ウェブページの寿命: 2001 年に存在した 1000 万ページを対象にした調査

宮田洋輔（常葉大学短期大学部） m@miyay.org
安形輝（亜細亜大学），池内淳（筑波大学）
石田栄美（九州大学），上田修一（前慶應義塾大学）

抄録

ウェブページの寿命を明らかにするため，2001 年に収集された 1,000 万件のウェブページを対象に，長期的な生存調査を実施するとともに，Internet Archive を利用して，ウェブページがいつ誕生し，いつ消失したのかに関する調査を行った。生存調査の結果，9 割以上のページが収集から 12 年を経た現在，消失していることが明らかになった。寿命調査の結果，2001 年に収集され現在消失しているウェブページの平均寿命は，1,108.2 日であることが明らかになった。

1. はじめに

ウェブは，公開後も容易に更新でき，削除できるなど，従来のメディアと異なる動的な特性を持っている。その関心のもとに，これまで，ウェブがどのように変化するか，あるいは，どのように消えていくかといった研究が行われてきた。

Internet Archive の創設者である Kahle は，1997 年時点で，ウェブページの寿命は 44 日であると述べている[1)]。ウェブページがどの程度の期間でアクセスできるか／できなくなるかについても実証的な研究が行われている。Koeher は 1996 年に収集した 361 件の URL について，4 年にわたり，ウェブページの生存や変化について調査を行い，2 年間で約半数のウェブページが消失することを明らかにした[2)]。最近の研究では，2007 年 2 月に存在していた，白人の権利擁護団体のウェブサイト 163 件を対象とした調査でも，2.4 年で半減期を迎えることが明らかになっている[3)]。

本研究グループでは，2001 年に収集された 1000 万件のウェブページを対象として，2003 年に生存調査（以下 2003 年調査）を行なった[4)]。その結果，Koeher による研究と同程度に，2 年を経ると 47%のページがアクセスできなくなるという結果が得られた。さらにその結果を用いて，2003 年当時の jp ドメインのウェブページ数を推計した。

ウェブの消失に関する異なるアプローチとして，Ainsworth らは，どれだけがウェブアーカイブに保存されているかを調査した[5)]。Ainsworth らは DMOZ，Delicious，Bitly，検索エンジンのインデックスから取得した URL 計 4,000 件を用いて，2010 年 11 月から 2011 年 1 月にかけて調査を行った。その結果，サンプル集合によって傾向は異なるものの，ウェブの 35%～90%程度はいずれか 1 つのウェブアーカイブにアーカイブされていることを明らかにした。

ウェブページの生存に関する既往研究では，ウェブページを取得した日付を基盤として，ウェブページの消失までの期間が調査されてきた。ウェブページが「誕生」した日を正確に取得することは困難であり，取得日を「誕生日」に近似して扱うほかない。一方で，取得日をウェブページの誕生日と設定することによって，実際の誕生から取得までの期間が除外されてしまい，消失までの期間が実際よりも短く評価されてしまう可能性がある。

さて，Ainsworth らの結果を踏まえると，ウェブページの多くはウェブアーカイブ中に存在している。ウェブアーカイブは，調査のための取得日よりも誕生に近い日時にウェブページを取得している可能性がある。そこで，ウェブアーカイブが最初に取得した日付を利用することでよりウェブページの誕生日に近い日付を得ることができる可能性がある。この関係を図 1 に示した。

