国立国会図書館
National Diet Library, Japan

海外におけるウェブアーカイブの現状とその利活用

国立国会図書館 関西館 電子図書館課
志村　努

NDLデジタルライブラリーカフェ　2020/12/10

目次

1. 世界のウェブアーカイブ
   - IIPC
2. 引用文献における利活用
3. 二次的データセット
   - 研究利用のためのデータ提供
   - 英国図書館の事例
   - The Archives Unleashed プロジェクト

1. 世界のウェブアーカイブ
IIPC (International Internet Preservation Consortium)

- ウェブアーカイブの国際的なコンソーシアム（2003年設立）
- 56機関が参加(2020年12月現在)
  - 国立図書館、アーカイブ機関、研究機関など
- 年に一度の総会
- 活動内容
  - 技術開発とツールの無償頒布
    - クローラ (heritrix), 閲覧ソフト (openwayback)
  - 共同収集、横断的な検索・閲覧の研究、利活用の検討

世界のウェブアーカイブ https://warp.da.ndl.go.jp/contents/reccommend/world_wa/index.html

IIPCホームページ https://netpreserve.org/

1. 世界のウェブアーカイブ IIPC 活動内容の変化

- 初期（2003-2006）は技術的な課題の克服
  - クローラ・閲覧ソフトの開発、WARCファイルフォーマットのISO化
  - オープンソースとして公開され、多数の機関が採用、標準化
- 現在は学術研究における利活用が議論の中心
  - バルク収集の多くは非公開 ＝ 一般利用は極めて限定的
  - 研究・分析で利用してもらうことで、新たな価値を生み出す
  - コンテンツデータは巨大過ぎて扱いが難しいため、二次的なデータセットを提供

前田直俊. ウェブアーカイブの利活用に向けた動き-世界の潮流とWARPの取組-. カレントアウェアネス. 2017, (331), CA1893, p. 9-13.
https://current.ndl.go.jp/ca1893
http://doi.org/10.11501/10317594

2. 引用文献における利活用
引用文献のリンク切れ、内容変化

- 引用文献のリンク切れ
  - 1997年から2012年の科学、技術、医学系論文を対象に引用文献のリンク切れを調査
  - 調査対象のリンク数：100万件以上
  - 調査結果：

出版年	arXiv	Elsevier	PMC
2012	13%	22%	14%
1997	34%	66%	80%

- 引用文献の内容変化
  - 調査対象の論文は同上
  - 調査対象のリンク数：約24万件
  - 調査結果：75%のリンク先が内容変化

Klein M, Van de Sompel H, Sanderson R, Shankar H, Balakireva L, et al. (2014) Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. PLoS ONE 9(12): e115253. doi:10.1371/journal.pone.0115253

Jones SM, Van de Sompel H, Shankar H, Klein M, Tobin R, Grover C (2016) Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content. PLoS ONE 11(12): e0167475. doi:10.1371/journal.pone.0167475

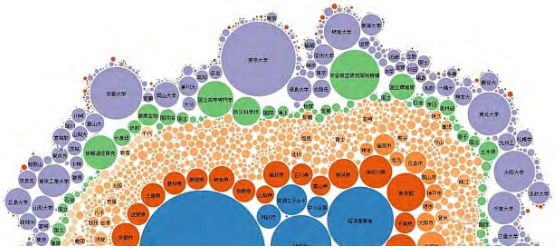
2. 引用文献における利活用
引用でのウェブアーカイブの活用

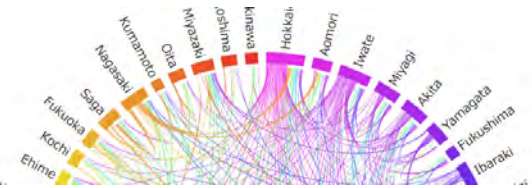
- Perma.cc（https://perma.cc/）
  - 法律分野における引用文献の保存サービス
  - ハーバード大学が開発、主に米国の法律関係図書館が参加
  - 引用文献のリンク切れ、内容変化を防ぐため、引用者自らウェブコンテンツを指定・保存し、永続的アクセスを保証する
- Wikipediaリンク切れ
  - *“Internet Archive（IA）によるWikipediaのリンク切れをウェブアーカイブされた情報に置き換える作業、900万件を突破”*
https://current.ndl.go.jp/node/36758

3. 二次的データセット
研究利用のためのデータ提供

- バルク収集の多くは非公開 ＝ 一般利用は極めて限定的
  - 選択収集においては権利処理を行うため公開の割合が高い
- 研究・分析で利用してもらうことで、新たな価値を生み出す
  - データ可視化、リンク解析など

特色あるコレクション
https://warp.da.ndl.go.jp/contents/recommend/collection/index.html





- コンテンツデータは巨大過ぎて扱いが難しいため、二次的なデータセットを提供
  - ファイルフォーマット情報、リンク情報など
- 分析用ツール開発も行われている

