UDC 007：930.25(0.068)：621.39：681.327.03：002.5：02

ボーンデジタル時代におけるウェブアーカイブとその活用基盤としての Socio-Sense

喜連川 優，豊 田 正 史，田 村 孝 之，鍛 治 伸 裕*

ウェブアーカイブに関して，2008 年 1 月 23 日国立国会図書館において開催された「ウェブアーカイビングの現在と展望―国際連携に向けて―」と題したシンポジウムで紹介されたインターネットアーカイブならびにヨーロピアンアーカイブの最新の動向に関して述べる。加えて，東京大学では，蓄積に加えて多様な分析・解析を目的とした『Socio-Sense』システムを開発してきており，その動機ならびに過去を遡るツールに関し，事例を用いて紹介する。ウェブアーカイブの多様な価値についても論ずる。

キーワード：ウェブアーカイブ，インターネットアーカイブ，ウェブ，クローリング，ブログ

1. はじめに

ボーンデジタル，すなわち，情報生成時点でデジタル媒体のみを想定して生まれた情報が今世紀に入り圧倒的に多くなってきた。そして，いわゆる，UCC（user created contents），あるいは CGM（consumer generated media）が大きく世界に影響を与えるようになった今日，ウェブは文化形成の舞台とも言え，その捕捉と永続的継承は極めて大きな意義を有する。ユネスコは，とりわけ，ボーンデジタル情報は脆く消失されやすいことから，2003 年，Charter on the Preservation of the Digital Heritage（デジタル遺産の保存に関する憲章）を採択している。ウェブアーカイブに関しては，国会図書館より国際動向と自身の取組みについて詳細な報告がなされており[1]，本稿ではその後の動き，ならびにわれわれの研究活動について紹介を行う。

2. ウェブアーカイブ

2.1 インターネットアーカイブとヨーロピアンアーカイブ

2008 年 1 月 23 日国立国会図書館において，「ウェブアーカイビングの現在と展望―国際連携に向けて―」と題したシンポジウムが開催され，インターネットアーカイブウェブアーカイブ部門のディレクターである Kris Carpenter 氏ならびに，ヨーロピアンアーカイブ ディレクターである Julien Masanes 氏より講演があり，その後，筆者の一人である喜連川がパネルチェアを仰せつかり，お二人と「ウェブアーカイブとデジタル化の国際的な潮流」についてパネル討論を行った。会場の定員 300 名に対し，事前申込みの打ち切りが早々になされるなど，シンポジウムは大変盛況であった。お二人はデジタル時代の文化とも言えるウェブのアーカイビングは人類の活動の歴史を記録する上でも大変重要であり，各国が各々のウェブ文化の保存に着手し，互いに協調すべきであることを強く訴えた。ボーンデジタルメディアであるウェブは極めて消失しやすいことから，今始めることが大切であると力説した。

インターネットアーカイブ（IA）[2]は Brewster Kahle によって 1996 年に設立され，現在に至るまで，ウェブページに止まらず，音楽，ビデオ，本，ソフトウエアまで種々のアーカイブを進めている。全体の容量は約 3 ペタバイト程度であり，その中で，2 ペタバイトがウェブアーカイブである。総ページ数は 100 億を超え，毎月 15–20 テラバイトのページを Alexa 等他の機関から譲り受けるなどして，常に成長をしている。IA 自身もクローリングをしている。収集対象は世界各国に及び，37 もの言語に及んでいる。最近は，本のスキャンによるアーカイブ化により多くのマンパワーを投入しているとのことである。現在，IA はカリフォルニア州立図書館とみなされている。膨大なデータの長期間保存はそれ自身が大きなチャレンジでもあり，IA は PetaBox なる低消費電力を特徴とするディスクストレージ装置を自らデザインしたり，天災や政治的攻撃など種々の事情でコンテンツの内容が失われる可能性があることに配慮し，複製を地理的に離れた場所に保存すること，いわゆるディザスタリカバリに実践的に取り組み，エジプトのアレクサンドリアにその一部を格納するなど，先進的な IT システム技術を駆使している。アーカイブには Wayback Machine なるソフトウエアが用意されており，URL を入力すると当該 URL の過去のページを年代順に閲覧することができる。

ヨーロピアンアーカイブは，本格的活動はこれからであり，現在，LiWA（Living Web Archives）[3]なるプロジェクトを EU として推進しようとしている。ここで Living とは単にコンテンツをフリーズすることを目的とするのではなく，進化するメディアに対して適切なアクセス手段を長期にわたって提供しようと試みている。スパムなどのノイズ除去もその研究対象に含まれている。