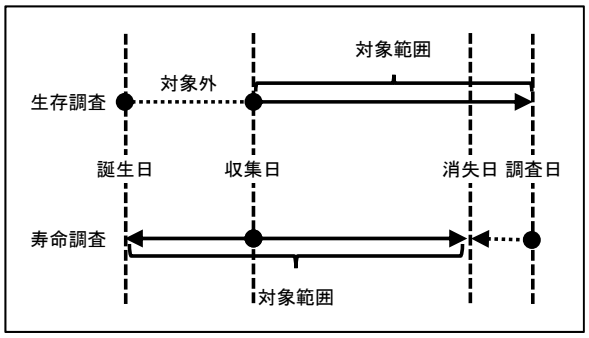

図 1 生存調査と寿命調査

そこで，本研究では，2003 年調査と同じデータ集合を対象として，前回の調査から 10 年後という長期的視点から大規模な生存調査を実施するとともに，Internet Archive を利用して，ウェブページがいつ誕生し，いつ消失したのかに関する調査を実施した。本研究の目的は，2 つの調査を組み合わせることで

ウェブページの寿命を明らかにすることである。

2. 調査方法

ウェブページの寿命を明らかにするために，調査対象のウェブページが，1)現在アクセスできる状態にあるのか，2)Internet Archive でいつからいつまでアーカイブされているかを調査した。

2.1. 調査対象

今回の調査にも，2003 年調査と同じく，NTCIR-3 Web タスクのために収集・構築された 100GB の文書集合（NW100G-01）を用いた。NW100G-01 は，NII のホームページを起点として，2001 年 8 月 29 日から 11 月 12 日にかけて（46 日間），収集された文書群の部分集合であり，1,100 万件以上のウェブページによって構成されている[6]。NW100G-01 から，jp ドメインのページのみを対象に無作為抽出した 1,000 万件のウェブページを調査対象とした。

2001 年頃の jp ドメインのページ数は，6,507 万ページと推定されており[7]，1,000 万件のウェブページは，当時の jp ドメイン全体の約 1/6 を占める文書集合であった。

2.2. 生存調査

調査対象となる 1,000 万件のウェブページについて，生存調査を行った。既に述べたように，筆者らは 2003 年に同一のウェブページ群を対象とした生存調査を行っていることから，結果の比較を行うことができるよう，調査方法は 2003 年調査に合わせた。

具体的な手順としては，ウェブページの URL を対象として，設定したタイムアウト時間内にアクセスが可能かどうかを調査した。アクセスできた場合を，「生存」とみなした。アクセスできなかった場合を「消失」とし，消失していた場合には，その理由を記録した。タイムアウトした場合には「タイムアウト」とし，サーバそのものが見つからない場合には「ホストエラー」とした。サーバにアクセスできたが該当ページへのアクセスできなかった場合にはレスポンスコードに基づき、それぞれのエラーとして扱っている。調査期間は 2013 年 7 月 22 日から 7 月 28 日の 7 日間である。

2.3. 寿命調査

生存調査と並行して，同じ 1,000 万件のウェブページを対象として，Internet Archive の登録データの取得を行った。

ウェブページがいつ誕生しいつ消失していったかに関する情報を正確に取得することは困難である。しかしながら，ウェブアーカイブがページを取得したデータを用いることで，生存分析のために取得した日時を利用するよりも，より正確な日時に近いデータを取得できる可能性がある。そこで，本研究では，現在 3540 億ページを保存している Internet Archive からデータを取得した。

Internet Archive からのデータの取得には，Memento API[8]を用いた。Memento API は，様々なウェブアーカイブ上に蓄積されたデータを透過的に取得できるように設計された API である。

2013 年 7 月に行った生存調査のデータと Internet Archive から取得したデータとを組み合わせてウェブページの寿命の近似値を得た。ウェブページの誕生と消失については，以下の手順によって設定した。

ウェブページの誕生については，Timemap での「first」（最初にアーカイブによって取得された日時）を用いた。その際に Internet Archive から当該ウェブページのデータが取得できなかった場合には，寿命調査の対象から除外した。Internet Archive の取得開始が，NTCIR による取得日よりも遅い場合は，NTCIR のクロール期間の最終日である 2001 年 11 月 12 日 00 時をウェブページの誕生の近似値として用いた。

