資料1

諸外国におけるウェブ情報アーカイブに関する調査報告の概要

2004年10月27日

UFJ総合研究所

市民生活室　広石拓司

内容

1. ウェブ情報アーカイブの背景と意義
2. 諸外国におけるウェブ情報アーカイブの取り組み状況
3. 諸外国のウェブ情報アーカイブの現状と課題、今後のわが国における推進の課題

1. ウェブ情報アーカイブの背景と意義

ウェブ情報アーカイブの定義と背景

- **本調査におけるウェブ情報アーカイブの定義**
  - ウェブサイトの情報を、長期保存を目的として、収集・蓄積・保存・（最終的には）公開すること
- **ウェブ情報アーカイブの背景**
  - 世界で推計7億人*が使うメディア
  - 情報量は推計で24PB**　（米国議会図書館 26TB）
  - 現代社会の情報環境を反映（多様な内容・形式・発信者）

➡ **後世の研究者が現代・これからの社会を研究するのに不可欠な情報源として計画的な保存が必要**

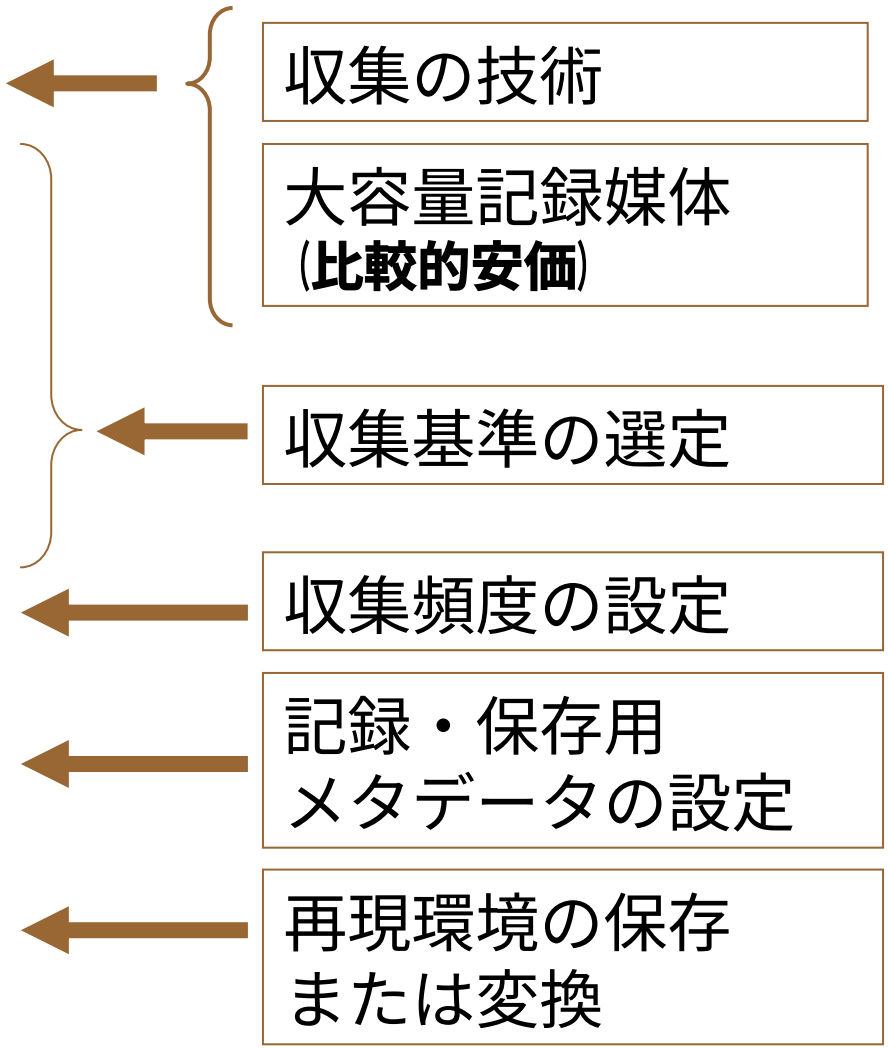
* 2002年末 「インターネット白書2003」より
** UFJ総研推計

保存対象としてのウェブ情報の特徴

- **実体がない**
  物理的な実体がなく、本のように「モノ」として保管・維持できない。
- **各主体が自ら情報発信の担い手である**
  出版社など決められた選別・編集プロセスがない。
  限りない数の発行者が対象になる。内容の質のばらつきが大きい。
- **更新が頻繁にされている**
  完成形の定義、初版などのバージョンの同定ができない。
- **データの種類・形式が多様**
  情報の構成要素、組み合わせ方が多様である。
- **リンク構造が情報の本質の一部を担っている**
  WEBサイトの範囲、情報のまとまりの単位を定義することが難しい。
- **国籍がない、多様な言語**