3. 二次的データセット 英国図書館の事例

- ３つのウェブアーカイブコレクション
  - Selective Archive (2004-)：選択収集
  - Legal Deposit Archive (2013-)：バルク収集
  - JISC UK Web Domain Dataset (1996-2013)
    ：Internet Archiveが収集した.ukサイト
- 二次的なデータセットを公開
  - Selective Archive (2004-)
    - Classification dataset
  - JISC UK Web Domain Dataset (1996-2013)
    - Format Profile
    - Geoindex
    - Host Link Graph
    - Crawled URL Index

UK Web Archive Open Data
https://data.webarchive.org.uk/opendata/

3. 二次的データセット
英国図書館の事例：データセット

- Classification dataset：主題、タイトル、URL

Arts & Humanities	Architecture	68 Dean Street	http://www.sixty8.com/
Arts & Humanities	Architecture	Abandoned Communities	http://www.abandonedcommunities.co.uk/

- Format Profile：ファイルフォーマット

application/x-csv	text/plain	application/octet-stream	2009	1

- Geoindex：テキスト中のpostcode

20080509162138/http://uk.eurogate.co.uk/contact_us	IG8 8HD
20080509162231/http://www.toolfastdirect.co.uk/acatalog/cable_Reels_and_Extensions_240_Volt.html	ML2 7UR

- Host Link Graph：ホストレベルのリンク関係

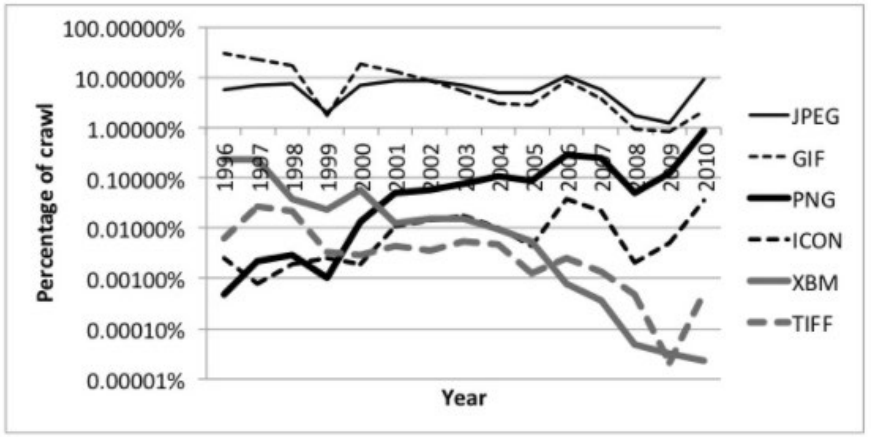
1996	appserver.ed.ac.uk	portico.bl.uk	1
1996	blaiseweb.bl.uk	blaiseweb.bl.uk	4

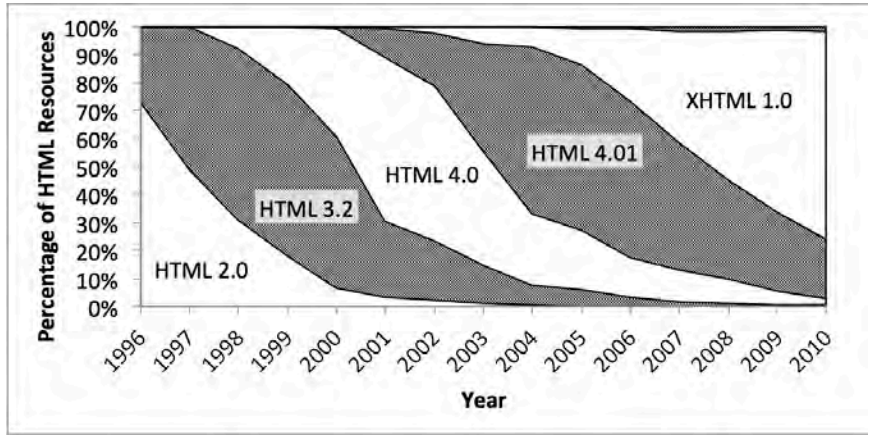
- Crawled URL Index：URL、収集日時、ファイルタイプ、ステータスコード、ハッシュ値…

vanguard.ntu.ac.uk/	19961018104851	http://vanguard.ntu.ac.uk:80/	text/html	200	2TAC6RS2DMTHHFVWCSDHNL6W6RIIOQIV	-	34954008	DOTUK-HISTORICAL-1996-2010-GROUP-AA-XABEGS-20110428000000-00000.arc.gz

3. 二次的データセット
英国図書館の事例：分析例

- UKドメインにおけるフォーマット利用率の時系列変化
  - 対象データ：JISC UK Web Domain Dataset のうち1996-2010
  - 分析内容：収集コンテンツに占める 画像フォーマットの割合、HTML バージョン、PDFバージョン/作成ソフトの割合の時系列変化
  - 結果（一部）：下図