次にウェブページの消失は，生存調査の結果，消失していることが分かった場合は，Timemap の「last」（アーカイブによって直近で最後に取得された日時）を用いた。生存調査で，現在も生存していたページは寿命の集計から除外した。

上記の方法で判定したウェブページの誕生日時と消失日時の間の日数をウェブページの寿命とした。

3. 調査結果

3.1. 生存調査

2003 年調査と比較して，生存調査の結果を表 1 に示した。ウェブページの状態は次のように判別した。

- タイムアウト: 接続時間内に応答がなかった
- ホストエラー: サーバに接続できなかった場合
- サーバエラー: ステータス・コード 500～505
- 認証失敗: ステータス・コード 401～403，407
- ファイル移動: ステータス・コード 301，303，307
- ファイルなし: ステータス・コード 404
- 他のエラー: その他のエラーコード
- 生存: ステータス・コード 200

2003 年の列が前回の調査の結果で，2013 年の列が今回の調査の結果を示している。調査の結果，

2001 年に存在した 1000 万のウェブページのうち，2013 年 7 月現在でアクセスできたのは，779,176 件（7.79%）で，9 割以上のウェブページが収集から 12 年の間にアクセスできなくなっていた。2003 年調査の 5,336,099 件（53.36%）よりもアクセスできるページの比率が減少していた。

アクセスできなかった原因は，2003 年調査では，「ファイルなし」がアクセスできなかった理由の多くを占めていた。本調査では，「ホストエラー」が全体の 46.5%で最も多くなっていた。ホストエラーに含まれるのは，ウェブサーバがなくなっている場合，DNS によってドメイン名が解決できない場合などが含まれる。2003 年調査で最も多かった「ファイルなし」は 35.02%で，2 番目に多かったが，2003 年調査と占める比率は同程度であった。

表 1 ページ集合の生存状況

		2003年調査		2013年調査	
		n	%	n	%
生存		5,336,099	53.36%	779,176	7.79%
消失	タイムアウト	84,255	0.84%	349,896	3.50%
	ホストエラー	878,467	8.78%	4,647,225	46.47%
	サーバエラー	17,254	0.17%	4,942	0.05%
	認証失敗	74,172	0.74%	66,795	0.67%
	ファイル移動	386,620	3.87%	240,721	2.41%
	ファイルなし	3,219,881	32.20%	3,501,616	35.02%
	他のエラー	3,252	0.03%	409,629	4.10%
	合計	10,000,000	100%	10,000,000	100%

つぎに，主要な SLD（セカンドレベルドメイン）別での集計を表 2 に示した。なお TLD（トップレベルドメイン）は全て jp ドメインである。これらの中では，go が最も生存率が低く，4.5%である。go ドメインのページにアクセスできなかった理由で最も多かったのは，「ホストエラー」で 70.7%であった。

3.2. 寿命調査

Internet Archive に登録されたデータを用いて，寿命の分析を行った。寿命調査は，Internet Archive のみに対してアクセスを行うこととなるため，サイトへの負荷を考慮しながら調査を行った。約 400 万件のデータが取得できた時点で，分析を行った。

はじめに，Internet Archive への登録と現在の生存の関係について表 3 に示した。466 万件のうち，4,006,158 件（86.0%）が Internet Archive に登録されていた。Internet Archive に登録されたページのうち 89.2%が現在はアクセスできなくなっているページであった。一方，Internet Archive に登録されていなかったウェブページは約 65 万件（14.0%）であった。そのうち，95.5%がアクセスできなくなっているページであった。フィッシャーの正確確率検定の結果，Internet Archive への登録とウェブページの生存との間に有意な関連が見られた（$p < 0.01$）。

表 3 Internet Archive への登録と生存状況

IAへの	消失		生存		合計	
登録	n	%	n	%	n	%
あり	3,572,998	89.2%	433,160	10.8%	4,006,158	86.0%
なし	624,331	95.5%	29,511	4.5%	653,842	14.0%
					4,660,000	100%

