電子化テキストの国際的共有

電子化テキストと書物との接点

——テキスト・データとページ構成——

家　辺　勝　文

キーワード：テキスト・データ、書物の形態、データの可搬性、マーク付け、WYSIWYG

要　旨

書物の形をとる原文献からテキスト・データの構築を行う時、文字列としてのテキストの電子コード化だけでは文書の論理構造や空間的な表現形態を取り残すことになる。文字コードのみならず文書単位での可搬性を確保する文書形式の規格化により、共有可能なデータとして書き言葉の表現形式をより豊かに捕捉したものが構築可能になる。反面、データ化にあたっての選択肢は増え、研究目的に応じたデータの構築・再構築が一層必要となる。

1.　問題の所在

1.1　テキストと書物の形態

テキスト・データベースは多くの場合、書物から出発しており、電子出版のプロセスにおいても、従来の書物の形態が様々な側面から出力形式を規定している。書物と全く独立したテキスト・データの処理を構想することは困難であろう。そのことは、従来の書物という形態が書き言葉としてのテキストの成り立ちに密接に関与しているからであり、テキスト理解に際しても有用な手がかりを提供しているからではないだろうか。

1.2　書物上の表現形態の電子データ化

書物上の表現形態をページ上の空間配置を含めて電子データ化する手段として構想または実用化されているものには、次のようなものがある。

（1）非分析的な画像データとしての取り込み

原文献資料の画像データベース構築の方式。

（2）分析的に付加情報を付け加える方式

1）マーク付け方式（本体となるテキストの中に一定の仕方で制御記号を書き込む）

a）ページの組み方のマーク付け

電算写植や TeX などの方式。テキストなどをページ上にどのように配置するかを細かく指定する。

b) テキスト内容の階層的構造のマーク付け

文書記述言語 SGML (Standard Generalized Markup Language) (ISO 8879 [1986], JIS X 4151 [1992]) の基礎となっている手法。テキストの階層的論理構造に着目して、その構成の区分化を記述する制御記号をつける。各部分の具体的配置方法は出力手続きの問題であるとして、文書構造を定義する記述自体には含めない。少なくともそのようなマーク付けを基本に考えてある。

2) WYSIWYG (What You See Is What You Get) 型の文書編集方式

テキストなどのページ上の配置を画面で見ながら組み立て、編集用ソフトの仕様によって記録する、パソコンを利用したデスクトップ・パブリッシング (DTP) で発達した方式。

実際には、画像データの検索に便利なように付加情報を付け加えたり、マーク付け方式の中でもページ上の配置とテキスト構造の記述を組み合わせる方式が開発されている。ここで注意したいのは、画像データの取り込み以外の手法は、主として新規のテキストを本ないし文書の形に印刷するという作業の流れに即して生まれてきたという点である。そのため、それらの手法を既存文献の電子データ化に利用することもできるが、道具としては不十分な点がある。

本稿ではデータの共有ないし共同利用という観点から、文書の形式を記述するための要件に立ち戻り、特にテキストとページ構成上の空間配置の関係について、その基本的な問題点を整理し、検討する。

2. 話し言葉・書き言葉・電子化テキスト

2.1 話し言葉を主とし書き言葉を従とするテキスト観

テキスト・データというとき、まずテキストを文字列としてのみ把握し、これを電子コード化することが考えられてきた背景には、単に技術的制約のみならず、一定の言語観・テキスト観が前提になっていると考えられる。

それは、1) 言語の表現形式は音声言語が基本であり、文字による (理想的には表音的アルファベットによって表記される) 書き言葉は音声言語の表記形式に過ぎず、従属的である、2) 音声言語の形式の基本的性格として線状性ないし逐次性がある、3) 書き言葉の実現形態の中でも線状のテキストのみが意味解釈にとって本質的である、という見方に要約されよう。

これは、19 世紀にヨーロッパで音韻変化の研究に重点を置く言語学が発達して顕著に主張されるようになった言語観であり、F. de Saussure の『一般言語学講義』に至ってもその端的な表明が見られる。[注1] テキストの電子データ化という局面において、書き言葉を出発点としながら、テキスト・ラインのみの抽出が特に疑問視されないとすれば、この言語観・テキスト観が暗黙の前提とされているからではないだろうか。