＊きつれがわ まさる，とよだ まさし，たむら たかゆき，かじ のぶひろ
東京大学 生産技術研究所 戦略情報融合国際研究センタ
〒153-8505 東京都目黒区駒場 4-6-1
Tel. 03-5452-6254
（原稿受領 2008.6.10）

その他の国のウェブアーカイブに関する動向は，文献 1)に詳細が取り纏められている。

2.2 IIPC（International Internet Preservation Consortium）[4]

インターネットアーカイブはウェブアーカイブの先鞭をつけその構築に大きな貢献をしたが，爆発的なウェブ空間の膨張に追いつくことは到底困難であり，一つの組織が全ウェブ空間の完全な収集をするには限界があることから，それぞれの国家が自国のアーカイブを進め，互いに連携することが妥当な解となるとの考えから，2003 年 7 月，オーストラリア，カナダ，デンマーク，フィンランド，フランス，アイスランド，イタリア，ノルウェー，スウェーデンの国立図書館，ブリティッシュライブラリ，米国議会図書館，インターネットアーカイブが集まり，IIPC なるコンソーシアムが結成された。本原稿執筆時点で 38 の組織が加盟している。

IIPC の活動は大きくは，ウェブページのクロール，永続保存，アーカイブデータのアクセス環境の提供の三つに分けることができるが，クロールなど主要な機能に関してはプログラムツールを開発・提供し，ウェブアーカイブ活動を開始する際の手助けをしている。

3. Socio-Sense

筆者らは，1999 年より日本語で書かれたホームページに特化した日本語ウェブアーカイブの構築を進めてきた。図 1 に，その蓄積量の推移を示す。現在おおむね 100 億ページのアーカイブとなっている。インターネットアーカイブが 1,000 億ページ程度有するのに対しては 10 分の 1 程度でしかないが，一国の言語に特化したアーカイブとしてはかなり大きな規模と言える。2003 年より文部科学省リーディングプロジェクトの支援を受け，Socio-Sense と名付けたウェブ解析システムの開発を進めてきた[7]。IA や IIPC 等，ウェブアーカイブ機関の主たる目的はウェブページの収集と保存に有るのに対し，筆者らは，むしろ，蓄積されたアーカイブを利用した利用価値を追求すべく，多様な解析・分析ツールを開発してきた。

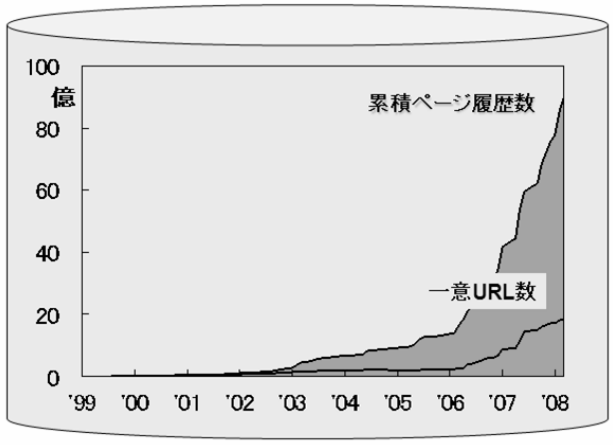

図 1 Socio-Sense 日本語ウェブアーカイブの容量推移

図 1 には 2 本の曲線が描かれているが，これらは収集した全ウェブページの数，ならびに，収集対象である URL の数を示す。すなわち，内容が高頻度で変化する場合には，当該ページを短い周期で何度も収集するのに対し，内容の変化がほとんどない場合には，収集頻度を低下させている。このような可変周期のページ収集により総ページ蓄積量に比べて，収集 URL 数は少なくなっている[5]。