ウェブ情報の特徴への対応策

- **数・情報量が莫大**
- **内容の質の差が大きい**
- **「サイト（情報のまとまり）」の定義が困難**
- **発行者の国の定義が困難**
- **内容の形式が多様**
- **更新が頻繁にされる**
  （完成形や第○版という定義が不可能）
- **発行の管理者がいない**
- **再現するためにソフト、プログラムが不可欠**



2．諸外国におけるウェブ情報アーカイブの取り組み状況

- 米国、オーストラリア、スウェーデン、デンマーク、フィンランド、イギリス、フランス
- 上記以外の国でのＷＥＢアーカイブ・電子出版物保存の事例
- 関連する国際プロジェクト

国名	機関	ﾌﾟﾛｼﾞｪｸﾄ名	収集手法	収集根拠の視点	情報収集対象	公開方法
米国	インターネット･アーカイブ		バルク	Fair Use	公開されている全て。ロボット除外サイト、登録拒否サイトは非対象	ネット上で公開
米国	議会図書館	MINERVA	選択	契約	職員選定サイト、特定サイトを収集。(特定の範囲のバルク収集も)	館内のみ
オーストラリア	国立図書館	PANDORA	選択	契約	自国文化を代表するサイトに関する基準を満たすものを選択	ネット上で公開
スウェーデン	王立図書館	Kulturarw3	バルク	納本制度	国名ドメイン、国内住所登録などのサイト	館内のみ
デンマーク	王立図書館		選択･納本	納本制度	納本制度に基づく発行申請を基に、重要度の高いサイトを選択	館内のみ
フィンランド	国立図書館		バルク	納本制度	国名ドメイン、.comなどのうち自国関連のもの	未検討
イギリス	The National Archive	UK Central Government Web Archive	選択	公文書館	51政府サイトを選択し、収集	ネット上で公開
フランス	国立図書館		選択	納本･契約	仏語サイトなどをロボット収集後、歴史的な資料性が高いと判断されたサイトを重点的に収集	非公開

インターネット・アーカイブ

- http://www.archive.org
- インターネット上に存在し、自由にアクセスできるウェブ情報の網羅的収集（バルク収集）
  （将来的価値は今判断できないので、できるだけ集める）
- 収集データ：約1PB（月20TBで増加）　（2004年夏）
- アレクサ・インターネットが社ロボット収集で集めたウェブ情報を提供されている
- パスワードが必要なサイト、個人情報を扱うサイト、所有者が登録を拒否したサイト（opt-out）などは収集しない
- ネット上の映像、音楽などのアーカイブ収集にも取り組む
- The Wayback Machine で、資源に誰もがアクセス可能

米国議会図書館：MINERVA

- http://www.loc.gov/minerva/
- 大統領選挙、９．１１同時多発テロ、イラク戦争など特定テーマを扱い、一定の選択基準を満たし、自由にアクセスできるサイトを図書館員が選択し、収集
  （収集はI.Aが協力）
- 情報量：約4万サイト
- サイト権利保有者と許諾協定を結び、公開条件を個別設定。権利保有者からの削除要求も受付ける。
- メタデータ：XMLベースのMODS

オーストラリア国立図書館：PANDORA

- http://pandora.nla.gov.au
- オーストラリアに関する情報、オーストラリアの社会・文化・科学・経済等に関するオーストラリア人の著作など国家にとって重要なサイトの選定基準を設け、その基準にあうサイトを図書館員が選択し、州立図書館と連携して収集
- 独自のアーカイビング・システム（PANDAS）を利用
- 情報量：５６６GB
- 収集にあたっては出版者の許諾を得る。アクセス制限、公開時期の設定の予防があれば個別に契約をする。
- 収集情報は目録化され、図書など他の情報資源と組み合わせて探し出せるように図書館目録などに登録

スウェーデン王立図書館：Kulturarw³

- http://www.kb.se./kw3/
- スウェーデンに関するオンライン公開情報の網羅的収集
  .se サイト、.com などで登記情報が自国内のものなどの基準を設け、ハイパーリンクをたどってバルク収集
- 情報量：約6TB
- 北欧諸国の共同プロジェクトNWAの収集ツールを基に、図書館で独自開発
- 王立図書館内のみで利用可能
- ネット情報の収集・図書館内での公開に関する法令が2002年に出され、それを法的根拠とする

