印刷メディアの電子化を計る指標

田口忠祐*, 三根慎二*, 長岡智子*, 石田栄美, 倉田敬子***, 上田修一*****
(*慶應義塾大学大学院, **国立情報学研究所, ***慶應義塾大学文学部)

1. はじめに

近年，図書や雑誌や新聞など，印刷メディアとして刊行されている多くのメディアを，World Wide Web（以下：WWW）や CD-ROM などを通じて利用することができる。では実際に印刷メディアはどの程度電子化されているのであろうか。そこで，印刷メディアの電子化を計ることが可能な指標を考案することが必要であると考えた。本報告では，予備的な調査から，印刷メディアの電子化を計る指標を考案，作成し，その指標に基づいた調査を行うことによって，印刷メディアがどれくらいの割合で電子化しているのかという割合を導き出すことによって，指標の妥当性を示すことにある。

2. 閾値方式による電子化指標

印刷メディアがどの程度の割合で電子化されているのかを指標化するにあたり，白書を対象とした予備調査を行った。まず，白書の中から電子化されているものを抽出し，全文，画像の再現性，文の書式，レイアウト等の観点について，実際の印刷媒体との比較を試みた。その結果,「印刷メディアの代替物となりうること」を大前提として条件を設定し，その条件（閾値）を満たしているものを電子化されているとみなすことにした。

実際に，どのような条件を設定するかについては，次の項目でそれぞれのメディアごとに説明する。単純な閾値方式では, 電子化率を以下のような式によって算出することができる。

$$\text{閾値方式による電子化率}(\%) = \frac{\text{条件を満たしてたものの総数}}{\text{印刷版の総数}} \times 100 \quad (1)$$

3. 電子化の現状

本報告において，電子化率を調査する印刷メディアは，白書，専門雑誌，新聞，小説の 4 種類である。この中から，それぞれ標本を抽出し，電子化の状況を調査した。以下では，4 種類の印脱メディアについて，個別に調査方法およびデータを示すことにする。なお，ここで用いた手法は，全て単純な閾値方式を採用している。

3.1. 白書

白書の場合，設定した条件は，最新版が提供されていることと，全文と図表が提供されていることとした。特に，白書に用いられている図や表が重要と判断した。そのため，図表がなく，全文のみが電子化されているものは，ここで用いている閾値方式による電子化率では計測されない。

調査対象には,「かんぽう」のサイト内の「最新/白書の一覧」[1] (調査日時：2002 年 9 月 11 日)中の「日本の主な白書」に掲載されている白書の全 41 点を対象に調査を行った。

3.2. 専門雑誌

雑誌の場合，次の号が刊行される前に，電子版で提供されているのであれば，印刷版の代替物になり得ると考えた。そこで，ここでは条件として，印刷版で提供されている最新号が

電子版でも提供されていることと，全論文が提供されていることとした。

調査対象は，国会図書館の Web サイトの雑誌記事索引採録誌一覧[2] に掲載されている雑誌, 国内刊行和文誌11,743誌と国内刊行欧文誌135誌を合わせた計11,878誌を母集団とし，等間隔抽出法により 100 タイトルの文献を抽出した。これを WWW で探索し，電子化の状況を調査した。次に, JCR Science Edition(2001)と JCR Social Science Edition(2001)から，国名を「Japan」で検索した。その結果，154 誌が検索され，電子版で全文が提供されているものの最新号を調べることによって，電子化の状況を調査した。

3.3. 新聞

新聞の場合，メディアの特性から速報性が重要であると考えた。そのため，印刷版の代替になりうるかどうかを考えた場合，その日の朝刊から夕刊までの間に電子的な形で提供されている必要があると判断した。そのため，条件には，調査時点で記事が電子的な形で提供されていることと，記事の全文が提供および再現されていることとした。なお，ここでは，新聞の見出しの再現性については考慮しなかった。