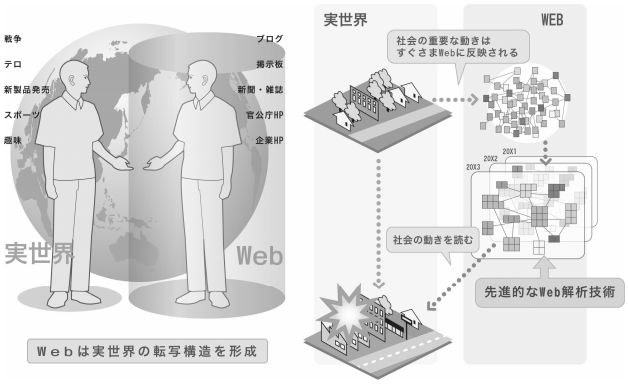

図 2 Socio-Sense（社会のセンサーとしてのウェブ）

図 2 に Socio-Sense なるシステムをわれわれが開発してきた動機を示す。多くの社会現象，すなわち，テロや地震のような事件，新製品の発表，あるいは，スポーツの勝敗など，社会の重要な情報は殆どすべて，しかも，瞬時にウェブに反映されるに到っている。このことから，ウェブは社会が投影されたもの，あるいは，ウェブは社会のセンサーと見做すことができると考えた。いったん，ウェブ上に社会現象を記述する情報が載せられると，当該空間は計算機処理可能であり，多様なマイニングによりある程度社会の動きを把握できるのではないか，あるいは，若干の予測が可能かもしれないと考え，Socio-Sense と名付けた次第である。すなわち，Socio-Sense は社会の動きを分析・解析するためのエンジンである。

ウェブアーカイブの利用において最大の特徴となる点は過去からの推移を把握可能となるという点である。現行のサーチエンジンは最新のスナップショットに関しての検索は可能であるものの，過去のコンテンツに関しての検索は不能である。図 3 は生協の白石さんという本にまでもなった大変ユーモラスな人物についてのブログ上での言及がどのように推移したかを示している。図 3(a)は 7 月時点での様子を示しているが，図中の一つの四角は一つの言及を示している。左中央にある「生協の白石さん」を紹介するページに対して極めて多くの言及がなされており，当該ページが話題の中心，いわゆるインフルエンサーであることが明らかであることが判る。約 3 週間前に遡ってその状況を示したのが図 3(b)であり，さらに 3 週間前が図 3(c)である。図 3(c)は図 3(a)において話題の中心となっているページそのものが発生した時点であり，当然のことながら，言及は極めて少なく，その後，わずか 1 月強で，極めて多くの注目を集めたことが判る。このようにサイバー世界において

CGM は大きな影響を与えることが理解できる。さらにさかのぼるならば，実は図 3(d)に示されるブログが最初に生協の白石さんを紹介した話題の震源地であったことまでさかのぼることができる。このように，歴史を紐解くことができることは大変魅力的である。本稿では，メディアが紙であるためスナップショットを紙芝居的に示さざるを得ないが，実際には，連続的な可視化が可能であり，サイバー空間のタイムマシンを実現することができる[6]。

なお，日本語のウェブアーカイブに関しては，国立国会図書館による WARP なるプロジェクトにより，限られたドメインに対して，蓄積がなされている。詳細は(1)を参照されたい。