デンマーク王立図書館

- http://www.pligtaflevering.dk
- 納本制度の対象が「媒体を問わずデンマークで発行された著作物」とされたことを受け、静的サイトを納本制度の対象となった。著作物の発行者が王立図書館に通知し、図書館で審査した上で収集
- 情報量：２３GB
- 収集の基点、対象範囲は発行者の通知に基づく。
  原本性を保証。登録情報の変更・削除は認めない。
- 収集情報は目録が作成され、王立図書館閲覧室で利用可
- 納本制度の徹底が課題（アーカイブされたサブドメインは0.1%程度）。バルク収集についても検討中。

フィンランド国立図書館

- 納本法の視点から、自国の文化資源を保存し、将来必要になった時のために備えて保存。
  ウェブを納本法の対象とする改正法施行は2005年春の予定
- 情報量：1TB弱（圧縮後）
- .fi、.com等でサイト情報収集団体の調査結果でフィンランドのものとわかったものをバルク収集。
  収集した情報は、ハッシュ関数によって重複がないかチェックされたものを圧縮して保存
- 現在は公開情報のみ。新しい納本法では収集時のアクセス用パスワードの提供など深層ウェブへの対応も検討中
- 公開はされていない

イギリス：UK Central Government Web Archive

- http://www.pro.gov.uk/webarchive/
- The National Archives（記録・写本省）が、ウェブ情報アーカイブへの実験的試行として2500ある中央政府のウェブサイトのうち、代表的な51を選定して、ロボット収集
- 技術はI.Aの協力によって実施
- 法的納本図書館法の2003年改正で電子出版物も法廷納本の対象とし、ウェブサイトについても研究レベルに到達しているものをThe National Archivesが収集することが定められた。

フランス国立図書館

- 将来世代に文化生産物の代表的なアーカイブを残すことを目的として取り組む。
- .fr、.comのうち登記住所が国内のものを対象に、ロボット収集をし、リンク数、用語の利用頻度などから機械的に重み付けをしたものから、図書館員が納本制度の視点から後世に残す価値を判断して対象を設定。
  深層ウェブについても、図書館員が高い重要性のあると判断したものは、個別に連絡して収集協力を依頼。
- 現在、研究開発段階

その他の国における取り組み

- **ノルウェー国立図書館　：Paradigma**

  URL　：http://www.nb.no/paradigma/

  2001年8月より3年間のプロジェクトとして技術的な課題、納本制度を研究

- **オーストリア国立図書館　：AOLA**

  URL　：http://www.ifs.tuwien.ac.at/~aola/

  1998年からWEBアーカイブを実験。2000年から2001年にかけて、2.1万サイト、270万ページのアーカイブを実施

- **チェコスロバキア国立図書館・Masaryk大学**
  ：charakteristika webarchivu

  URL　：http://webarchiv.nkp.cz/

  2000年から2001年にWEBアーカイブを試行した。

電子出版物アーカイブへの取り組み

- **オランダ国立図書館** (KB)
  - EU**の電子情報保存に関する** NEDLIB**プロジェクトの主幹**
  - e-Depot：**デジタル出版物のアーカイブの先進的システム**
  - **しかし、ウェブ情報アーカイブの本格的な取り組みは行っていない。**
    - **定型メタデータがなく、メタデータの処理が困難**
    - **何を収集するか・しないかという選択の議論が必要**
- **ドイツ国立図書館** (Die Deutsche Bibliothek)
  - DEPOSIT.DDB.DE. ：**オンライン出版物ならびにデジタル化された出版物を保存**
- **中国国立図書館**
  - **電子出版物の保存**
- **カナダ国立図書館**
  - Canada Electronic Collection**において、電子出版物の収集**

デジタル生産物保存の国際プロジェクト

- Electronic Resource Preservation and Access Network（ERPANET)
  - 2001年11月から3年間、欧州の図書館などによる文化遺産と科学的デジタル生産物の保存のための情報交換と知識の基盤を構築
- Networked European Deposit Library (NEDLIB)
  - 1998年～2001年に、EU予算により、欧州各国の8図書館を含む11機関と出版社3社が、電子出版物のデジタル納本に関する技術と運営手法を共同研究開発。
  - ウェブ情報アーカイブについては、NEDLIB Harvester（収集ツール）を開発
- Open Archive Information System（OAIS参照モデル）
  - ISOによるデジタル情報の長期保存に関する共通用語と概念に関する国際標準。ISO14721として、国際標準規格化されている。
- PREservation Metadata: Implementation Strategies（PREMIS）
  - OCLC（Online Computer Library Center：84カ国の図書館ネットワークからなる非営利研究活動団体）などによる電子情報の長期保存のためのメタデータの勧告と良い実践例を研究するプロジェクト。2003年6月～2004年6月

ウェブ情報アーカイブ国際共同プロジェクト

- Nordic Web Archive

  デンマーク、フィンランド、アイスランド、ノルウェー、スウェーデンの国立図書館によるウェブ情報アーカイブのためのツール（the NWA toolset）の共同開発プロジェクト。（2000年9月～2002年6月：予算総額約25万ユーロ）。
  2003年3月から、NWA IIがスタート。

