する許諾の管理がスムーズに行えるなどのメリットがあります。

(ii) 保存したサイトの見せ方

ターゲット単位で保存したウェブサイトの多くは、そのまとまりで公開します。例えば WARP では、ターゲットのもとに保存日ごとに分けて公開しています（図 7）。

保存したウェブサイトを閲覧している際、リンク先のファイルが同じ日に保存されていない場合には、メッセージ画面が表示されます（図 8）。他の保存日や、他のターゲットの中に保存されている可能性がありますので、画面に従って検索し直すと見つかる場合があります（図 8①）。また、オリジナルサイトへのリンクもありますので、現在のページへ飛ぶこともできます（図 8②）。

図 6：ターゲット単位での収集のイメージ

図 7：WARP の詳細画面

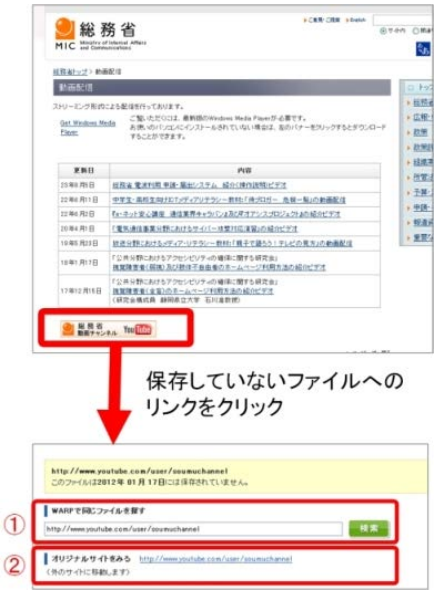


図 8：WARP のメッセージ画面

(2) ページ単位

(i) 大規模な収集

ページ単位の管理は、バルク収集に多く用いられる方法です。「.fr」や「go.jp」などのドメインレベルで大規模な収集を行い、URL のみで管理します（図 9)。ターゲット単位とは異なり、検索用のメタデータを付与することはあまりありません。

(ii) 保存したサイトの見せ方

ページ単位の多くは URL をキーにして検索・閲覧をします。例えばインターネットアーカイブの Wayback Machine（https://archive.org/）は、URL で検索をするとカレンダーに保存日が表示されます（図 10)。

表示されるコンテンツはページ単位で管理されているためターゲットの枠はなく、リンク先のファイルが全体アーカイブデータのなかに保存されていれば表示されます。リンク先

図 9：ページ単位での収集のイメージ

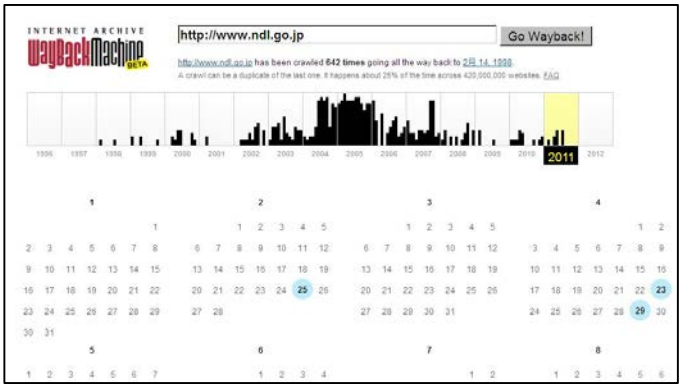


図 10：Wayback Machine のカレンダーナビゲーション

図 11：Wayback Machine の画面遷移

のページは、Wayback Machine の場合、保存しているなかで一番近い日付のページが表示されます（図 11）。

保存したサイトの見せ方においては、ウェブサイトをかたまりとして見るかどうかが、ターゲット単位とページ単位の大きな違いと言えるでしょう。保存日を固定して見せるかどうかは選択事項ですので、システム要件などを考慮して決定します。

（3）収集の効率性

収集ロボットを使用してウェブサイトをクロールする際、同時に走行できる収集ロボットの数（プロセス数）はシステム規模に応じて限りがあります。そのため、1 プロセスで複数の URL をまとめて収集すると、より効率よく収集することができます。収集単位を設定する際には、こうした効率性の観点も必要です。

WARP は、ウェブサイトを発信している機関ごとに複数の URL を設定しています（図 12）。例えば、総務省の場合、総務省のメインサイト（http://www.soumu.go.jp/）のほか、統計局

（http://www.stat.go.jp/）、電子政府の総合窓口（http://www.e-gov.go.jp/）を始め十数の URL が存在します。それらを起点に設定して、1 プロセスで収集を行います。収集したファイルは全体をまとめて「総務省」として管理しています。

ただし、起点として設定する URL の数が増えるほど 1 プロセスあたりの収集にかかる時間は長くなるため、プロセス数と起点 URL 数のバランスを考慮する必要があります。特にバルク収集や大規模な選択収集の場合には、収集する単位と効率性を総合的に勘案した上で収集スケジュールを立てることになります。

WARP でも、同時走行が可能なプロセス数、ターゲット単位での収集保存、公的機関サイトの網羅的収集などを所与条件とし、この条件下で最も効率よく収集ができるように収集スケジュールを設定しています。

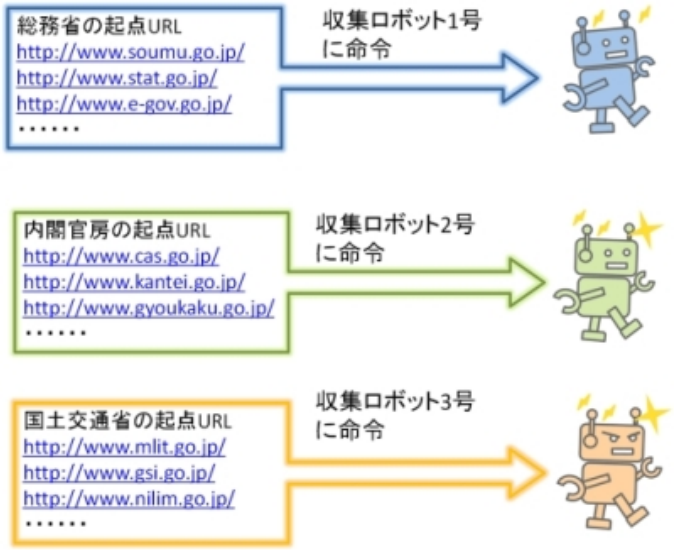


図 12：複数の起点 URL をまとめて収集設定

5. 収集する頻度

ウェブサイトを収集する単位が決まると、次に収集する頻度（収集間隔）を決める必要があります。その前に収集するタイミングについて考えてみましょう。

(1) 収集するタイミング

ウェブサイトを効率的に収集するには、どのようなタイミングで収集するのが理想的でしょうか？答えはウェブサイトの更新直後です。なぜなら、一度ウェブサイトを収集した後、更新前に再度収集を行えば、前回収集したウェブサイトと同じ内容（状態）のページを収集することになりますし、逆に更新後しばらく間を置いてから収集を行うと、収集前に再度更新されたり、ページ自体が削除されたりする恐れがあるからです。

(2) ウェブサイトの更新頻度

では、ウェブサイトの更新頻度はどのくらいでしょうか？ウェブコンテンツの平均寿命については、75 日や 100 日など諸説あります[3]。しかし、ニュースサイトのように毎日更新されるものもあれば、１年に一度程度の更新しかされないものもあり、更新頻度はウェブサイトによってまちまちです。

(3) ウェブサイトの収集頻度

以上のことから、それぞれのウェブサイトが更新される度に収集するのが最も理想的なのですが、そのためにはウェブサイトの更新をリアルタイムで検知する仕組みが必要になります（図 13）。

図 13：リアルタイム検知のイメージ

[3] Michael Day. Collecting and preserving the World Wide Web：a feasibility study undertaken for the JISC and Wellcome Trust. Joint Information Systems Committee, 25 February 2003, 7p. http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf

大学研究機関においてビッグデータの解析などのために、そのような仕組みを搭載した高性能クローラを開発しているところもありますが、ウェブアーカイブの運用機関でこうしたクローラを使用している例はあまりありません。

多くの場合は、予め収集頻度を決めて定期的に収集する方法を採用しています（図 14）。

バルク収集を行っている機関の多くは年平均 1～3 回の頻度で収集を行っています。しかし、そのような低頻度の収集では、ニュースサイトをはじめとする更新頻度の高いウェブサイトを収集しきれないという問題があります。そのため、バルク収集を行っている機関の多くは選択収集を併用しており、必要なサイトを選択して高頻度の収集を行っています。

WARP では大規模な選択収集を行っており、それぞれのウェブサイト（ターゲット）毎に収集頻度を定めています。

（4）WARP の収集頻度

WARP では収集頻度をターゲット単位で予め決め、定期的（毎月等）に収集を行っています。原則として「国の機関」のウェブサイトは月 1 回（年 12 回）、それ以外は四半期ごとに 1 回（年 4 回）です（図 15）。この収集頻度は以下の 3 つの観点から決められました。

(i) 法律に基づく収集

2010 年 4 月から国立国会図書館法に基づいて、公的機関ウェブサイトの網羅的な収集を開始しました。中でも「国の機関」のウェブサイトで発信される情報を可能な限り保存できるよう、高頻度（毎月）に収集を行うことが適当と考えました。

(ii) 相手先サーバへの考慮

相手先サーバへ負荷をかけないようにするため、ターゲット内の各ページをダウンロードする間隔を 1 秒以上空けることを前提に収集頻度を算出する必要がありました。

(iii) システムにおける収集能力

システムの能力に見合った収集を行う必要がありました。システム上、同時に収集できるターゲット数は 50、全てのターゲット数は約 5,000 という条件下で、最も効率的に収集できる頻度を算出しました。

