

技術協力調整班
平成11年3月15日

電子図書館推進委員会
委員長 原 田 公 子 殿

技術協力調整班
班 長 田 屋 裕 之

「出版データの活用に関する調査報告書」の提出について

標記について、技術協力調整班は、平成10年6月24日（水）以来、12回の会合をもち、電子図書館推進に伴う技術的課題について鋭意検討を重ねてきましたが、今般、別紙のとおり検討結果を取りまとめたので報告します。

検討経緯

○平成10年		第7回	10月29日（木）
第1回	6月24日（水）	第8回	11月25日（水）
第2回	7月9日（水）	第9回	12月25日（金）
第3回	7月22日（水）	○平成11年	
第4回	8月5日（水）	第10回	1月27日（水）
第5回	9月2日（水）	第11回	2月22日（月）
第6回	9月30日（水）	第12回	3月1日（月）

技術協力調整班

（管理職以外の職員名を含む一覧であるため、被覆）

出版データの活用に関する調査報告書

1. 「印刷組版情報」と電子図書館の『蔵書』

国立国会図書館は、平成 10 年 5 月に「国立国会図書館電子図書館構想」（以下「電子図書館構想」）を発表し、それ以降の当館の電子図書館の具体的な計画策定および事業実施の基盤となっている。この「電子図書館構想」では、第 2 章「電子図書館への『蔵書』の構築－文化の保存と提供－」において電子図書館の蔵書の構築にあたり、「可能な限りの一次情報の提供が望ましいが、著作権等の制度的課題、財政的課題、外部機関との連携・協力等を考慮して実施する」と述べている。

実際に提供する資料として、①国立国会図書館が収集する電子出版物、②印刷物を電子化して蓄積する電子アーカイブ、③インターネット上で提供される情報資源等を主なものと想定している。（この他に当館は所蔵せず、アクセスを保障する情報も考慮に入れている。）また、「上記のほか、学協会を含む出版社との協力事業として、出版のプロセスで作成される電子的な印刷組版情報の収集・蓄積及び独自にサーバを用意しない出版社の電子出版物の当館のサーバへの蓄積とその提供等を、当館所蔵の電子情報として想定する」と述べ、出版社の持つ印刷組版情報の電子図書館への活用の可能性を示している。

一方、納本制度調査会は電子的な媒体の納本について、「答申 21 世紀を展望した我が国の納本制度の在り方－電子出版物を中心に－」（平成 11 年 2 月 22 日）で、パッケージ系の電子出版物は納本の対象とし、ネットワーク系の出版物については積極的な収集を図るものの当分の間は納本の対象外とするという方向を示している。この「電子的な印刷組版情報」については、出版物ではなく、また特定の媒体やファイル形式に「固定する前のデータ」であって、納本の検討対象となる「出版物として固定されたデータ」とは異なるため、同調査会での検討の対象とはなっていない。

これまで当館の電子図書館の各種事業や実験では主として既存の紙の出版物を電子化（主に画像情報）する方法を用いてきた。これは既に紙で出版された情報を遡及して電子化するには有効な方法である。しかし、当館では既に国会会議録検索システムや電子化英文政府刊行物を提供しており、これらテキスト情報は検索やデータ伝達の面で、画像情報よりはるかに優れた特徴をもっている。また、現在多くの出版物は、印刷までの工程が電子的に行われ、印刷組版情報もほとんどが電子化されている。このような電子化されたデータを活用する方法も電子図書館の『蔵書』の構築には有効な手段であると考えられる。

つまり、

印刷組版情報→紙の出版物→図書館での電子化

というプロセスだけでなく、

印刷組版情報→電子図書館の『蔵書』

のように、印刷組版情報を電子図書館の『蔵書』として生かすプロセスができれば、迅速かつ容易に電子図書館の『蔵書』を充実させることができると考えられる。

しかし、現在使用されている電算写植システムは印刷に特化したものであり、そこから出力される印刷組版情報は、そのままでは電子図書館の『蔵書』として利用できるものではない。そのため、印刷組版情報を電子図書館の『蔵書』として活用できる形にデータを変換する工程が必要となってくる。