調査対象としては, 朝日新聞と日本経済新聞の 2002 年 8 月 21 日の朝刊を対象とした。調査方法としては，朝日新聞の場合，WWW 版の asahi.com(http://www.asahi.com)との照合を行った。日本経済新聞の場合も，同様に WWW 版の NIKKEI NET(http://nikkei.co.jp)との照合を行い，それぞれの新聞についての電子化の状況を調査した。

3.4. 小説

小説の場合，条件として，全文が提供されていることとした。調査対象は,『ちくま日本文学全集』1-60[3] に収録されている作品を，NDL Web-OPAC(http://webopac2.ndl.go.jp/)で検索し，見つからなかったものは, 紀伊国屋の Web(http://bookweb.kinokuniya.co.jp/)で検索した。得られた 699 作品から等間隔抽出法で 100 作品を抽出し，青空文庫[4]，電子文庫パブリ[5]，電子書店パピレス[6] それぞれのデータベースで検索し，無いものについては，Google で著者名および作品名で検索した。さらに小説の場合は，CD-ROM でのみ提供されているものも存在すると判断したため，『世界 CD-ROM 総覧 2001』[7] から，同じ標本を対象として、CD-ROM でのみ電子化されているものの有無を調査した。それぞれメディアについての調査結果を表 1 で示す。

表 1 閾値方式による電子化率

メディアの種類	電子化率	注
白書	65.2%	
専門雑誌	3.0%	国会
	34.4%	JCR
新聞 (朝刊のみ)	31.5%	朝日新聞
	10.5%	日本経済新聞
小説	39.0%	WWW +CD-ROM

4. ポイント方式による電子化指標

単純な閾値方式による電子化率の算出方法には問題点がある。それは，多くの場合，全文および最新のものが提供されているという条件を設定しているからである。この場合，たとえ過去のものが電子化されていても，電子化されているとはみなされず，計測されないことになる。そこで，閾値方式による電子化率を補完するためにポイント方式による電子化率を考案することが必要であると考えた。

そこで，電子化率を計る指標として，ポイント方式を採用した。単純な閾値方式でなく，ポイント方式を採用することにより，電子化されているそれぞれのメディアごとの比較が可能になると考えたからである。先ほどの調査からの問題点を考慮した上で，白書，専門雑誌，新聞，小説それぞれのメディア全てに適用可能な共通の尺度を考案した（表 2 を参照）。

表 2 各項目に割り当てられたポイントの値

			メディアの種類			
			白書	専門雑誌	新聞	小説
			ポイント	ポイント	ポイント	ポイント
物理レベル	媒体	WWW	1.0	1.0	1.0	1.0
		CD-ROM	0.3	0.3	0.3	0.6
		オンライン DB			0.2	
	ファイル形式	HTML	1.0	0.8	0.8	1.0
		PDF	0.8	1.0	0.6	0.8
		Text	0.3	0.2	0.3	0.9
		その他		0.2-0.7	0.1	0.5-0.8
構造レベル	内容の再現性	全文+図表	1.0	1.0	0.9	
		見出し+全文+図表 or 写真			1.0	
		全文のみ（図表なし）	0.5	0.5	0.8	1.0
		テキストの一部	0.3		0.3	
		テキストの一部+図表 or 写真			0.6	
		テキストの一部(改変)+図表 or 写真			0.5	
		目次のみ	0.1	0.1		
		目次+抄録		0.2		

それぞれの項目におけるポイントの値は最小を 0.0，最大を 1.0 とした(なお，各項目中のポイントを累積するという方式はとらないため，各項目に付与される最大ポイントは 1.0 を超えない)。このポイント方式による電子化率には，総ポイント数と平均ポイント数の 2 つから算出することとした。それぞれの算出方法は以下の通りである。ここでの総ポイント数とは，電子化されている総量を表している。一方の平均ポイント数は，同じメディア内での比較ができるように正規化したものである。