図 14：定期的な収集のイメージ

コレクション名	頻度
国の機関	毎月（年12回）
都道府県	四半期ごと（年4回）
政令指定都市	四半期ごと（年4回）
市町村	四半期ごと（年4回）
独立行政法人等	四半期ごと（年4回）
大学	四半期ごと（年4回）
電子雑誌	刊行頻度に合わせる
イベント	随時

図 15：WARP における収集頻度

「祭り」や「映画」といったイベントサイトは、それぞれのイベント開催直後に収集を行っています。また、東日本大震災などの大規模災害時には、頻繁に更新される情報を確実に保存するため、通常よりも頻度を上げて収集をしています。

6. 差分収集

ウェブアーカイブでは、同じウェブサイトを定期的に収集していきます。そのため、収集するファイルのなかには、過去に収集した時点から更新されているファイルもあれば、過去と全く同じファイルもあります。

収集するたびに全てのファイルを保存する方法をフル収集と言い、変更があったファイルのみを保存する方法を差分収集と言います。

フル収集では、同じファイルを重複して保存することになりますので、必要なストレージ（電子書庫）の容量が大きくなります。一方、差分収集では同じファイルは保存しないため、ストレージを節約することができます。

(1) フル収集と差分収集のしくみ

フル収集と差分収集について、模式図で詳しく見てみましょう。

オリジナルのウェブサイトが以下であると仮定します（図 16）。

・1 回目の収集時には、A.html、B.pdf、C.docx、D.png が存在。
・2 回目の収集時には、A.html、B.pdf、D.png には変更がなく、C'.docx が変更、E.xlsx が追加。
・3 回目の収集時には、D.png と E.xlsx には変更がなく、A'.html、B'.pdf、C''.docx が変更。
（「'」はファイル名の変更ではなく、データ内容の変更を表す。）

図 16：オリジナルのウェブサイト

(i) フル収集

フル収集（図 17）では、ファイル変更の有無に関わらず、収集するたび全てのファイルを保存します。そのため、重複して保存されるファイルがあり、ファイル数の合計は「14」になります。

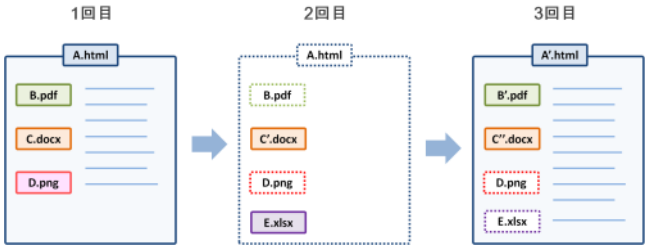
ファイル	保存数
A.html	2
A'.html	1
B.pdf	2
B'.pdf	1
C.docx	1
C'.docx	1
C''.docx	1
D.png	3
E.xlsx	2
合計	14

図 17：フル収集

(ii) 差分収集

差分収集では、過去に収集したのと同じファイルがある場合、そのファイルは保存しません。図 18 のように、同一ファイルの点線部分は保存せずに、実線部分のみを保存します。

その結果、各ファイルを保存する回数は 1 回のみで、ファイル保存数の合計は「9」になります。フル収集時の保存数「14」と比べると、少なくなっているのが分かります。



ファイル	保存数
A.html	1
A'.html	1
B.pdf	1
B'.pdf	1
C.docx	1
C'.docx	1
C''.docx	1
D.png	1
E.xlsx	1
合計	9

図 18：差分収集

(2) ハッシュ値による比較

差分収集において、同一ファイルかどうかの判定は、ハッシュ値を比較して行います (図 19)。

ハッシュ値とは、電子データを一定の計算方法（ハッシュ関数）で操作して得られる値のことです。異なる電子データのハッシュ値が同じになることは殆どないため、電子データにおける指紋に例えられます。電子データに僅かでも変更を加えると、ハッシュ値も変わります。

新たに収集したファイルを保存する際には、前回の収集ログに同名のファイルが存在しなければ、新たに保存します。同名のファイルが存在する場合には、ハッシュ値を比較して異なる場合のみ保存します。

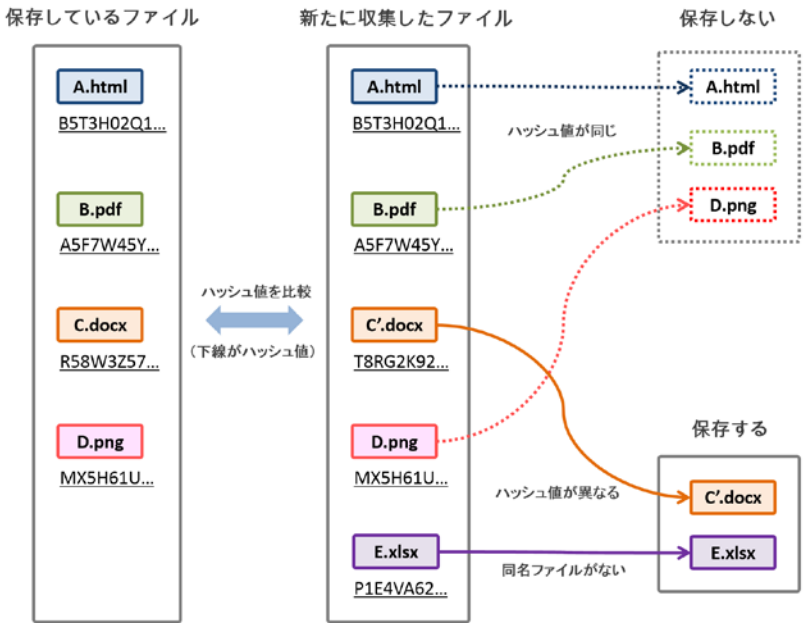


図 19：ハッシュ値の比較

（3）差分収集したウェブサイトの再現

差分収集で保存したウェブサイトを再現する際、収集した時点のファイルを保存している場合はそのファイルを表示し、その時点のファイルがない場合は、一番近い過去に保存した同名のファイルを表示します。これは、収集の際にハッシュ値を比較して同値だったファイルですので、収集時点が異なっていてもオリジナルの状態を保ったままで再現することができます（図 20）。

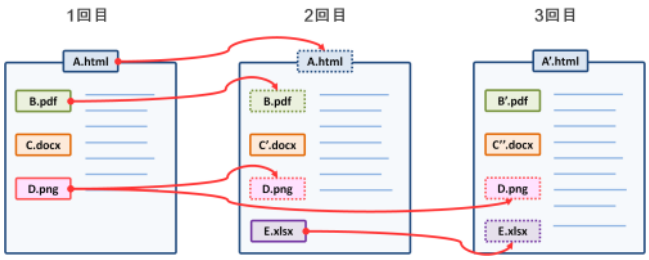


図 20：差分収集したウェブサイトの再現

（3）ストレージの節約効果

差分収集をすることで、保存するファイルを少なくすることができ、ファイルの保存に必要なストレージの容量を削減することができます。

「第 1 章 5. 収集する頻度」（p.16）で紹介したように、WARP では国の機関を毎月、その他を概ね年 4 回の頻度で収集しています。これらを差分収集した場合、フル収集に比べて約 7 割の削減効果があることが分かっています。つまり、必要なストレージの容量が、フル収集の 3 割程度で済みます。

このように、膨大なデータを扱うウェブアーカイブにおいては、差分収集がストレージの節約に大きな効果を発揮するのです。

7. 収集したウェブサイトの組織化

収集したウェブサイトの組織化について、URL 管理、メタデータの付与、全文検索、データマイニングに分けて紹介します。

(1) URL 管理

収集したウェブサイトの URL は、オリジナル URL との関係を保ちつつ、オリジナル URL とは違うものである必要があります。多くのウェブアーカイブでは、日付や識別子とオリジナル URL を組み合わせて URL を付与しています。

(i) 日付と URL の組み合わせ（図 21）

（例）http://web.archive.org/web/20040618115539/http://www.meti.go.jp/

2004 年 6 月 18 日 11 時 55 分 39 秒に収集した http://www.meti.go.jp/のページということを表しています。

(ii) 識別子と URL の組み合わせ（図 22）

（例）http://warp.da.ndl.go.jp/info:ndljp/pid/285403/www.meti.go.jp/

保存日ごとに割り振られた識別子（info:ndljp/pid/285403）と、オリジナル URL の http://www.meti.go.jp/を組み合わせています。

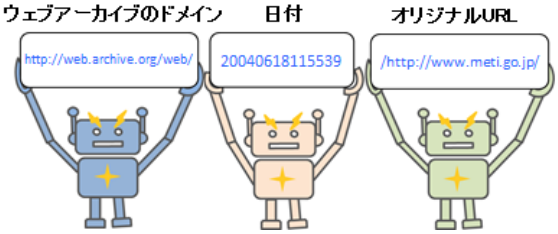


図 21：日付と URL の組み合わせ

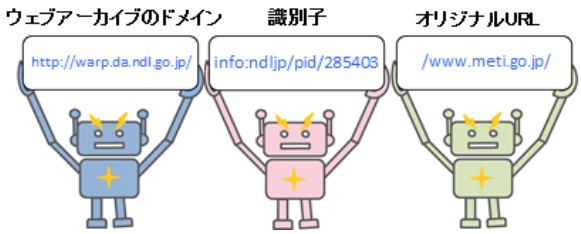


図 22：識別子と URL の組み合わせ

(2) メタデータの付与

メタデータについては、どのような単位＝粒度で付与するかという問題があります。以下の 3 つの観点から考えてみましょう。

(i) アーカイブの規模

バルク収集の場合、膨大な収集を行うため粒度の細かいメタデータを付与しづらいのに対し、選択収集の場合、収集量は小規模なため、比較的細かいメタデータを付与しやすい傾向があります。

(ii) 需要と供給のバランス

利用する側が求めるメタデータの粒度に対し、提供する側がそれに対してどこまで応えられるかという視点は重要です。

利用する側が詳細なメタデータを求める場合、提供する側はできる限りその要求に沿った粒度のメタデータを用意するのが理想です。

しかし、ウェブコンテンツの量が膨大であること、人的・財政的制約があることなどから、全てのウェブコンテンツに詳細なメタデータを人的に付与するのは現実的ではありません。