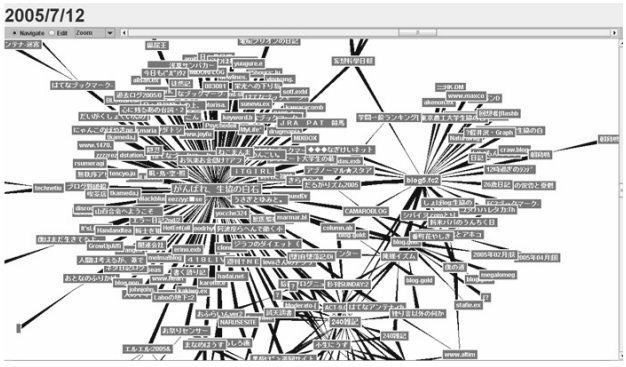

図 3(a)　爆発的人気になった「生協の白石さん」を紹介するページ

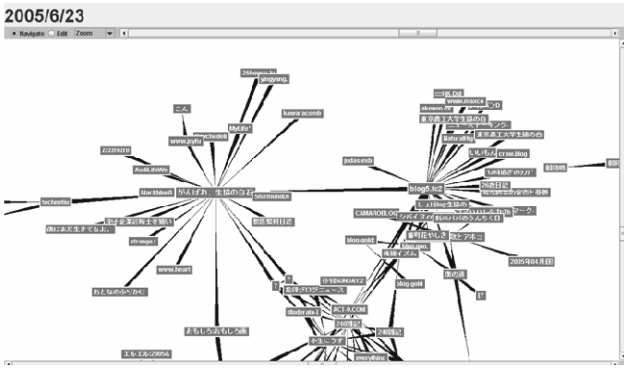

図 3(b)　「生協の白石さん」を紹介するページが次第に認知される段階

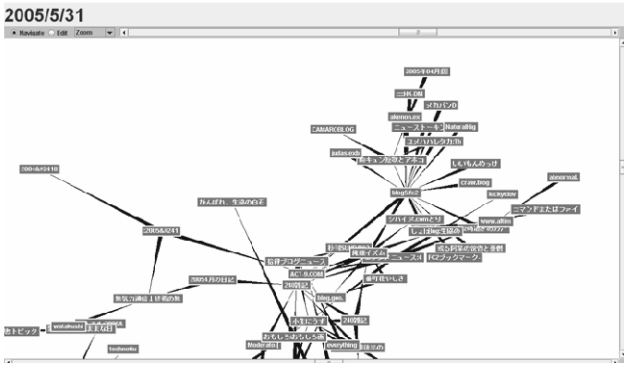

図 3(c)　「生協の白石さん」を専門に紹介するページの出現

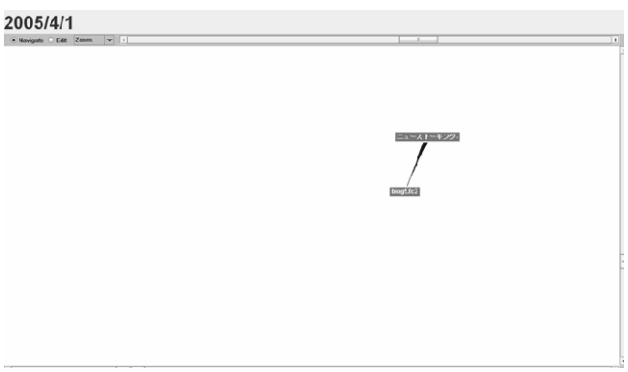

図 3(d)　生協の白石さんを最初に紹介したブログの出現

3. ウェブアーカイブの重要性

ウェブアーカイブは多様な価値を有している。

3.1　文化の保存

すでに述べたように，多くの人間活動がウェブに反映されており，ウェブ空間自体がデジタル社会の文化とも見做せよう。

3.2　学術基盤として

社会学者にとって，ウェブアーカイブは重要な研究資源となりつつある。筆者らはお茶の水大学ジェンダー研究センター舘教授と共同研究を行い，99 年に施行された男女共同参画社会基本法がどのように社会に展開していくか，あるいは，ジェンダー問題に関し，その受け止められ方が時代とともにどのように変化しているかについて研究をした[8]。言語学の観点からも言語資源としての活用が期待される。筆者らは，ブログ等のメディアでは辞書にない新しい造語が頻繁に用いられることに着目し，新造語抽出の研究を進めてきた。新造語には一過的なものと定着するものとが存在するが，アーカイブを利用することにより，新造語の時間的な普及の度合いを確認することができる。また，ググるという言葉は当初，Google を利用するという意味で使われ始めたが，近年では，その意味から汎化がなされ，サーチエンジンを利用するという意味で利用されることが増えている[9]。こうした現象もアーカイブの利用によって観測できると考えている。社会学，言語学以外の学術研究の利用も広く利用されると予見される。

3.3　情報処理分野の研究基盤として

言うまでもないが，ウェブデータはサーチエンジンを始めとする情報処理分野の種々研究において貴重な研究資源と言える。今までには利用できなかった膨大なテキスト情報が利用可能となり，自然言語処理の研究では，センチメント解析，機械翻訳を始め，すでに大きな変革をもたらしつつある。今後，動画が広く利用されるにつれて，ビジョン研究者の対象としても重要な資源となることは間違いない。

価値あるページの保存には異論はないものの，不良コンテンツに関してはその蓄積の意義が問われることが多い。しかし，検索ランキングに付随するスパムの研究には極めて重要な資源とも言える。時代と共に変化する多様なスパ

ム活動を把握するには，一件無駄にも見えるウェブ空間の収集も望まれる。ウェブ資源の広い潜在的価値を示唆する事例と言えよう。

3.4 マーケティングなどビジネスを指向した研究・利用の基盤として

専修大学新井教授ならびに電通は筆者らが開発したツールを利用し，消費材の市場投入にともなうブログの反応を詳細に解析して，社会の受け止められ方の分析において極めて有用であることを確認した。文献 10)は，マーケティングなる学問的観点からの解析であるが，ブログの時系列解析はすでに種々商用利用されている。新製品開発のヒントにも有効であることが確認されている。また，種々のトラブルが与える社会への影響についての解析にも利用されるなど，価値は高い。