総ポイント数 = 資料1点1点が獲得したポイントの合計　(2)

$$\text{平均ポイント数} = \frac{\text{個々のメディアごとの 総ポイント数の合計}}{\text{印刷版の総数}} \quad (3)$$

5． ポイント方式による電子化率の例

ここでは，実際にポイント方式を用いた計測を，例示的に行う。全てのメディアに対して共通なものとして，以下のような手順を踏む。始めに印刷版で刊行されている数の計測を行

う。次に，電子化されているもののみを，ポイント方式の各項目を参照しながらポイントを付与していく。それぞれポイントが付与されたものを対象に，ポイント方式の式(2)および(3)を適用することで，電子化率を算出する。

以下では，白書，その中でも『外交青書』を例に取って計測方法を例示的に示す。まず，『外交青書』の刊行状況を調査する。その結果，1957 年の創刊から 2002 年の最新号まで，合計 45 点刊行されているという結果を得た。次に，WWW や外務省の HP(http://www.mofa.go.jp/mofaj/)を調査した結果，1996 年から 2002 年までの合計 7 点が電子化されていた。電子化されているものから，最初に，ポイント方式の「物理レベル」から「媒体」の調査を行った(表 2 を参照)。『外交青書』の場合，すべて WWW で提供されていたため，1.0 ポイント(以下：p.)を付与した。同様に「ファイル形式」についても調査した。外交青書の場合，全て HTML 形式で提供されていたため，全てに 1.0 p.を付与した。もしここで，仮に HTML 形式と PDF 形式双方で提供されていたとする。白書の場合では，HTML に 1.0 p.，PDF に 0.8 p.が付与されていることから，HTML の 1.0 p.と PDF の 0.8 p.を合計した 1.8 p.ということはせずに，一番高い値である HTML の 1.0 p.が付与される。最後に，「構造レベル」から「内容の再現性」を調査する。2001 年版から 1999 年版までと 1996 年版には全文が提供されているが，図表は提供されていないということがわかった。そのため，これらには，全文のみ（図表なし）の 0.5 p.が付与される。それ以外のものは，全文と図表が提供されていることから、1.0 p.を付与した。その後，式(1)および(2)を適用して，ポイント方式による電子化率を算出する。さらに，専門雑誌，新聞，小説に対しても同様の方法で調査を行った（それぞれのメディアについての例示的な調査結果は，表 3 で示す）。

表 3 ポイント方式による電子化率

メディア	対象	総ポイント数	平均ポイント数
白書	外交青書	19.0 point	0.42 point
	経済財政白書	105.0 point	1.91 point
雑誌	Chemistry Letters	153.4 point	0.42 point
新聞	朝日新聞	755580 point	0.79 point
	日本経済新聞	1264677 point	0.86 point
小説	抽出した 100 件	114.5 point	1.15 point

6． 結論

我々の考案した指標を利用することによって，おおよその電子化率を計測することができると言える。ただし，ポイント方式による電子化率の場合は，メディア自体の特性などの点から，メディア間での比較は難しいが，メディア内での比較を行うことは可能である。今後の課題としては，メディア間での比較が可能な指標を考案する必要があると言え，指標を改良していくことで,様々なメディアのより詳細な電子化率を計測することができると言える。

参照文献

(1)最新/白書の一覧. http://kanpo.net/hakusho00.htm. (2)国立国会図書館-National Diet Library: 電子図書館 - 雑誌記事索引採録誌一覧. http://www.ndl.go.jp/library/magazine_index.html. (3)ちくま日本文学全集. 筑摩書房. (1-60). (4)青空文庫. http://www.aozora.gr.jp. (5)電子文庫パブリ. http://www.paburi.com. (6)電子書店パピレス. http://www.pappy.co.jp. (7)共同計画株式会社出版事業部 Data Net 編集部. 世界 CD-ROM 総覧. 共同計画出版事業部, 2001. vol. 13.