将来的には、セマンティック・ウェブなどの技術を活用して、収集したウェブサイトの内容をシステムが理解し、自動でメタデータを付与することが期待されます。また、ソーシャルタギングのように、閲覧しているコンテンツに対しユーザ自身が主題等のメタデータを付与する仕組みを導入することも考えられます。

(iii) 対象コンテンツ

対象コンテンツによっても適切な粒度は異なります。

ターゲット単位で収集を行う場合、収集前にターゲットのタイトル・公開者（出版者）・起点 URL 等のメタデータをターゲットに付与します。このメタデータ、若しくは項目が追加されたメタデータが、そのままウェブサイト閲覧時の検索においても流用されます。効率よく収集ができる単位で付与されているため、粒度は粗くなりがちです。

一方で、精度の高い検索結果が求められるコンテンツに対しては、粒度の細かいメタデータを集中的に付与することもあります。例えば、ウェブアーカイブの多くは国立図書館が実施していますが、図書館としては、電子雑誌の論文記事の単位でメタデータを付与するのが望ましいでしょう。

国立国会図書館では、保存したウェブサイトの中から白書、会議資料、報告書、年報、論文などの刊行物を取り出し、それらに対しメタデータを付与しています（図 23）。こうすることで、ウェブサイト中に散在する刊行物を効率よく検索・閲覧でき、従来の紙の刊行物との連続的なアクセスも保障されます。

図 23：国立国会図書館デジタルコレクション：電子書籍・電子雑誌
http://dl.ndl.go.jp/#internet

(3) 全文検索

(i) メタデータと全文検索の補完関係

検索をする場合、メタデータだけではそれより細かい情報はヒットしません。そこで、全文検索インデクスを作成し、ウェブサイト本文をより細かく検索できるようにします。世界のウェブアーカイブ機関の約 6 割が、この全文検索の機能を提供しています[4]。

多くのウェブアーカイブで使われている全文検索エンジン Solr は、検索結果を適合度順に表示することができますが、同時に不要な情報も数多く表示されます。

メタデータによる粗くてノイズの少ない検索と、全文検索による細かくてノイズの多い検索、それぞれの特徴を活かしながら検索サービスを提供することが必要です（図 24）。

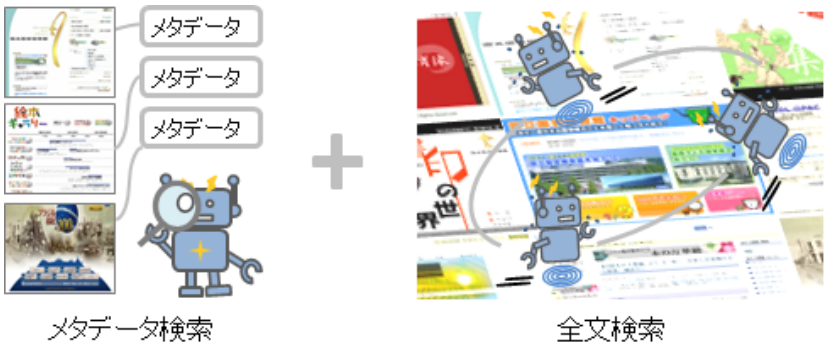


図 24：検索のイメージ

(ii) WARP の検索サービス

WARP では、メタデータ検索と全文検索を組み合わせた検索サービスを提供しています。検索する際は、トップページのキーワード検索欄に単語を入力し検索します（図 25）。

検索結果画面では、メタデータの検索結果と全文検索の結果が同時に表示され、タブで両者を切り換えることができます（図 26）。

図 25：WARP 検索画面

図 26：WARP 検索結果画面

[4] Daniel Gomes, Joao Miranda, and Miguel Costa. A survey on web archiving initiatives. http://sobre.arquivo.pt/about-the-archive/a-survey-on-web-archiving-initiatives

（iii）インデキシングの維持

全文検索サービスを提供するためには、高い全文検索能力を備え、かつ高速なインデキシングが可能な検索サーバが必要です。また、日々保存されるウェブサイトは膨大な数に上るため、それを処理・保存できるリソース・技術も必要です。

（4）データマイニング

収集した膨大なアーカイブデータを解析し、データの相関関係、パターンなどを探し出す技術のことをデータマイニングと言います。ウェブアーカイブ機関でも、様々なデータマイニングの試みが行われています[5]。

UK Web Archive の「Visualization」[6]はその好例で、ある特定の単語がアーカイブ中に出現する頻度を時系列で表示する「N-gram Search」[7]や、収集したウェブサイトを収集年ごとに解析し、ファイルフォーマットの流行を分析する「Format Analysis」[8]などがあります。

[5] Emily Reynolds. Web Archiving Use Cases. http://netpreserve.org/sites/default/files/resources/UseCases_Final_1.pdf

[6] http://www.webarchive.org.uk/ukwa/visualisation

[7] http://www.webarchive.org.uk/ukwa/ngram

[8] http://www.webarchive.org.uk/ukwa/visualisation/ukwa.ds.2/fmt

8. ウェブアーカイブの長期保存

(1) 長期保存とは

書物であれデジタルコンテンツであれ、アーカイブにおいては、そこに記録された内容を百年単位、千年単位の長期間にわたって利用できる状態で保存しなければなりません。そうした対策や試みは「長期保存」と呼ばれます。

デジタルコンテンツの長期保存にあたっては、考慮しなければならない特有の要点があります。それは、「ビット列の保存 (Bit-stream preservation)」と「論理的な保存 (Logical preservation)」です。「ビット列の保存」とは 0 と 1 からなる文字列を欠損なく完全に保存することで、「論理的な保存」とは 0 と 1 からなる文字列の意味を人が読み取り理解できるように再現することです。

(i) ビット列の保存

ビット列の保存はデジタルコンテンツの長期保存における最も基本的な対策で、冗長化とバックアップがその主な手段です。

(ii) 冗長化

RAID（Redundant Arrays of Inexpensive Disks）などの技術を用いてデータを冗長化して保存することで、ハードディスクの故障によるデータの消失に備えることができます。

RAID とは、複数のハードディスクを組み合わせて一つの仮想的なハードディスクとして使用し、ディスク故障時のデータ復旧を可能にする技術です。複数ディスクにデータを分散して記録（striping）、複数ディスクに同じ内容を記録（mirroring）、それらの組み合わせ、誤り訂正符号データ（parity）の使用などのヴァリエーションにより、いくつかの RAID レベルがあります。

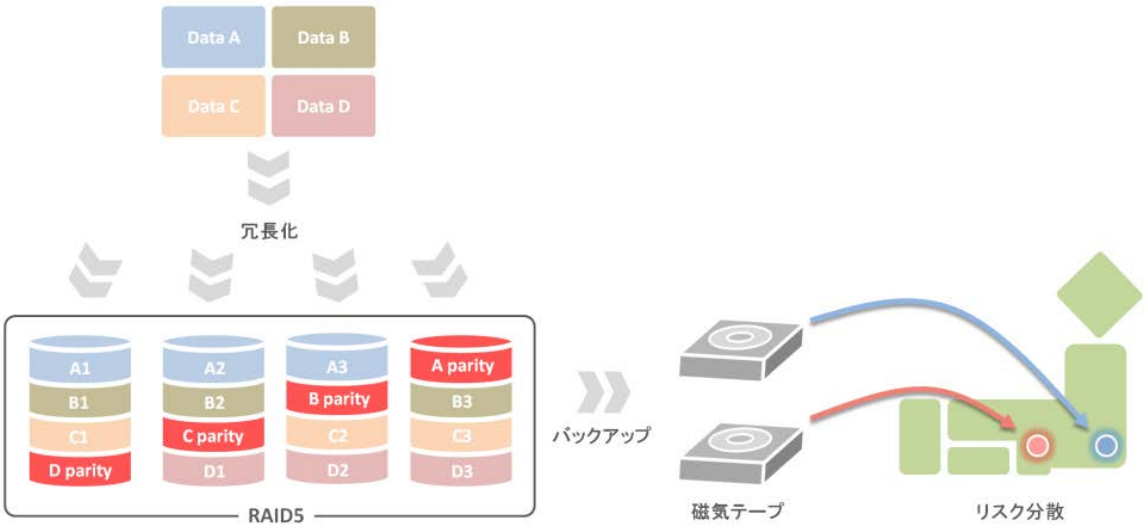


図 27：冗長化、バックアップ、リスク分散のイメージ

(iii) バックアップ

ハードディスクのデータを磁気テープなどの記録媒体に定期的にバックアップを取って複数世代を残します。さらに保存場所を分けてリスク分散（Disaster Recovery）を図ることも重要です。データを格納した記録媒体は物理的に安定した環境で保存し、記録媒体の劣化に対しては媒体変換などの対策が必要です。

(2) 論理的な保存

ビット列が完全に保存されていたとしても、人がその内容を理解できなくては意味がありません。論理的な保存の有力な方法がマイグレーション（Migration）とエミュレーション(Emulation)です。

(i) マイグレーション

ハードウェアやソフトウェアの環境の変化によりファイルが技術的に読めなくなってしまう前に、フォーマットを変換したり別の記録媒体へ移行したりする方法で、最も広く行われている対策です。

例えば、旧式のワープロソフトで作成したファイルを最新のワープロソフトのデータ形式にフォーマット変換をしたり、再生機器の変化にともないフロッピーディスクから光ディスクへと媒体を変えてデータを移行したりします。

ただし、環境が変化するごとにマイグレーションを繰り返す必要があるため、ウェブアーカイブのように大量なデータ群のマイグレーションには膨大なコストがかかります。また、マイグレーションを繰り返すことで起こりうるデータ改変や作業ミスによるデータ消失などの危険性もはらんでいます。