しかし、書き言葉は話し言葉の表現形式をすべて忠実に表記しうるものではなく、反面、話し言葉では表現できない要素を書き言葉の表現形態のうちに含むことが可能である。実

際上、上記の言語観・テキスト観は書き言葉の諸相とその媒体である本などの空間的・立体的形式の軽視につながる。

2.2 書誌学、書物の歴史の観点からの書物の形態と社会的伝搬様式の重視

書物の形態とテキストの読解の関係についての関心が、すでに書誌学、社会史といった分野で、近年高まりを見せている。特にコンピュータで読める大量のテキスト・データの蓄積という点で先行している欧米の学者の間で、ディスプレイ上の文字列主体のテキストに対比して伝統的な書物の形態への関心の高まりが見られることは興味深い。[注2]

書物の形態を重視する観点に立てば、書かれたものとしてのテキスト解釈において、その媒体（書物など）のあり方と媒体上でのテキストの物質的、空間的実現形態は無視できないものであり、このことは書物の編集過程でのページ構成上の様々な配慮の重要性の認識とも軌を一にする。

コンピュータ上では必然的に新しいテキストの形態が生まれるが、書物上のテキストの研究にコンピュータを利用するという観点からは、書物の形態をもデータとして取り込むことが課題となる。

2.3 何のために何を電子データ化するのか

書物の形態の特徴を最大限電子データ化しようとすれば、コンピュータ上のデータとしては効率が悪くなるかもしれない。しかし、狭いテキスト観に立ち、文字列のデータ化に終始する場合には、もとの文献から多くの情報を取り残すことになるだろう。

そこで問われるのは、どういう目的のために、もとの文献をどのような方式で選択的にデータ化するのかという、作業内容についての明確な認識であろう。何のために何をコンピュータであつかえる形にするのかという基本的な問題がそこで問われてくる。

3. テキストとページ構成

3.1 テキスト構成要素および空間配置上の諸要素の分類

次に、書物上の、テキスト自体を構成する要素と、空間配置上の諸要素について、いく

表 1 テキスト構成要素と空間配置上の諸要素

	特定の言語の構造への依存性が高い	依存性が低い
テキスト	文字種，言語単位， 表記法（わかち書き，文の終止など）， 文法情報 〈規則性が高い〉	文章の構成またはテキスト構造（要素の属性〔見出し，本文，注記など〕，階層的論理構造）など 〈伝統的規範〉〈自由な表現〉
空間配置	横書き，縦書き 進行方向（左→右，右→左） 〈規則性が高い〉	改行，改ページ，横組，縦組， 行間の使い方，段組，ブロック化， その他のページ・レイアウト 〈組版の原則〉〈視覚的要求〉 〈伝統的規範〉〈自由な表現〉

つかの観点から分類し、その電子データ化にともなう問題点を検討する。

まず、テキスト構成要素ならびに空間配置上の諸要素について、大まかに、特定の言語の構造への依存性の高いものと低いものとにわけ、それぞれについて規則性の高さ、様々な制約の有無と自由な表現の可能性とについて検討してみる(表1を参照)。電子データ化に際しては、規則性の高いものについては、コード化ないし処理システム一般の問題として取り扱うことができるだろうし、自由な表現の可能なものについては、それぞれの表現の分析に立脚した個別的な記述が必要となるだろう。また組版上の慣習に基づくと見なされる形式などについては必要に応じてその取り扱い方を判断することになる。

文章の構成またはテキストの論理構造といったものはまさに個別的な記述の対象となる。他の言語に翻訳してもほぼ同等に維持されうるという限りにおいて、特定の言語への依存性は低いものと言えるが、一つの言語文化の中で文章の書き方の規範といったものが存在する場合、文字や表記の規則性とは性格が異なるものの、その規範に則って書かれているかどうかは、文章の構成を分析する上で無視することはできない。

空間配置の方では、左から右へといった一行の中の進行方向は、特定の言語への依存性の高い要素と言える。ただし、日本語のように横書きも縦書きも許される場合には、どちらかを選択することで一つのスタイルを表現することもできる。その他、改行、改ページ、行間の使い方、テキストのブロック化などの空間配置上の手法は、特定の言語への依存性は低いものの、組版上の技術的制約とか、組版方法の伝統的規範、さらには人間の視覚に対する配慮などの様々な条件に依存しつつ組み立てられている。しかし、実際には、なおかなり多くの選択肢の中から配置方法を選べるものであり、その意味で自由な表現と解釈の余地のあるものである。