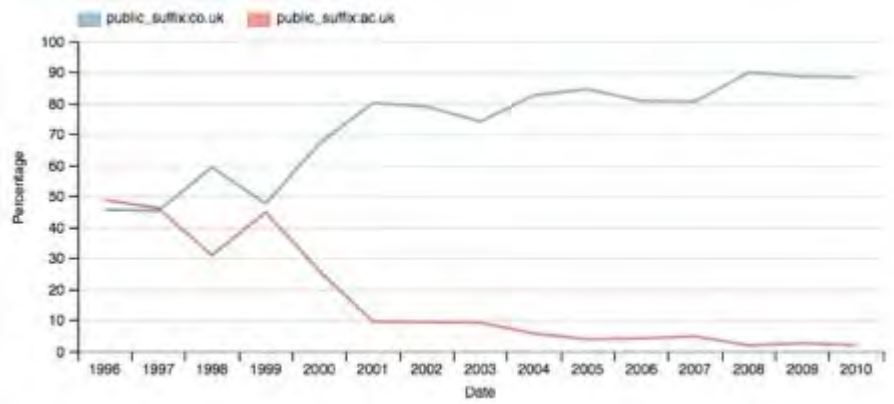


Andrew N. Jackson(2012) Formats over Time: Exploring UK Web History arXiv:1210.1714

https://www.slideshare.net/andrewnjackson/formats-over-time-exploring-uk-web-history-14601282

3. 二次的データセット
英国図書館の事例：ツール開発

- Big UK domain data for Arts and Humanities プロジェクト
  - ビッグデータに関するプロジェクト。ロンドン大学のInstitute of Historical Researchや英国図書館などが参加。
  - *https://buddah.projects.history.ac.uk/*
- UK Web Domain Dataset (1996-2013)を利用
  - 65TB、35億URL以上（画像含む）
- 開発されたツール SHINE
  - 検索語の出現頻度比較
  - 検索例：co.uk hosts(青)と
    ac.uk hosts(赤)の時系列変化
  - *https://www.webarchive.org.uk/shine*



https://anjackson.net/2015/04/27/what-have-we-saved-iipc-ga-2015/

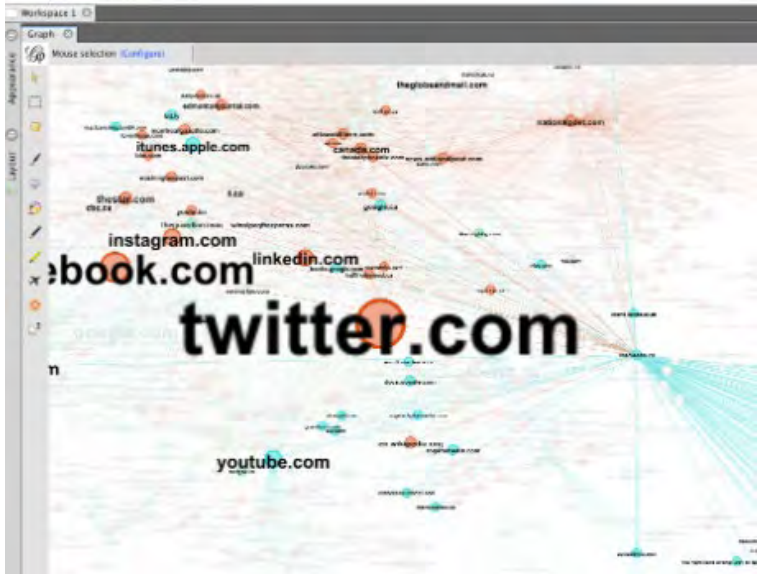
3. 二次的データセット
The Archives Unleashed プロジェクト

- ウェブアーカイブを用いた研究における課題
  - データが巨大であるため分析ツールが必要
  - プログラミングや可視化のためにコンピュータの知識が必要
- The Archives Unleashed Project の目的
  - ユーザフレンドリーなツールの開発
    - the Archives Unleashed Toolkit
    - the Archives Unleashed Cloud
  - 研究者のコミュニティ構築
    - Datathonの実施

The Archives Unleashed Project (https://archivesunleashed.org/)

3. 二次的データセット
the Archives Unleashed Toolkit の概要

- 学術研究におけるワークフローに沿った機能開発
  - フィルター (Filter)
    - 例：特定ドメイン、ファイルタイプ(html, 画像など)、特定文字列
  - 抽出 (Extract)
    - 例：テキスト抽出
  - 集計 (Aggregate)
    - 例：ファイルタイプ別のカウント、
      各ページへのリンク数
  - 可視化 (Visualize)
    - 例：表形式、Gephi(可視化ツール)対応形式



Archives Unleashed Toolkit
(https://aut.docs.archivesunleashed.org/)

https://aut.docs.archivesunleashed.org/docs/toolkit-walkthrough

まとめ

1. 世界のウェブアーカイブ
   - IIPCの活動内容の変化
     - 技術開発 → 利活用
2. 利活用：引用文献
   - 永続的なアクセスを保証
3. 利活用：二次的データセット
   - 研究利用のためのデータ提供
     - 扱いやすい二次的データ
   - 分析用ツール開発
     - 研究者自身が抽出データを選定