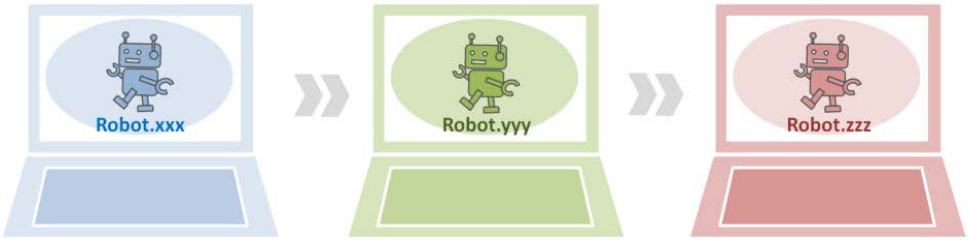


図 28：マイグレーション(フォーマット変換)のイメージ

(ii) エミュレーション

エミュレータ（Emulator）と呼ばれるソフトウェアを使って、旧式化したファイルやソフトウェアの再生環境を新しいハードウェア・ソフトウェアの環境下で模擬的に再現することです。

例えば、Windows3.1 でしか動作しないソフトウェアでも、エミュレータを使って最新の OS 環境上で Windows3.1 の環境を再現することで利用できるようになります。また、市場

から消えてしまった専用ゲーム機でのみプレイできたゲームソフトも、エミュレータを使えばパソコン上でプレイすることができます。

オリジナルのファイルに変更を加えないためデータ改変などの危険性が少ないのが利点ですが、旧ファイルの再生環境を完璧に再現するのが難しいケースもあり、近似の再現に留まるエミュレータもあります。エミュレータの開発にはコストがかかりますが、マイグレーションのコストと比べて低く抑えられるため、今後、より洗練されたエミュレーション技術の確立が期待されます。あわせて長期的な保存に耐えうる信頼性の高い記録媒体が開発されれば、より有用な手段になると考えられています。

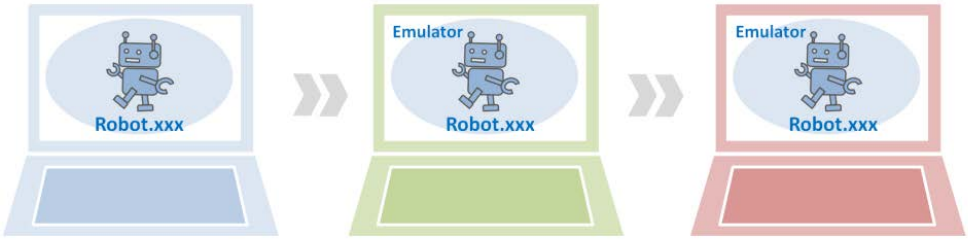


図 29：エミュレーションのイメージ

(3) 保存メタデータ

マイグレーションやエミュレーションを効果的に管理・実施するためには、データをアーカイブする時点で再生機器や再生環境、作成アプリケーション、ファイルフォーマットのヴァージョンなどを記録した保存メタデータを作成しておくことが求められます。保存メタデータに記録された情報と最新の技術動向を照らし合わせることで、ファイルの旧式化を適時に把握してマイグレーションを行ったり、エミュレーションの環境を適切に用意したりすることができます。

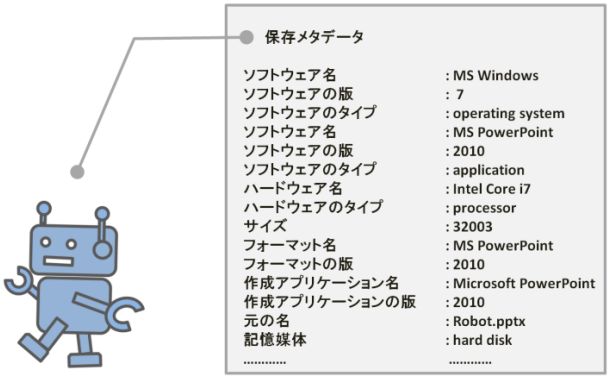


図 30：保存メタデータのイメージ

(4) ウェブアーカイブの長期保存

以上、デジタルコンテンツの長期保存の技術についてみてきましたが、それではウェブアーカイブの長期保存はどこまで進んでいるのでしょうか？

ウェブアーカイブのデータはファイルの数量や種類が膨大なため、マイグレーションやエミュレーションを包括的に行うのは容易ではありません。それらをどのように施していくのがよいのか、使用するソフトウェアはなにが最良なのか、より効果的な別の方策はないのかなどについて、いくつかの調査や試行が行われていますが、未だ確実な方策は定まっていません。結果、多くのウェブアーカイブにおいては本格的な長期保存の対策がなされていないのが現状です。

世界中でウェブアーカイブが行われるようになって既に十数年が経過しますが、果たして古い時期のウェブサイトは今でも問題なく閲覧できるのでしょうか？例えば、WARP で保存している最も古い 2000 年 9 月のウェブサイト（図 31）を見る限り、少なくとも html ファイルや画像ファイルについては現在でも読み取ることができます。

一見すると問題ないように思えますが、ウェブサイトのなかには多種多様なフォーマットのファイルが含まれているため、なかには既に旧式化して読めなくなっているファイルがある可能性もあります。時間の経過とともにそのようなファイルは確実に増えていくでしょう。また、そもそも将来にわたってウェブ技術そのものが現在と同じであり続ける保障はどこにもありません。

このようにウェブアーカイブの長期保存は不安定な状況におかれていると言えるのです。ウェブ情報の多くは他の媒体には記録されていないため、長期保存に失敗した場合のインパクトは極めて大きいでしょう。そうした危機感のもと、IIPC を始めとする関係機関により長期保存の検討や研究開発が進められています。

ウェブアーカイブはウェブサイトを集めて保存すれば終わりではありません。長期間にわたって人が理解できるように保存し続けるためには、絶え間ない取り組みが欠かせないのです。

図 31：WARP で保存している最も古いウェブサイト
大地の芸術祭 越後妻有アートトリエンナーレ 2000
保存日：2000 年 9 月 13 日
http://warp.da.ndl.go.jp/info:ndljp/pid/236618/www.artfront.co.jp/art_necklace/jp/index.htm

9. 保存したウェブサイトの公開

世界各国のウェブアーカイブでは、そこで保存されているコンテンツのすべてが無条件でインターネット公開されることは珍しく、アクセス可能な場所や資格、範囲など、何らかの制限を設けて公開されるのが一般的です。

（1）公開方法の変遷

ウェブアーカイブが本格的に実施され始めた当初は、コンテンツがまったく公開されないことも珍しくありませんでした。

当時は安定した収集技術の確立に力が注がれており、先ずは日々消えていくウェブサイトを消滅する前に収集することが最優先の課題とされていました。相対的に、公開については未検討あるいは優先度が低いとする考え方があり [9]、それは当時の状況下では決して不自然なものではありませんでした。

一方で、一部には公開を意識していたウェブアーカイブもあり、システムにより様々なパターンのアクセス制御（公開期間の設定、対象ユーザの設定など）を可能にすることで、コンテンツの性質に応じた公開を行っていました [10]。

その後、世界各国のウェブアーカイブ機関の連携・協力により、収集ロボットを始めとする基本的な技術が開発され収集が安定的に行えるようになると、徐々に公開に焦点が当てられるようになりました。

2009 年 10 月に開催された IIPC オープンミーティングにおいて、インターネットアーカイブの代表であるケール（Brewster Kahle）氏は、『ウェブアーカイブがたとえ長期保存されても、利用に供されない「ダークアーカイブ」では意味がなく、常に利用を前提とした保存を考えることが重要である』と述べています [11]。

ダークアーカイブとは非公開のアーカイブのことで、その他、制限を設けて公開されるものはグレイアーカイブ、インターネット上で公開されるものはホワイトアーカイブと呼ばれることがあります。

（2）公開に制限を設ける理由

インターネットアーカイブの Wayback Machine では、収集したコンテンツを原則インターネット上で公開しており、申し出があれば公開を停止する方式（オプトアウト方式）を採用しています。ただし、これは米国著作権法に定めるフェアユースの考えに基づいており、ウェブアーカイブの中でも特異な例と言えます。

ほとんどのウェブアーカイブは、公開にあたって何らかの制限を設けざるを得ません。その理

---

9 北欧諸国におけるウェブ・アーカイビングの現状と納本制度
http://dl.ndl.go.jp/view/download/digidepo_1052103_po_NDLM_490.pdf?contentNo=1&alternativeNo=

10 ウェブ・アーカイビング―オーストラリア・オンライン出版物国家コレクション「文化資産としてのウェブ情報」http://www.ndl.go.jp/jp/publication/proceedings/web_archive/websympo2002j.pdf

11 IIPC オープンミーティング及びワーキンググループ＜報告＞ http://current.ndl.go.jp/e995

由として以下のような観点が挙げられます。

(i) 著作権

ウェブアーカイブの公開にあたっては、自国の著作権法を遵守し発信者の著作権を侵害しないよう注意する必要があります。

法律により複製に係る著作権を制限することで、事前に発信者からの許諾を得ることなく収集が可能であっても、インターネット公開については法律による権利制限が設けられていない場合が多くあります。特に「.fr」や「.uk」などのトップレベルドメインでバルク収集したコンテンツがインターネットで無条件で公開されることはほとんどありません。その多くが施設内に限定したり研究目的に限定したりして公開されています。

(ii) 個人情報

ウェブサイト内には個人情報が多く含まれています。それらをインターネットで公開することは非常にセンシティブな問題をはらんでいますので、極めて慎重に扱う必要があります。

(iii) 許諾条件

インターネットで公開されているウェブアーカイブの多くは発信者から許諾を得たものです。選択収集は、事前に発信者から収集・保存・公開について許諾を得た上で行いますので、バルク収集とは対照的にインターネット公開できるものが多い傾向にあります。もちろん発信者が施設内に限定しての公開を望む場合には、その条件に従わなければなりません。