3.2 結果としての書物上の「空間配置」と表現手段としての「レイアウト」

空間配置の中でも、組版処理の結果、偶然的な実現結果としてあらわれてくるものと、テキストの構造に即した区切りや配置などの意図的な表現手段として組み立てられるレイアウトとは、一応概念的に区別して考えることができる。

偶然的な実現結果といえども、ページ、行などによるテキストの位置情報は、テキストの読解の便宜上一定の役割を果たしており、テキスト・データベースを構築する場合でも重要な付加情報となっている。

レイアウトの効果的用法として、振り仮名などの行間の使い方や、テキストブロックどうしの組み合わせ方次第で、テキストの一直線的な流れを変更して、複線的な流れを視覚的に表現したり、テキストの内容だけからは必ずしも自明でないテキストの中の離れた部分どうしの関係を表現したりする場合もある。こうした表現形態に忠実なデータを構築するためには、技術的原則やテキスト構造に還元できないレイアウトについて個別的に記述する必要がある。

意図的なレイアウトと言うとき、誰の意図なのかということも問題になる。多くの場合、編集者や出版する側の意図が反映されるだろうが、そもそもテキストの中味と無関係ではありえず、また著者がレイアウトの細部に至るまで細かく注文をつけることもしばしばあ

る。どこまでが表現手段として意図されたレイアウトと解釈されるかという問題と合わせて、その判断はテキスト成立に関する研究成果に依存し、電子データ化に際してもその判断が反映される。

3.3 テキストの構成や空間配置に関するデータの性格

テキストの文字列については、一字の欠如、誤りがデータ上の重大なエラーとなるといった性格を持つ。一方、テキストの構成や空間配置の分析については、既存文献の場合、研究者によって解釈がわかれる可能性もある。また、分析の精度の設定やデータ化の内容には選択の余地があり、研究目的と研究者のテキスト解釈に応じたデータ構築が要求される。

新規に文書を作成する場合に準拠すべきテキスト作成のためのルールと既存文献の電子化にあたっての分析・記述の手法とでは自ずからその性格が異なる。前者にあっては、ルールに従ってテキストの書き方を直していくが、後者にあっては、研究目的に応じて、また様々な表現手段を駆使した豊かな表現内容の読み取りに対応できるよう、電子データ化の手法の方が改良の対象となる。

4. 文書処理規格の標準化とデータの多目的利用ないしは可搬性（portability）

4.1 データの可搬性

文書処理システムの標準規格を作れば、機種・ソフト間の互換性が確保され、データの再利用に道が開けるだろう。データの可搬性（異なる機種・環境上での互換性）という観点から見て、文字列の電子化だけなら、使われる文字コード・文字種の共通化・規格化が主たる関心事になるが、テキストを含む拡張された文書形態の可搬性という観点からは、文書の論理的・形式的構成の電子データ化も必要となり、それに応じた方式の共通化・規格化が問題となってくる。

4.2 マーク付け（markup）方式 ― SGML

国際的規格となった文書記述言語 SGML は、テキスト構造の分析・記述方式として一般性があるため、これを使って各種の文献の電子データ化を推進しようとする動きがある。しかし、新規文書の作成にとどまらず、既存文献の電子データ化のためには、そのままではうまく使えない点がある。

SGML では、階層的論理構造をもったテキスト主体の文書（図や表を含んでもよい）から出発して、結果として簡便にページ構成を合成する出力工程として空間配置が考えられており、テキスト構造が最重要の情報で、空間配置は出力装置に依存した可変の実現手続き程度にしか位置付けられていない。こうしたマーク付けの基本理念の特性から、例えばテキスト・データへの便利な付加情報となる原文献のページ・行番号などの位置情報を論理構造と同一の階層構造の中には書き込むことができない。ただし、SGML では記述上複数の文書型の併存の余地が残されており、論理的なマーク付けに加えて出力文書の配置情報をも書き込み、複数の視点をもたせることも不可能ではないが、そもそも階層的構造のパターンにあてはまらないような複線的なテキスト構造などは記述が困難である。[注3]