- International Internet Preservation Consortium（IIPC）
  - インターネット・アーカイブの呼びかけに、11の国立図書館（米国議会図書館、オーストラリア図書館、フランス国立図書館、英国図書館、NWA参加図書館 等）が参加。
  - 共通の目的と標準の開発、収集とアクセスのためのオープンソース・ツールの開発、共通ツールを基盤にした各国ウェブコレクションの構築
  - 2003年から3年間の期間限定の協働
  - 予算（計画）：クロール費用として1百万ドル／年。また、3年間の開発費用としてツール開発：50万ドル、ソフト技術：60万ドル、マネジメント：45万ドル
  - Heritrixと呼ばれる新世代ウェブアーカイブツールの開発。（Heritrixは、オープンソースアプリケーションとなる予定）

3．諸外国の現状と課題、今後の日本での推進の課題

取り組みの現状と課題（1）

- ウェブ情報アーカイブの基本的理念の整備が進む
- ウェブ情報アーカイブ用システムの基本的な形が整備されてきている。ただし、深層ＷＥＢ、動的コンテンツ、メタデータ添付への対応に課題が残る
- 収集対象・範囲の設定、選択方法は、未だ模索中
- 法定納本制度によるデジタル情報の収集・保存・公開について、北欧を中心に法制度の整備が進む
- メタデータなど国際的な標準化に関する活動が増えてきている。

取り組みの現状と課題（2）

- 収集・保存に努力が集中し、長期的な保存・アクセス、活用のあり方という課題への取り組みは遅れている
  - 表示再現に必要なシステム・プログラムの収集・保存
  - エミュレーション、マイグレーションの手法の開発
  - 厳格な保存、定期的なチェックやレンダリング
  - 著作権の課題から、活用への取り組みが遅れている
- 「オンライン出版物の納本」の視点が中心になり、商用サイト、広告、掲示板など“使われているサイト”が対象外に
- パートナーシップの重要性が高まっている。
  - European Conference on Digital Libraries（ECDL）Workshop on Web Archiving

ウェブ情報アーカイブと著作権

- 3つの論点

  (1) 収集、(2) 公共のアクセス、(3) 保存

- 欧州を中心に納本制度を根拠とする動き
  - オンライン出版物を対象とする納本の制定法、適切な法的プロセスを有する国

    北欧4カ国、イギリス (本格運用は今後)
  - 電子出版物の納本制度はあるが、オンライン出版物が対象となっていない国

    フランス、オーストリア
- オーストラリアは発行者との契約
- I.A：Fair Use思想を基本に、opt-out、個人情報の除外

メタデータの状況

機関	収集方法	資源管理のためのメタデータ	資源利用のためのメタデータ
I.A (米)	ﾊﾞﾙｸ	対象サイトの URL、日付、アーカイブバージョンの組みあわせ	なし (URL 検索、全文検索で対応)
MINERVA (米)	選択	MODS (一部 METS を試行)	MODS
オーストラリア	選択	説明情報、許可情報、タイプ、状態、テーマ･コレクション名、規制情報、収集スケジュール	ダブリン･コア
スウェーデン	ﾊﾞﾙｸ	収集過程、サーバ情報、対象物に関する情報を含む独自ファイルネーム	なし (基盤ページからのリンクで閲覧可能)
デンマーク	選択	納本制度に基づく発行者の通知情報	ダブリン･コア
フィンランド	選択	url、収集日、サーバ応答記録、MD5によるハッシュ情報、タイムスタンプ	ダブリン･コア
イギリス	選択	e-GMS2 (英国電子政府政策全般のメタデータ標準)	なし (I.A のシステムを利用)
フランス	選択	site-delta (URL、収集日の組からなる永続的識別子)	なし (非公開のため)

＊ 収集時間･歴史的変遷を管理する時間パラメータについては未だ検討中

今後の日本における開発のポイント

- WEBアーカイブの目的について十分な議論を
  - 現在、WEBサイトは文化、コミュニケーション、生活の記録、デジタルアートの発表の場など幅広い役割を担っている
    WEBアーカイブの実施主体、目的は多様化していくものと考えられる。
- アーカイブ間のリンクが重要に
  - 収集物の意味付けや整理のコストの面から、各アーカイブが精度高く収集するためには、地域や実施主体にとっての重要性など何らかの選択が必要となる。アーカイブ間のリンクについて配慮する必要がある。
  - 国際的な動向との連携を常に考慮して取り組む必要がある。
- 共同プロジェクトの重視を
  - 深層ウェブまでを対象とした収集技術、長期保存、再現性確保のための環境整備などの共同プロジェクト