また、バルク収集でも範囲が限定（例えば政府サイト限定など）されている場合には、公開に係る著作権が法的に制限されていたり、発信者から許諾を得たりするなどして、インターネット公開を実施しているウェブアーカイブもあります。

国名	組織	収集方法	公開方法
アメリカ	インターネット・アーカイブ	バルク収集	インターネット公開
アメリカ	米国議会図書館	選択収集	発信者から許諾が得られたもののみインターネット公開、その他は館内公開
イギリス	英国図書館	バルク収集（「.uk」ドメイン） 選択収集	6つの納本図書館内で公開 インターネット公開
オーストラリア	オーストラリア国立図書館	バルク収集（連邦政府サイト）・選択収集	インターネット公開
オーストリア	オーストリア国立図書館	バルク収集（「.at」ドメイン）・選択収集	館内公開
オランダ	オランダ国立図書館	選択収集	館内公開
カナダ	カナダ国立図書館・文書館	バルク収集（連邦政府サイト）・選択収集	インターネット公開
スイス	スイス国立図書館	選択収集	スイス国立図書館，カントナル図書館等の館内で公開（メタデータはインターネット検索可）
スウェーデン	スウェーデン国立図書館	バルク収集（「.se」ドメイン）・選択収集	館内公開
チェコ共和国	チェコ共和国国立図書館	バルク収集（「.cz」ドメイン） 選択収集	館内公開 発信者から許諾が得られたもののみインターネット公開、その他は館内公開
デンマーク	デンマーク王立図書館	バルク収集（「.dk」ドメイン）・選択収集	博士号以上の学位を持つ研究者のみインターネットアクセス可
日本	国立国会図書館	バルク収集（公的機関サイト）・選択収集	発信者から許諾が得られたもののみインターネット公開、その他は館内公開
ニュージーランド	ニュージーランド国立図書館	バルク収集（「.nz」ドメイン） 選択収集	非公開 インターネット公開
フィンランド	フィンランド国立図書館	バルク収集（「.fi」・「.ax」ドメイン）・選択収集	6つの納本図書館内で公開
フランス	フランス国立図書館	バルク収集（「.fr」・「.re」ドメイン）・選択収集	研究者のみ館内アクセス可

表 2：各国ウェブアーカイブの公開状況

（3）各国ウェブアーカイブの公開状況

世界各国の主なウェブアーカイブの公開状況を表 2 にまとめました。

（4）制限と利用のバランス

ウェブサイトにはありとあらゆる情報が詰まっており、その膨大な蓄積がウェブアーカイブです。それらを公開するためには、上で述べた著作権や個人情報なども含めて多様な観点から十分に検討しなければなりません。一方で、収集・保存をしても利用されなければアーカイブとしての意味がないと言うのもその通りです。

公開にあたっては、各国の事情や時代の要求に合わせつつ、制限と利用のバランスが取れた最適の公開方法を模索することが必要なのです。

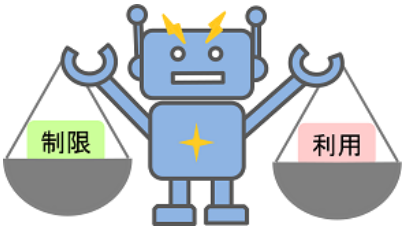


図 32：制限と利用のバランス

10. ウェブアーカイブの技術的な課題

ウェブアーカイブでは、全てのウェブコンテンツを完璧に収集できるわけではありません。収集ロボットの技術的な限界により収集が難しいコンテンツがあるためです。代表的なものとして、動的コンテンツやストリーミングファイルがあります。

（1）動的コンテンツ

データベースの中に格納され、検索を実行して初めて表示されるようなデータは、収集ロボットで収集することができません。

これらは動的コンテンツと呼ばれ、検索を実行したり画面をスクロールしたりするなど、ユーザの操作により要求（クエリ）がサーバに送信され、サーバ側のプログラムで結果が生成されてデータが返信される仕組みです。また、JavaScript を使ってクライアント側で実行して生成されるコンテンツもあります。表示される内容や URL は、クライアントの要求によって異なったものになります。

一方、html ページ、画像ファイル、文書ファイルなどが固定した URL で置かれ、誰がいつ見ても同じように表示されるものは、静的コンテンツと呼ばれます。

収集ロボットはトップページを起点としてリンクをたどりながら、URL をもとにファイルを収集していく仕組みのため、動的コンテンツは静的コンテンツに比べて収集し難いのです。

ただし、サーチエンジンが使用している収集ロボットのなかには、Java Script を実行する機能を備えたものがあると言われています。世界各国のウェブアーカイブで広く使用されている Heritrix でも、補助ツール[12]を実装することでクライアントサイドのスクリプトを実行し、動的コンテンツを収集する試みがなされています。

図 33：動的コンテンツのため収集できなかった例

KNM Gallery（京都国立博物館）

保存日：2014 年 4 月 6 日

http://warp.da.ndl.go.jp/info:ndljp/pid/8562444/gallery.kyohaku.go.jp/public/index_jp.php

12 Introduction to Umbra. https://webarchive.jira.com/wiki/display/ARIH/Introduction+to+Umbra

（2）ストリーミングファイル

動画ファイルも収集が困難なコンテンツのひとつです。近年、動画の多くはファイルをそのままウェブサイトに置くのではなく、ファイルをダウンロードしながら再生する方法で配信されています。

その配信方法には、専用のプロトコルとサーバを用いて配信する「ストリーミング」と、http プロトコルを用いてファイルをクライアント側に一時的に保存しながら再生する「プログレッシブダウンロード」の 2 種類があります。

ストリーミングを一般的な収集ロボットで収集することはできません。収集するためには、専用プロトコルを用いてデータを受信し､それを蓄積するソフトウェアを利用する必要があります。

プログレッシブダウンロードは、ダウンロード用の URL を抽出するなどして収集できる場合もあります。収集ロボットで収集するためには、ソースコードを自動的に解析してダウンロード用 URL を抽出する機能が必要となりますが、Heritrix にはそのような機能は実装されていません。動画サービスの技術仕様が頻繁に変更されるため、解析機能の仕様を固定し難いことがその理由として挙げられます。

また、動画サービスのなかには利用規約によりファイルのダウンロードを禁じているものもあり、課題は技術的な側面だけにとどまりません。

図 34：ストリーミングファイルのため収集できなかった例

山梨インターネット放送局

保存日：2014 年 2 月 6 日

http://warp.da.ndl.go.jp/info:ndljp/pid/8424198/www.pref.yamanashi.jp/webtv/

（3）新技術への常なる対応

こうした状況に対して、ウェブアーカイブも手をこまねいているわけではありません。動的コンテンツの項で紹介したように、収集ロボットの技術がより進歩して汎用的になれば、収集できるものが増えてくると期待されます。また、収集ロボットではない方法でストリーミングファイルを収集する試みもなされています[13]。

[13] Hockx-Yu, Helen et al. Capturing and replaying streaming media in a web archive：a British Library case

一方で、ウェブ技術は急速に進化しており、新たなフォーマットやプロトコル、プラットフォームが生まれています。仮に動的コンテンツやストリーミングファイルが収集できるようになったとしても、その先には新しい技術によるコンテンツが待ち受けているでしょう。ウェブアーカイブはウェブ技術の進化に常に対応し続ける必要があり、そのチャレンジが終わることはありません。

study. iPRES 2010. http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hockxyu-44.pdf

第 2 章 ウェブアーカイブをささえる技術

1. 収集ロボット Heritrix

「Heritrix」[14]は数あるクローラの中の 1 つです。クローラは、「ロボット」や「ボット」と呼ばれることもあります。

クローラとはインターネット上のウェブページを巡回し、画像や PDF ファイルなどを自動的に集めてくるプログラムのことをいいます。Google や Bing などの検索エンジンは独自に開発したクローラを使って、インターネット上の情報を収集し、それらを検索できるようにしています。

Heritrix は、インターネットアーカイブや国立国会図書館の WARP をはじめとして、大英図書館(British Library)、米国議会図書館(Library of Congress)など、世界中の国立図書館のウェブアーカイブ事業で使用されています。

(1) 特徴

Heritrix は Java 言語で開発されているオープンソースソフトウェアです。インターネットアーカイブによって開発されました。Apache License, Version 2.0 を採用していますので、ユーザはそのライセンスのもとで自由に修正、再頒布等を行うことができます。

Heritrix にはウェブベースのインタフェースがありますので、ブラウザを使っての設定や収集状況の確認ができます。

また、Heritrix は収集したウェブコンテンツをウェブアーカイブの保存用ファイルフォーマットである ARC 形式や WARC 形式で保存します。これらのフォーマットで保存されたコンテンツ

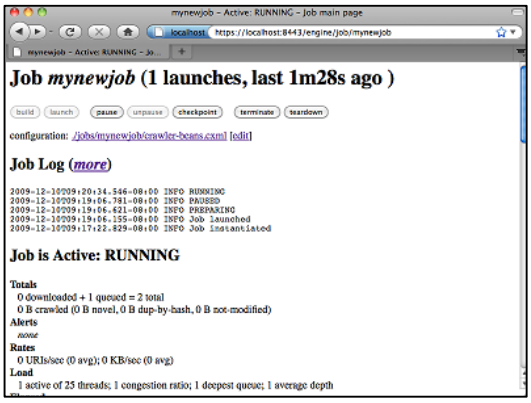


図 35：Heritrix のウェブベースインタフェース
「Heritrix 3.0 and 3.1 User Guide[15]」

[14] https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

[15] https://webarchive.jira.com/wiki/display/Heritrix/A+Quick+Guide+to+Running+Your+First+Crawl+Job