4.3 WYSIWYG 型の文書処理

WYSIWYG 型の文書処理システムによる作業では、出力結果をディスプレイ上に表示しながら入力、編集作業を行うのが普通であり、作成しようとする文書は、常に空間上の配置を伴って組み立てられていく。このタイプのシステムは多くの場合、1)テキストの論理構造を分析・記述しているわけではなく、2) 空間配置を表示できるといっても、これをテキストの内容との関係の上から分析・記述しているわけではない、3) 何よりもシステムとして特殊化し、特定の機種・環境に依存している。確かに、目的とする文書形態に応じて柔軟な処理が可能であり、便利ではあっても、一度そのシステムで作られた文書は、多く場合、そのシステムによる印刷・改訂版の作成以外に使い道がない。

WYSIWYG の前提となる GUI (Graphical User Interface) の流れの中で、最近、「文書指向の」 (document-oriented) ないしは「文書中心の」と称される文書単位での記録・表示の互換フォーマットがいくつか開発されつつある。一例は、Adobe Systems 社の Acrobat と呼ばれる一群のソフトウェアであり、これは Macintosh、Windows、UNIX のいずれの GUI 環境上でも、文字の書体、テキストや図・表の配置、データ間のリンクまで含め、ほぼ同一の文書形態を維持したままデータ交換を可能にするというもので、同社のページ記述言語 PostScript を利用する。[注4] さらには文書をユーザ・インタフェースの中心に置くようなシステムソフトウェアレベルでの文書指向インタフェース環境が開発されつつある。[注5] こうした可搬文書フォーマットの出現の背景には、文字列としてのテキストより拡張された内容をもつ文書形態でのやりとりの必要性についての認識があり、将来、こうしたもののいずれかが事実上の標準規格となっていく可能性も考えられる。

5. 結論

電子化データについては、一度出来上がってしまえば不変という捉え方ではなく、利用目的に応じて、原文献に立ち戻って柔軟に再構築される道具という捉え方のほうが、よりその性質にふさわしい見方なのではないだろうか。文書形式の可搬性が確保されたとしても、文書処理規格の将来における改訂の可能性も含めて、電子化データは研究支援ツールとして必要に応じて増補・改訂を受けうるものである。その意味で、電子化データは原資料にとってかわる性質のものではなく、その利用形態においても、原資料たる書物と密接不可分である。

データの構築・再構築に際しては、テキスト構造ならびに原文献上の空間配置などを書き言葉における表現手段という観点から整理して、必要に応じてデータとして付け加えることも課題となる。文書形態のデータの可搬性を考えるなら、従来の狭い意味でのテキスト主体のデータ形式にとどまらず、デスクトップパブリッシングで馴染みの深い手法にも、文書規格上ないしはインタフェース上の問題として活用の余地があるのではないかと思われる。

注1 音声言語を主とする言語観については Paul (1880 [1975] : pp. 373-389)、Saussure (1916

[1972]: p. 45, 103, 165)を参照。『一般言語学講義』における書き言葉の理論的位置付けについては Harris (1987: pp. 14-32, 69-78)に論考がある。

注2　フランスの社会史の分野では Chartier (1991; 1992)などを参照。イギリスにおける書誌学の分野での問題提起としては McKenzie (1986)を参照。いずれの場合も、コンピュータで読めるテキストは書物の物質的制約から解放され、新たなメディアの特性に従うため、新しい形態のテキストが生まれるという見方で、むしろコンピュータ上では書物からの離脱を基本的与件とみなしている。

注3　国際規格の「開放型文書体系 (Open Document Architecture) 及び交換様式」(ISO 8613-1, 2, 4-8 [1989]; JIS X 4101, 4102, 4104-4108 [1993])では、文書概念の基礎として論理的な見方（章、節、段落への分割）と割付け的な見方（ページ、区画への分割）の二つの補完的観点があることを明示している。そこでは「編集、書式付け及び表示に関し、文書発信者の意図が最も効率的に伝えられるように」(X 4101, 5.1) 文書の交換を促進することが目的の一つとして掲げられており、データ構造の表現では SGML の応用、論理構造と割付け構造の連結などが規定されている。

注4　類似のソフトウェアとして、Macintosh 上で実用化されている、Common Ground (No Hands Software 社)、Replica (Farallon 社)などがある。

注5　Macintosh 上では OpenDoc (Apple Computer 社)、Windows 上では OLE (Object Linking and Embedding) (Microsoft 社)など。

参考文献

伊藤雅光 (1992)「電子化テキストと画像データ」,『国語学』, 170 集.