つぎに，2001 年に収集され，現在は消失していたページ 3,572,998 件に対して，ウェブページの寿命の分析を行った（表 4）。ウェブページ全体の平均寿命は 1,108.2 日であった。標準偏差は，879.5 日で大きくばらつきがあった。2001 年に収集されたウェブサイト中で最も早くに Internet Archive に登録されていたのは，1996 年 4 月 4 日からであった。

表 4 対象ページ全体の寿命

平均	標準偏差	中央値
1108.2	879.5	948.9

表 5 に，寿命を 1 年毎にグループ化した集計を示した。1 年程度の寿命を持つページが 87 万件で最多であった。6 年以降は概ね年とともに減少傾向に

表 2 SLD ごとの生存状況

	生存		消失														
			タイムアウト		ホストエラー		サーバエラー		認証失敗		ファイル移動		ファイルなし		その他のエラー		
SLD	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	小計
co	346312	7.5%	116060	2.5%	1949518	42.0%	3089	0.1%	21271	0.5%	141002	3.0%	1845171	39.7%	223812	4.8%	4646235
ac	130344	8.8%	125295	8.5%	684984	46.4%	1336	0.1%	23588	1.6%	23345	1.6%	443689	30.1%	43392	2.9%	1475973
or	125490	10.3%	31283	2.6%	498731	41.1%	101	0.0%	5739	0.5%	22198	1.8%	491830	40.5%	39120	3.2%	1214492
ne	66203	6.6%	29977	3.0%	579100	58.1%	126	0.0%	8098	0.8%	17675	1.8%	233068	23.4%	61907	6.2%	996154
gr	32746	10.4%	5530	1.7%	153649	48.6%	37	0.0%	1625	0.5%	2609	0.8%	114667	36.3%	5262	1.7%	316125
go	14054	4.5%	8595	2.8%	219731	70.7%	2	0.0%	1201	0.4%	7807	2.5%	52616	16.9%	6784	2.2%	310790
ed	12620	8.5%	2059	1.4%	74461	50.4%	12	0.0%	390	0.3%	4770	3.2%	50279	34.0%	3206	2.2%	147797

あった。一方で17年間継続していたページも僅かではあるが存在した(111 件)。累積比率で見ると, 2 年から 3 年の間に, 存在しているウェブページは半減していた。2 年程度で半減期を迎えるという結果は, 先行研究での結果と大きく異なることはなかった。

表 5 寿命 1 年ごとの集計

年数	n	%	累積%
0	300792	8.4%	8.4%
1	850287	23.8%	32.2%
2	592358	16.6%	48.8%
3	622572	17.4%	66.2%
4	464425	13.0%	79.2%
5	210479	5.9%	85.1%
6	207094	5.8%	90.9%
7	127516	3.6%	94.5%
8	73474	2.1%	96.5%
9	41475	1.2%	97.7%
10	30853	0.9%	98.6%
>10	51,673	1.4%	100.0%
合計	3,572,998	100.0%	

SLD ごとの寿命の平均値と標準偏差, 件数を表 6 に示した。表から, ac が 1,445.5 日と最も長く, 4 年近い寿命があったことが分かる。一方で, gr や ed は, 比較的寿命が短く, 誕生から 3 年程度で消失していることが分かる。

表 6 SLD ごとの平均寿命

SLD	平均	標準偏差	n
ac	1445.5	1075.5	507,177
co	1408.9	1259.5	125,488
go	1329.4	1018.4	547,215
ne	1231.7	1017.0	124,806
or	1206.6	1017.0	1,921,908
gr	1113.8	978.4	384,173
ed	1057.0	875.5	56,586