は、オープンソースソフトウェア wayback を使って閲覧することができます。

2004 年 8 月にバージョン 1.0.0 がリリースされ、2013 年 3 月時点でバージョン 3.1.1 がリリースされています。

(2) 処理フロー

Heritrix の動作の流れは「第 1 章 3.ウェブを収集するしくみ」(p.8) に概要が書かれています。ここでは、もう少し詳しく見ていくことにします。

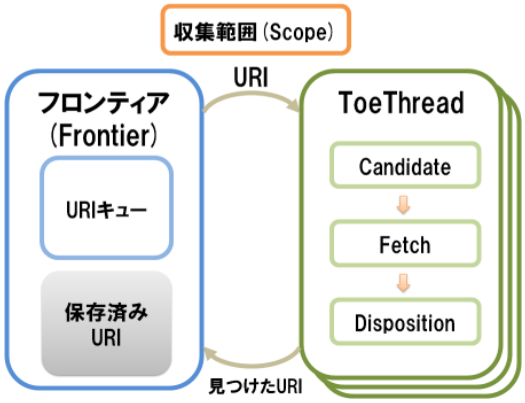


図 36：Heritrix の処理フロー

Heritrix は大きく分けて 3 つの機能からなります。図 36 にあるように、上部の収集範囲(Scope)、左側のフロンティア(Frontier)と右側の ToeThread です。

(i) 収集範囲(Scope)

収集範囲では収集する範囲を管理しています。

・ドメイン単位

・http://warp.da.ndl.go.jp/contents/ 配下の URL のみ

といったような柔軟な指定が可能です。

(ii) フロンティア(Frontier)

フロンティアでは URI の管理をしています。「URI キュー」にはまだ保存していない URI がためられています。最初は、収集する起点となる URL が登録されています。また、すでに保存した URI は「保存済み URI」にためられていきます。まず、URI キューの中から次に保存すべき URI を 1 つ選び出します。これはあらかじめ設定していた内容に従い自動的に抽出されます。選び出した URI は ToeThread へと渡され、その URI の収集が始まります。

(iii) ToeThread

ToeThread では URI が示すウェブコンテンツの保存を行います。Heritrix はマルチスレ

ッド処理に対応していますので複数の URI を同時に保存することができます。その際、それぞれの URI に対して ToeThread の処理が同時に行われることになります。広範囲にわたるドメインのアーカイブを行う場合は、何百もの URI を同時に保存することになります。ToeThread では 3 つの処理が順に行われます。それらの概要は次の通りです。

・Candidate 処理
URI にアクセスする前の処理を行います。例えば、その URI が収集する範囲内かどうかを確認します。
・Fetch 処理
URI にアクセスしてコンテンツを取得します。また、取得したコンテンツを解析し、そこに含まれているリンク(URI)を抽出します。
・Disposition 処理
取得したコンテンツをファイルに保存します。保存の際は通常 WARC 形式が使われます。また、抽出されたリンクのうちまだ保存していないものを URI キューに追加します。

※上記 ToeThread の概要は Heritrix バージョン 3 系列について説明しています。バージョン 1 系列では、処理が 5 段階に分かれています。

こうして URI キューに入っている URI がすべてなくなるまで処理が続けられます。URI キューが空になれば収集は完了です。

（3）モジュールによる機能追加

Heritrix はモジュールを追加することで機能を追加することができます。現在、以下のようなモジュールが公開されています。

・DeDuplicator：差分収集機能 [16]
・Crawl-By-Example：主題に従って収集したページを分類する機能 [17]

国立国会図書館は DeDuplicator モジュールを追加して差分収集を行っています。

（4）参考文献

Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery and Michele Kimpton. "An Introduction to Heritrix" 4th International Web Archiving Workshop (2004)
https://webarchive.jira.com/wiki/download/attachments/5441/Mohr-et-al-2004.pdf?api=v2

16 http://deduplicator.sourceforge.net/index.html
17 https://webarchive.jira.com/wiki/display/SOC06/Crawl-By-Example

2. 全文検索エンジン Solr

「Apache Solr」[18]（以下 Solr）は、Java で記述された高速検索サーバの 1 つです。

（1）特徴

CNET Networks 社により"Solar"として開発された後、2006 年 1 月に Apache コミュニティに寄贈され、"Solr"と名前を変えました。2007 年 1 月より Apache Lucene のサブプロジェクトとなっています。

以下のような特徴が挙げられます。（本稿は Solr3.x を基に記述しています。）

・高い全文検索能力
・負荷の高い Web 環境に最適化
・XML、JSON や HTTP 等のオープンスタンダード技術を基に開発
・高速なインデキシングが可能
・XML による設定や定義が可能
・AND や OR、NOT 等の演算子及びワイルドカードが利用可能
・全文検索に加えて、範囲指定や重み付け等の検索が可能
・ウェブベースのインタフェースがあり、ブラウザを使っての検索や検索結果の確認が可能

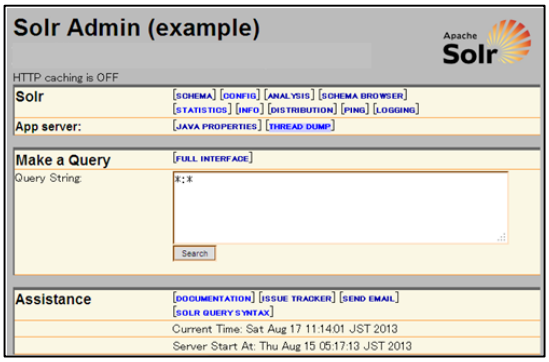


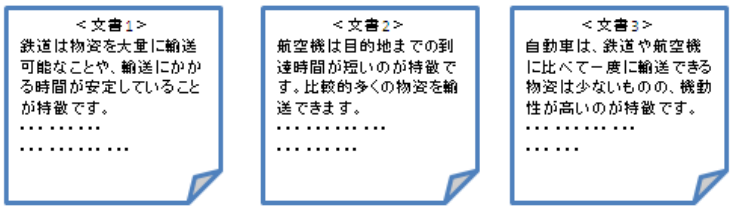
図 37：Solr のウェブベースインタフェース

（2）転置インデクス方式

Solr は、インデクス作成時に文書を解析して単語に切り分け、単語ごとにそれが含まれている文書の情報を記録しておきます。このようなインデクスを転置インデクスと言います（図 38）。

あらかじめ転置インデクスを作成しておくことで、ある単語で検索した時に、その単語がどの文書に含まれているのかをすぐに探しだすことができます。インデクスの作成には時間を要しますが、一旦インデクスを作成した後は、高速に検索できるのが利点です。

[18] http://lucene.apache.org/solr/



〈インデクスデータ〉

単語	文書		
鉄道	1	3	
航空機	2	3	
自動車	3		
物資	1	2	3
目的地	2		
時間	1	2	
…	…	…	…

図 38：転置インデクスの例

(2) 解析方法

インデクスを作成する際に必要となるのが文書の解析です。以下、主に日本語文章の解析に用いられる形態素解析と N-Gram 法について説明をします。システムへの実装においては、それぞれの特徴を生かしながら、両者を組み合わせて使用することも可能です。

(i) 形態素解析

「形態素」とは、言語において意味のある最小の単位を意味します。形態素解析は、辞書などを用い文脈や単語などを解析し、「形態素」に区切ってインデクスを作成する方法です。後述する N-Gram に比べて、生成されるインデクスのサイズが小さいのが特徴です。また、検索時に意図しない結果がヒットすることが少ない半面、検索漏れが発生することがあります。

◇解析対象文

「国立国会図書館は、出版物を中心に国内外の資料・情報を広く収集しています」

◇解析結果

1	2	3	4	5	6	7	8	9	10
国立	国会	図書館	は	出版	物	を	中心	に	国内外
11	12	13	14	15	16	17	18	19	
の	資料	情報	を	広	く	収集	して	います	

図 39：形態素解析の例

（ii）N-Gram

N-Gram は、文脈や単語を考慮せず、文字数の単位（N=2 の場合は 2 文字毎）で文字を分解する方法です。検索に漏れが発生しない一方で、検索結果にノイズが多くなったり、インデクスのサイズが肥大化したりする傾向があります。

◇解析対象文

「国立国会図書館は、出版物を中心に国内外の資料・情報を広く収集しています」

◇解析結果

1	2	3	4	5	6	7	8	9	10
国立	立国	国会	会図	図書	書館	館は	は出	出版	版物
11	12	13	14	15	16	17	18	19	20
物を	を中	中心	心に	に国	国内	内外	外の	の資	資料
21	22	23	24	25	26	27	28	29	30
料情	情報	報を	を広	広く	く収	収集	集し	して	てい
31	32								
いま	ます								

図 40：N-Gram(N=2)解析の例

（3）適合度（Score）

Solr は、検索結果を「適合度（Score）」によって順序付けます。適合度の算出は、「tf-idf」と呼ばれる方法で、文章中の単語の重み付けを行います。「tf-idf」とは、単語の出現頻度 TF（Term Frequency）と、全文書中の単語の集中度合い IDF（Inverse Document Frequency）を掛け合わせることにより適合度を算出する方法です。

TFとIDF

	TF					IDF
単語	文書1	文書2	文書3	文書4	文書5	
鉄道	3	0	0	0	1	0.397
航空機	0	1	0	0	0	0.698
自動車	0	0	3	1	0	0.397
物資	2	1	1	0	5	0.096

（IDF=log(全文書数 ÷ 単語の出現する文書数)）

適合度

適合度(TF×IDF)					
単語	文書1	文書2	文書3	文書4	文書5
鉄道	1.191	0	0	0	0.397
航空機	0	0.698	0	0	0
自動車	0	0	1.191	0.397	0
物資	0.192	0.096	0.096	0	0.48

図 41：tf-idf による適合度

例えば、「鉄道」と「物資」の二つの単語に注目してみましょう。「鉄道」は文書 1 に 3 回出現し、他は文書 5 に 1 回のみ出現します。一方、「物資」は文書 5 に 5 回出現しますが、他の文書にも多く出現しています。文書 1「鉄道」と、文書 5「物資」を比較すると、出現回数は、