佐竹秀雄 (1993)「「書き言葉」の空間学」, 月刊『言語』, Vol. 20, No. 8.

芝野耕司 (1993)「SGML の現在」,『情報の科学と技術』, Vol. 43, No. 12.

進藤咲子(1982)「ふりがなの機能と変遷」,『講座日本語学 6　現代表記との史的対照』, 明治書院.

土屋俊 (1991)「テキストデータベースの現状と展望――これから始めようとする人のために――」,『日本語学』, Vol. 10, 1991 年 8 月号.

當山日出夫(1991)「読書とコンピュータ――テキストデータで本が読めるのか――」,『情報処理学会研究報告』, Vol. 91, No. 89, 91-CH-11.

當山日出夫 (1992)「電子化テキスト総論――国語学の立場から――」,『国語学』, 170 集.

水谷静夫 (1991)「言語研究におけるデータの意義」,『日本語学』, vol. 10, 1991 年 8 月号.

安永尚志(1989)「国文学原文献資料データベースシステムの開発」,『国文学研究資料館紀要』, 15.

安永尚志 (1991)「日本古典文学本文データベース」,『日本語学』, vol. 10, 1991 年 8 月号.

Chartier, Roger (1991) *Ecriture, livre, lecture : trois histoires et un futur.* (1991 年 3 月 28 日、日仏会館での講演。テキストは邦訳の形で『読書の文化史―テクスト・書物・読解』[福井憲彦訳, 新曜社, 1992] の中に「書くこと・書物・読むこと――三つの歴史と一つの未来――」と題して収録。)

Chartier, Roger (1992) *L'Ordre des livres : Lecteurs, auteurs, bibliothèques en Europe entre XIVe et XVIIIe siècle*, Aix-en-Provence : Alinea. (邦訳:『書物の秩序』長谷川輝夫訳, 文化科学高等研究院出版局, 1993)

Harris, Roy (1987) *Reading Saussure : A critical commentary on the Cours de linguistique générale*, London : Duckworth.

Knuth, D. E. (1990) *The TeXbook*, 19 th printing rev., Reading, MA: Addison-Wesley. (邦訳:『改訂新版 TeX ブック―コンピュータによる組版システム』斎藤信男監修・鷺谷好輝訳, アス

キー出版局，第2版，1992）

McKenzie, D. F. (1986) *Bibliography and the Sociology of Texts : The Panizzi Lectures 1985*, London : The British Library.

Paul, Hermann (1880) *Prinzipien der Sprachgeschichte*, 5. Auflage (1920); 9., unveränderte Auflage (1975), Tübingen : Max Niemeyer.（引用ページは1975年版による。）（邦訳：『言語史原理』福本喜之助訳，講談社学術文庫版，1976）

Saussure, Ferdinand de (1916) *Cours de linguistique générale*, Paris : Payot.（引用ページはTullio de Mauroの校訂本［1972］による。）（邦訳：『一般言語学講義』小林英夫訳，岩波書店，1972［改版］）

＜付記＞

本稿は、1993年10月30日国語学会秋季大会において口頭で発表した内容をもとに加筆したものである。発表要旨集の原稿と比べて、第2節はその後の知見を含めて大幅に書き直し、第4節は見出しを変え内容を追加した。第5・6節は最後の結論にまとめた。注記は、新たに追加した内容の補足にとどめ、他は整理・割愛した。その他、若干の表現を改め、参考文献にも論旨に沿うように若干の入れ替えがある。（1994年3月28日）

本稿締切後に、SGMLを応用して各種文献の可搬電子データ化を目指すTEI（Text Encoding Initiative）の指針の第3版である *Guidelines for Electronic Text Encoding and Interchange* (TEI P3)（ACH, ACL, ALLC, 1994）が公表されたが、本稿ではその内容についての検討は追加せず、締切時点の構成のままとした。なお、4.2でふれた非階層的構造の取り扱いの困難さについては、TEI P3の第31章（pp. 751-760）でも言及されており、同時にいくつかの事例に即した手法が列挙されている。（1994年8月21日追記）

——日仏会館フランス事務所——