そこで当館では、BBCC（新世代通信網実験協議会）と連携して、この印刷組版情報を変換する実験を行った。出版社の作成する出版データが、電子出版や電子図書館に活用可能であることは一部では指摘されてきたが、実際にデータを入手し変換する実験は、我が国では初めてのものである。今回の実験は、出版社や印刷会社にとっても、これまでの資産を広く活用する方策を示すものであり、大きな意義を持つ実験であるといえる。

2. 出版データの提供実験

2. 1 実験の趣旨

国立国会図書館では BBCC と連携して、これまで電子図書館システムの実証実験を行ってきた。この実証実験では、電子図書館の実現に向けて「入力に関する実験」と「検索に関する実験」の両面から実験を実施している。このうち、平成 10 年度における「入力に関する実験」では、出版データの有効活用を図るため、社団法人日本図書館協会と株式会社出版ニュース社から出版データの提供を受けて、インターネット提供用データとして HTML ファイルへ変換するプログラムの開発と、インターネット上での提供実験を行った。

電子図書館推進委員会の下に設置された技術協力調整班では、その活動の一環として、その変換における技術的な課題の検証と、インターネット上での提供に対する評価を行った。また、この実験を検証する過程で、技術的な課題の他に、いくつかの制度的な課題が浮き彫りになってきた。この報告書では、実験の概要を説明し、それに付随して明らかになった技術的・制度的課題と、今後の当館の取り組みについて述べる。

2. 2 実験の概要

本実験は、『図書館雑誌』（日本図書館協会発行・平成 10 年 2 月～12 月、11 号分）と『出版ニュース』（出版ニュース社発行・平成 10 年 2 月～12 月、33 号分）のデータを用いて行った。実験は大きくインターネット提供用データへの変換実験と、モニターによる提供実験の 2 段階に分けられる。

モニターによる提供実験は、平成 11 年 1 月から 3 月にかけて行われている。この実験は変換実験によって HTML ファイルに変換した『図書館雑誌』『出版ニュース』を日本図書館協会・出版ニュース社の各ホームページ上（<http://www.jla.or.jp/>
<http://www.snews.net/>）で公開し、モニターを募り、表示データの見やすさや画面構成などの評価、課金の可能性やその方法に関するアンケート等を行うものである。

この提供実験は現在実施中であり、また、アンケートの実施や結果の解析等は日本図書館協会と出版ニュースによって行われている。その内容は日本図書館協会及び出版ニュースが近く公表する報告に記されることになっている。

2. 3 データ変換作業手順

データ変換は、次のような作業手順で行われた（資料「ワークフローおよびプログラム概要」参照）。

2. 3. 1 印刷会社からの印刷組版情報の入手

実験の当初は印刷会社から電算写植システムで作成した印刷組版情報をそのまま入手することを想定していたが、これはシステム独自の日本語コードを使用しているため、実験の初期の段階でそのままではコンピュータで扱うことができないことが判明した。そこで、印刷組版情報からコンピュータで扱える JIS コード体系のファイルに変換したもの（以下入手データ）を使用した。この変換作業は印刷会社が行なった。

2. 3. 2 作業用中間ファイルの作成

入手データから HTML ファイルに変換するためには、文書構造等を定義した汎用性の高い作業用中間ファイル（以下中間ファイル）の作成が必要となった。これは、印刷組版情報から入手データへの変換の過程で、自動変換の手掛りとなる組版情報の相当量が抜け落ち、そのままでは自動変換することができなかったためである。中間ファイルは、次の処理を手作業によって行い、作成した。

- （1）入手データの記事単位による切り分け
- （2）記事属性のチェック、記事属性を表すヘッダ部分の作成と自動変換プログラム用のタグの付与
- （3）文字の校正

2. 3. 3 HTML ファイルへの自動変換

中間ファイルから、HTML ファイルを自動生成するプログラムを用い、以下の変換処理を行った。

- （1）非 JIS コードの JIS コード（X0208）上への展開、漢数字から算用数字への置き換え

- (2) 中間ファイルのヘッダによる記事の種類の判定、その記事の種類に従うレイアウトテンプレートをを用いた HTML タグの書き込み
- (3) 中間ファイルのヘッダを利用したリンク情報の書き込み
- (4) 中間ファイルのヘッダに基づいた所定のフォルダへの保存