5（文書 5「物資」）＞3（文書 1「鉄道」）

ですが、適合度は以下のように逆転します。

1.191（文書 1「鉄道」）＞0.48（文書 5「物資」）

また、「鉄道」と「物資」で掛け合わせ検索をした際の適合度は、合算して、

1.383（文書 1）＞0.877（文書 5）

となり、文書 1 の適合度が高くなります。

Solr で検索をした際には、この適合度の順で、検索結果の表示を並び替えることができます。

3. 保存用ファイルフォーマット WARC

ウェブアーカイブでは、ウェブページから収集したファイルをそのまま保存するのではなく、ウェブアーカイブに適した保存用ファイルフォーマットにして保存します。その理由としては、収集時の情報やファイルのメタデータが同時に保存できるため長期保存対策が可能であること、差分収集に対応しているフォーマットであることなどが挙げられます。

(1) WARC ファイルとは

WARC は世界のウェブアーカイブ機関で広く採用されている保存用ファイルフォーマットで、その名称は「Web Archiving」に由来します。IIPC の主要メンバーであるインターネットアカイブが採用していたファイルフォーマット ARC をもとに、2004 年に IIPC により汎用的に使える形式に拡張されました。2009 年 5 月には、国際標準機構(ISO)の国際規格 ISO 28500:2009 となっています [19, 20]。

WARC 形式で保存されたファイルは、そのままではブラウザで閲覧することはできません。オリジナルのサイトと同じように表示するためには、Wayback などの WARC 形式に対応したツールが必要です。

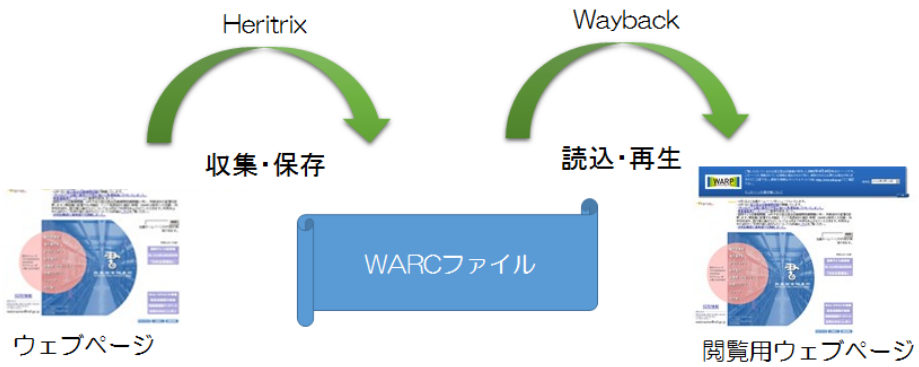


図 42：収集から再生までのイメージ

(2) WARC ファイルの構造

WARC 形式のファイルは、1 つあるいは複数の「WARC レコード」で構成されます（図 43)。「WARC レコード」は、「WARC レコードヘッダー」と「コンテンツブロック」のセットから成っています。「WARC レコードヘッダー」には、WARC のバージョン及び「WARC フィールド」が格納され、「WARC フィールド」にはレコード ID やレコードタイプ、ファイル（コンテンツ）の収集先や収集日、ファイルのサイズなどの情報が収められています（表 3)。「コンテンツブロック」には収集したファイルそのものが格納されます。

---

19 IIPC のウェブアーカイブ保存形式"WARC"が ISO 規格に. http://current.ndl.go.jp/node/13149

20 ISO 28500:2009. Information and documentation : WARC file format. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717

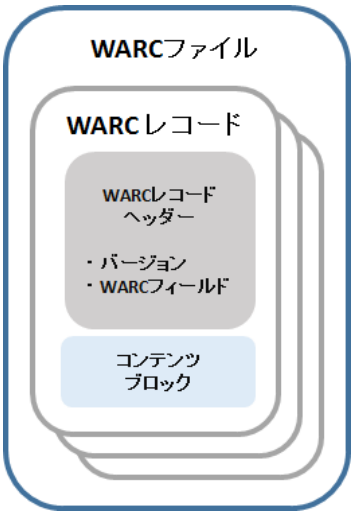


図 43：WARC ファイルの構造

Name Field	項目	内容(サンプル)
WARC-Type	レコードタイプ	warcinfo、response、request、metadata など
WARC-Date	ウェブアーカイブ収集日時	YYYY-MM-DDThh:mm:ssZ 2013-08-05T16:49:22Z
WARC-Filename	WARCファイル名	レコードタイプがwarcinfoの場合のみ
WARC-Record-ID	レコードID	<urn:uuid:b3d73c5c-839c-945f5d6672d7>
Content-Type	レコードブロックの MIME Type	text/html
Content-Length	レコードブロック長	バイト長
WARC-IP-Address	コンテンツ取得先の IPアドレス	xxx.xxx.xxx.xxx
WARC-Target-URI	取得先のURI	dns:www.libraryikaruga.jp
WARC-Payload-Digest	コンテンツのハッシュ値	sha1:WQ3NILTDFV2BAC4CN2HBRETN3

表 3：WARC フィールドに格納されている主な情報

(3) ウェブページの保存

ウェブサイトの各ページは、html ファイルや画像ファイル、文書ファイルや Java スクリプトなど、複数の URL（ファイル）によって構成されています。Heritrix などのクローラを用いてウェブサイトを収集する場合、これら URL の単位で収集を行います。そしてそれらを WARC 形式のファイルとして保存する際には、URL ごとに複数の「WARC レコード」が作成されます。

基本的に、1 つの URL に対して以下の 3 つの「WARC レコード」が作成されます。

・Request レコード
・Response レコード
・Metadata レコード

Request レコードには該当 URL を収集した際の情報が、Response レコードにはファイルそのものが格納されます。Metadata レコードには、URL のメタデータ情報が格納されます。

例えば図 44 の左のようなウェブページに対しては、右のようなレコードが作成されます。

WARC 形式のファイルを閲覧する際には、URL ごとに格納されたこれらの情報を Wayback などのツールで読み解くことで、ウェブページを元の形で再生することができるのです。

WARC ファイルのサイズは、1GB 以下に抑えることが推奨されています。このため、WARP では 100MB を目安として WARC ファイルを分割して格納しています。

また、ストレージ領域の削減のために WARC ファイルは圧縮することが推奨されています。GZIP による圧縮が推奨されています。

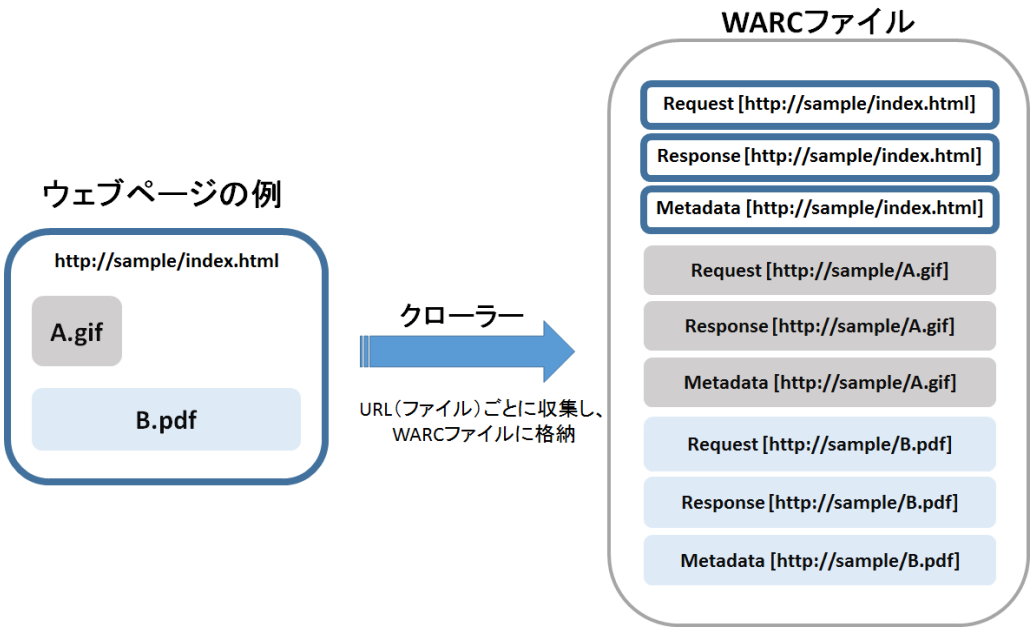


図 44：WARC ファイルの格納例

4. 閲覧アプリケーション Wayback

Wayback は WARC フォーマットで保存されたウェブアーカイブを表示するためのツールです。米国の非営利団体インターネットアーカイブは収集したウェブサイトを閲覧できるよう Wayback Machine というサービスを提供しており、そのオープンソース版が Wayback です。インターネットアーカイブを中心として開発が進められてきました。

(1) 特徴

Wayback を使うと、Heritrix などのクローラで収集された WARC フォーマットのアーカイブを表示することができます。つまり、収集時点のウェブサイトをブラウザで閲覧することができます。なお、WARC フォーマットが国際規格化される以前に使われていた ARC フォーマットにも対応しています。

Wayback は Java 言語で開発されているオープンソースソフトウェアです。ソフトウェアライセンスとして Apache License, Version 2.0 を採用していますので、ユーザはそのライセンスのもとで自由に修正、再頒布等を行うことができます。動作にあたっては、UNIX 系の OS、Java Runtime Environment 1.5 以上、Apache Tomcat 6.0 がシステム要件です。

2005 年にバージョン 0.2 が公開され、2013 年 12 月時点でバージョン 1.6.0 が公開されています。現在、インターネットアーカイブに代わり IIPC が中心となり、2014 年中の公開に向けバージョン 2.0.0 の開発が進められているところです [21]。このページの説明はバージョン 1.6.0 を基にしています。