4. その他の課題

コピーライトに関して, シンポジウムでも質問が続いた。マサネス氏は大変興味深い指摘をした。ユネスコによると，1 年間に出版される本は 100 万冊とのことで，仮に，1 冊 1,000 ページとすると 10^9 ページ印刷されることとなる。これに対してウェブのページ数は，曖昧なところもあるがざっくりと 10^{15} 程度はあると想定され，これを単純に比較すると百万倍の差がある。すなわち，コピーライトについての法律は旧来のメディアを対象として設けられたものであるが，現時点においては，ボーンデジタルなページが旧来の出版物のそれに比べて圧倒的に大きくなってきており，百万分の 1 の媒体に対してのルールをウェブに適用するというよりも新しく考えなおすことも必要ではないかと発言した。共感を覚えた方々も少なからずいたと感じる。わが国では，サーチエンジンに関連して著作権制限に関するパブコメが文化庁よりとられた[11]。インターネットアーカイブも現時点ではグレイな運用となっているが，米国ではフェアユース法があることからより柔軟な対応が可能である。ヨーロッパは国によって事情が異なるなど，乗り越えていかねばならない課題も多い。

また，ウェブアーカイブを実施する組織についても，その利用使途が潜在的に広いことから，妥当な機関がただ一つに決められるものでもない。国会図書館など保存を得意とする組織と，次々に生まれる新しいメディアを効率良く収集し，解析機能を開発する研究組織との連携が不可欠であろう。

5. おわりに

ウェブというメディアは社会に大きな影響を与えた。100 年後に計算機の歴史を振り返ったとしても，ウェブのインパクトは他の計算機技術に比して見劣りすることはないと想像される。ウェブは情報発信を限りなく容易にし，社会を大きく変えてきた。ボーンデジタルメディアとしてのウェブは人類の新しいデジタル文化の中心的存在であり，その蓄積と活用は，誰の目からも見ても，極めて重要であることは間違いない。

参 考 文 献

1) 廣瀬信己．ウェブ情報のデジタルアーカイブ：WARP を中心に．情報管理．2005，vol.47，no.11，p.721-732.
2) インターネットアーカイブ（Internet Archive）：http://www.archive.org/ [accessed 2008-06-09].
3) LiWA - Living Web Archives : http://www.liwa-project.eu/ [accessed 2008-06-09].
4) インターナショナル・インターネット・プリザベーション・コンソーシアム：http://www.netpreserve.org/
5) 田村孝之，喜連川優．大規模 Web アーカイブ更新クローラにおけるスケジューリング手法の評価．電子情報通信学会論文誌．2008，vol.J91-D，no.3，p.551-559.
6) Masashi Toyoda and Masaru Kitsuregawa. A System for Visualizing and Analyzing the Evolution of the Web with a Time Series of Graphs. Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia (Hypertext 05), 2005, p.151-160.
7) 喜連川優．文部科学省リーディングプロジェクト e-Society 基盤ソフトウェアの総合開発：先進的なウェブ解析技術の開発. http://cif.iis.u-tokyo.ac.jp/e-society/ [accessed 2008-06-09].
8) 小山直子，増永良文．Web マイニングツールを用いたジェンダー関連 Web コミュニティの通時的分析．DBSJ Letters. 2004，vol. 3，no.3，p.21-24.
9) 福島健一，鍛治伸裕，喜連川優．機械学習を用いたカタカナ用言の獲得. 言語処理学会第 13 回年次大会. 2007, p.815-818.
10) 馬渡一浩，富田英裕（株式会社電通），新井範子（専修大学），豊田正史, 鍛治伸裕, 喜連川優（東京大学）. ブログからレピュテーション分析の可能性を探る－Web2.0 時代の新しい方法論へのトライアルー．日本広報学会 第 12 回研究発表大会予稿集<統一論題＞メディアの変化と広報の近未来．2006，p.60-63.
11) 「文化審議会著作権分科会法制問題小委員会中間まとめ」に関する意見募集の実施について：案件番号 185000283，2007

Special feature: Web archiving: the present and challenges. Web archiving and Socio-Sense.

Masaru KITSUREGAWA, Masashi TOYODA, Takayuki TAMURA, Nobuhiro KAJI (Center for Information Fusion, Institute of Industrial Science, The Univ. of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505 JAPAN)

Abstract: National Diet Library held Symposium on Web Archive on Jan 23rd, 2008. This article introduces the current status on the activities of Internet Archive and European Archive. In addition, Socio-Sense system which is currently being developed at Univ. of Tokyo is also introduced. Tools to show the historical overview which run over the archive are briefly described with examples. Web archive is very valuable for various kinds of researches.

Keywords: web archive / internet archive / web / crawling / blog