次に, カプラン・マイヤー法を用いて生存関数を推定した。現在も生存していたページは打ち切りデータとした。分析には, R2.15.2 の survival パッケージを用いた。SLD ごとの中央値, 95%信頼上限値、95%信頼下限値は表 7 の通りである。最も寿命が長い or の場合, 1,263 から 1,269 日で半数が消失に至る。最も短い ne の場合, 895 から 903 日で半数が消失する。次に, ログ・ランク検定を用いて, SLD ごとの生存時間に有意差があるか否かを確認したところ, 有意水準 1%で有意差が検出された。

表 7 SLD ごとの寿命の要約

				信頼区間	
SLD	n	イベント数	中央値	下側95%	上側95%
or	507,177	436,095	1,266	1,263	1,269
ac	547,215	475,263	1,155	1,151	1,158
その他	338,805	311,674	1,088	1,085	1,092
gr	124,806	108,374	1,083	1,077	1,087
go	125,488	114,992	1,044	1,037	1,053
co	1,921,908	1,730,975	1,023	1,022	1,024
ed	56,586	49,555	928	917	937
ne	384,173	346,070	899	895	903

4. まとめ

本調査では, 収集から12年を経過したウェブページ群の生存分析と, Internet Archive を用いた寿命の調査を行った。その結果, 2001 年に収集されたウェブページの 9 割以上が消失していることがわかり, またウェブページの平均寿命は, 1,108.2 日であった。

引用文献

1) Kahle B. Preserving the Internet. Scientific American. 1997, vol. 276, no. 3, p82-83, http://web.archive.org/web/19970215093036/http://www.sciam.com/0397issue/0397kahle.html, (accessed: 2013-09-04)
2) Koehler W. Web page change and persistence: a four-year longitudinal study. JASIST. 2002, vol. 53, no. 2, p. 162-171.
3) McCluskey, M. Website content persistence and change: Longitudinal analysis of pro-white group identity. Journal of Information Science. 2013, vol. 39, no. 2 188-197.
4) 池内淳ら. "ウェブの動的変化に関する調査". 2003 年度三田図書館・情報学会研究大会発表論文集. 慶應義塾大学, 2003-11-8. 三田図書館・情報学会, 2003, p. 19-22
5) Ainsworth S G. How Much of the Web Is Archived?. JCDL '11. 2011, p 133-136, http://arxiv.org/abs/1212.6177(accessed: 2013-09-04)
6) 江口浩二ら. NTCIR-3 WEB : Web 検索のための評価ワークショップ. NII journal. 2003, no. 6, p. 31-56.
7) 内田斉. メディアとしての Web の成長を測る : サーチロボットを使った Web コンテンツ統計調査の試み. http://www.a-brain.com/result/report/015/
8) Memento: Adding Time to the Web. http://mementoweb.org/

1 Kahle B. Preserving the Internet. Scientific American. 1997, vol. 276, no. 3, p82-83, http://web.archive.org/web/19970215093036/http://www.sciam.com/0397issue/0397kahle.html, (accessed: 2013-09-04)
2 Koehler W. Web page change and persistence: a four-year longitudinal study. JASIST. 2002, vol. 53, no. 2, p. 162-171.
3 McCluskey, M. Website content persistence and change: Longitudinal analysis of pro-white group identity. Journal of Information Science. 2013, vol. 39, no. 2 188-197.
4池内淳ら. "ウェブの動的変化に関する調査". 2003年度三田図書館・情報学会研究大会発表論文集. 慶應義塾大学, 2003-11-8. 三田図書館・情報学会, 2003, p. 19-22
5 Ainsworth S G. How Much of the Web Is Archived?. JCDL '11. 2011, p 133-136, http://arxiv.org/abs/1212.6177(accessed: 2013-09-04)
6江口 浩二ら. NTCIR-3 WEB : Web 検索のための評価ワークショップ. NII journal. 2003, no. 6, p. 31-56.
7内田斉. メディアとしての Web の成長を測る : サーチロボットを使った Web コンテンツ統計調査の試み. http://www.a-brain.com/result/report/015/
8Memento: Adding Time to the Web. http://mementoweb.org/