(2) 機能

(i) 概要

Wayback は大きく 2 つの機能を持っています。1 つ目は、クローラが収集したアーカイブファイルから URL と収集日時を取り出しインデクスを作成するインデキシング機能です。2 つ目は、閲覧要求に応じてアーカイブを表示する機能です。

- インデキシング機能
  - インデキシング機能：アーカイブからインデクスを作成する機能
- 表示機能
  - 表示機能：アーカイブをブラウザへ表示させる機能
  - アクセス制御機能：アクセス制限を提供する機能
  - UI カスタマイズ機能：検索結果の表示形式などを変更する機能

図 45 はこれらの機能を図示したものです。

[21] http://netpreserve.org/about-us/news/iipc-re-launches-open-source-wayback

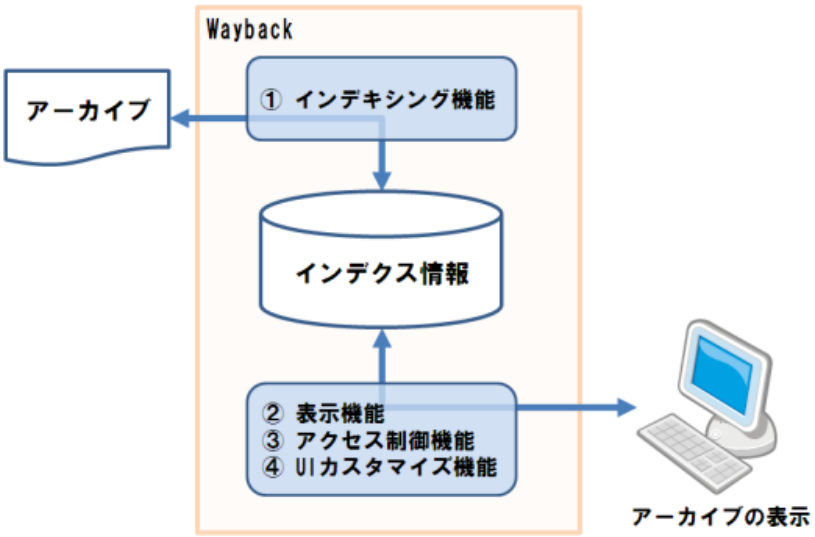


図 45：Wayback の機能

(ii) インデキシング機能

Wayback は事前にアーカイブファイルから URL と収集日付を取り出し、インデクスを作成します。ウェブサイトのアーカイブは容量が TB（テラバイト）～PB（ペタバイト）と非常に大きくなるのが特徴的です。表示の度ごとに規模の大きなアーカイブを直接探すのでは非効率的ですので、インデクスを作成することで効率的にアーカイブの中から要求のあった日時のデータを抽出することができるようにします。

インデキシング処理の流れは図 46 の通りです。

処理内容は、大きく 2 つに分けられます。1 つは、アーカイブの格納場所をロケーションデータベースへ登録する処理です。もう 1 つは、アーカイブからインデクスを作成しリソースインデクスと呼ばれるデータベースへ登録する処理です。その際、作られるインデクスは CDX ファイルと呼ばれます。

図 46 の各処理の概要は表 4 の通りです。

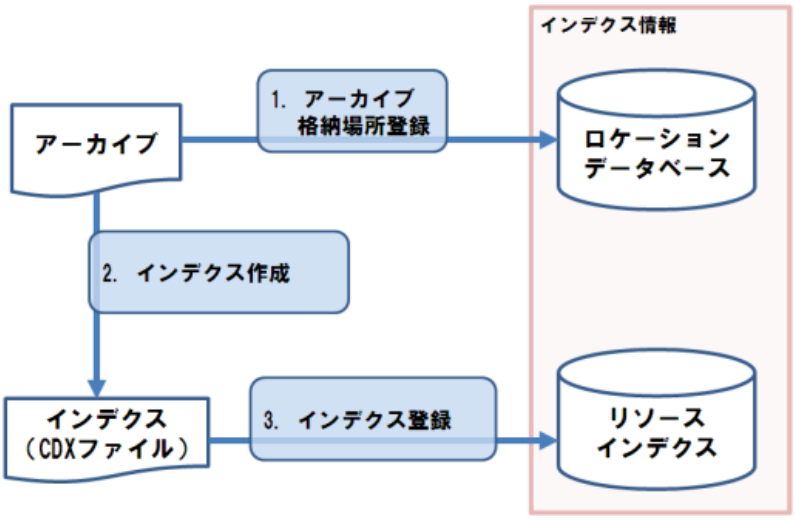


図 46：インデキシング処理の流れ

処理名	概要
1. アーカイブ格納場所登録	ウェブサイトの収集が終わりアーカイブが作成されると、アーカイブのファイル名とその格納場所をロケーションデータベースへ登録します。
2. インデクス作成	アーカイブから CDX ファイルを作成します。CDX ファイルは URL、収集日時そしてその URL が含まれているアーカイブファイル名などがまとめられたファイルです[22]。
3. インデクス登録	CDX ファイルの内容をリソースインデクスに登録します。

表 4：インデクス機能

（ii）表示機能

Wayback の表示機能により、ブラウザでアーカイブを閲覧することができます。表示機能には 3 つの表示モードがあり、いずれかを選択します。

表示モード	概要
1. Archival URL Replay モード	・専用 URL でアーカイブを表示します。 ・アーカイブページ内のリンクを専用 URL に書き換えて表示します。 ・専用 URL でアクセスするため、JavaScript などの動作が正常に再現されない場合があります。
2. Proxy Replay モード	・ブラウザのプロキシ設定に Wayback サーバを指定することで、元の URL でアーカイブを表示できます。 ・元の URL でアクセスするため、JavaScript の問題は起こりません。
3. DomainPrefix Replay モード	・DNS を設定することで、ブラウザのプロキシ設定なしに Proxy Replay モードと同様に元の URL で表示できます。 ・実験的なモードです。

表 5：表示機能

[22] http://archive.org/web/researcher/cdx_file_format.php

以下、ウェブアーカイブ機関で多く用いられている「1. Archival URL Replay モード」について説明します。

アーカイブを表示する際、ブラウザからは次の Archival URL Replay モード専用の URL を利用します。

http://HOSTNAME:PORT/CONTEXT/ACCESS-POINT/TIMESTAMP/URL

URL の各項目は次の通りです。

・HOSTNAME：Wayback が動作しているホスト名
・PORT：アクセスポート番号（80 番なら省略可）
・CONTEXT：Wayback をデプロイしたときのコンテキスト名（ROOT の場合省略）
・ACCESS-POINT（アクセスポイント）：異なる設定を適用するために付与する名前（省略可）
・TIMESTAMP：検索対象の日付（年月日時分秒）
・URL：検索対象ページの URL

Wayback Machine を例に、実際の URL を見てみます。

http://web.archive.org/web/20130204110456/http://warp.da.ndl.go.jp/

この URL にアクセスすると 2013 年 2 月 4 日 11 時 4 分 56 秒に保存された http://warp.da.ndl.go.jp/が表示されます。また、表示ページの HTML ソースを表示すると、リンクがすべて専用 URL に書き換わっていることが分かります。

なお、日付の指定は完全に一致している必要はありません。指定した日付に近い日のアーカイブが表示されるようになっています。

（iii）アクセス制御機能

Wayback はページとアクセスポイントに対してアクセス制御を行うことが可能です。アクセス制御は表 6 の通り 4 種類あります。これ以外のアクセス制御はできません。例えば、各ページに対して、ユーザごとにアクセス制御を設定することはできません。

項番	制御対象	対象者	制御方法
1	ページ	利用者全員	アーカイブ元の robots.txt による制限
2		利用者全員	URL 指定による制限
3	アクセスポイント	指定した IP アドレス	IP アドレスによる制限
4		指定したユーザ	ユーザ認証による制限

表 6：アクセス制御機能

(a) ページに対する制限（robots.txt）

この制限を適用すると、利用者が閲覧を要求したサイトの元サイトに robots.txt が設置されている場合に、その robots.txt に記載されている内容に従ってアクセス制御が行われます。また、保存時点で robots.txt が設置されていない場合でも、その後 robots.txt が設置されれば、アーカイブの閲覧時にはその robots.txt に基づいたアクセス制御が行われます。

(b) ページに対する制限（URL 指定）

設定ファイルにアクセス制限の対象とする URL を記述することで、該当ページの表示を制限することができます。

(c) アクセスポイントに対する制限（IP アドレス）

設定ファイルにアクセスを許可する IP アドレスを記述することで、該当 IP アドレス以外からのアクセスを禁止することができます。

(d) アクセスポイントに対する制限（ユーザ認証）

設定ファイルに対象とするアクセスポイントを記述し、BASIC 認証による制限ができます。ユーザ名及びパスワードはあらかじめ設定ファイルに追加しておく必要があります。

(e) UI カスタマイズ機能

ユーザインタフェース（UI）の一部をカスタマイズすることができます。カスタマイズできる UI は表 7 の通りです。

カテゴリー	概要
Query UI	Wayback はアーカイブを URL で検索することができます。その際、検索結果表示を次の 3 パターンの中から選択することができます。 ・カレンダー形式 ・一覧表示形式 ・xml 形式
Replay Inserts	アーカイブの表示画面に、コメントやバナー、デバッグ用メッセージなどを表示することができます。また、設定ファイルを変更することで、表示内容をカスタマイズすることができます。
Exception	アーカイブの表示時に発生するエラー画面をカスタマイズすることができます。
Localization	利用者の利用言語（ブラウザの設定）に応じて、検索結果表示画面やエラー画面の言語を切り替えることができます。ただし、切り替えるためには、その言語用の設定ファイルを用意する必要があります。初期は英語が設定されています。なお、日本語の設定ファイルは用意されていません。

表 7：UI カスタマイズ機能

ウェブアーカイブのしくみ

発行日　平成 26 年 10 月 15 日
編集・発行　国立国会図書館関西館電子図書館課
〒619-0287 京都府相楽郡精華町精華台 8-1-3