上記の各プロセスを自動化するために、HTML 変換用のプログラムに加え、漢数字から算用数字に変換するための置換表、記事ごとにふさわしいレイアウト表示をするためのレイアウトテンプレート、さらにリンク・ファイルの格納のためのファイル構成図などを作成した。

2. 3. 4 HTML ファイルのカスタマイズ

HTML ファイル変換後、デザイン等のカスタマイズと、レイアウトやリンクのチェックなどを手作業にて行った。

2. 3. 5 画像データの組み込み

雑誌にはテキストデータだけでなく、写真や図表などの画像データも用いられている。印刷時に使用された電子化画像データが入手できない場合は、スキャナ等による雑誌本体からの画像入力が必要になる。本実験においては、画像データはスキャナにより取り込み、2. 3. 4 の段階で HTML ファイルに組み込む形を取った。

2. 4 実験に関わる技術的課題

2. 3. 2 で述べたように、本実験においては、当初想定したよりも多くの手作業が発生したため、以下のようないくつかの中間段階が生じた。

当初の想定：印刷組版情報→（調整）→（自動化プログラム）→HTML ファイル
本実験：印刷組版情報→（出力）→入手データ→（手作業）
→中間ファイル→（自動化プログラム）→HTML ファイル

また文字の取扱に関する課題、画像処理における課題等も発生した。これらは、次に挙げる要因によるものである。

2. 4. 1 データの入手

本実験は、当初、印刷会社の電算写植システムで作成する印刷組版情報をそのまま利用することを想定していた。印刷組版情報には、いわゆる本文の情報の他、印刷実行時に必要となる行送り、書体の指定など印刷上の体裁制御に関する情報が付加されている。その制御情報を元に HTML タグ付きファイルに自動変換することが実験の当初の目的であった。しかし、印刷会社の持つ印刷組版情報は電算写植システム独自のコードを使用しているため、入手したデータはそれをテキストデータとして出力したものとなった。

この印刷組版情報からテキストデータへの出力は、それぞれの印刷会社により出力ソ

フトの仕様が異なるため、印刷組版情報の制御情報がかなり抜け落ちてプレーンテキストに近いものや不完全ながらも制御情報が残っているものなど、出力後の入手データの状態がさまざまとなった。

2. 4. 2 レイアウトの設定

インターネット上で表示する際、記事ごとにふさわしい表示をするためにレイアウトテンプレートを用意した。このテンプレートにより、HTML 変換後インターネットブラウザで表示した時の見やすさなどが決まるため、できるだけ雑誌や記事にあわせた複数のテンプレートが必要になる。ただし、一旦それぞれのレイアウトが定まれば、記事ごとに自動的に適用できるものである。また、雑誌の連載記事など、同じレイアウトの記事間でリンクを自動的に作成することも可能となる。

2. 4. 3 中間ファイルの作成

今回入手したデータは、プレーンテキストに近いものであったため、中間ファイルを作成するためにかなりの手作業が必要となった。ただし、本文の改行など記事の規則性を利用し、自動的にタグ付けを行うことができたものもあった。

まず、入手データが、雑誌1冊で1つのファイルであったため、記事毎にファイルを切り分ける作業が発生した。この作業には1冊あたり約8時間を費やした。その後、それぞれの記事属性をチェックし、記事属性を表すヘッダ部分の作成と自動変換プログラム用のタグ付けの作業を行った。特に記事のタイトルなどは、本文とは別に作成されることが多いため、入手データから抜け落ちていることが多く、印刷物である雑誌本体を見ながら入力した。この作業には1冊あたり約12時間かかった。

2. 4. 4 数字の変換

『出版ニュース』の場合、HTML への変換によって縦組みから横組みに変わるため、見易さの点で漢数字から算用数字への置き換えが必要となったが、全てを自動的に置き換えるわけにはいかなかった。例えば、日付・通貨・ページ・住所などは、漢数字から算数字への置き換えても問題ないが、文章中の地の文に現れる漢数字（「一斉に」の「一」など）は漢数字のままとしなければ意味が変わってしまうことになる。よって、漢数字の前後の文字（年・月・日・円・ページなど）により変換してよいかどうかを判断し、自動的に置き換えるプログラムを作成、使用した。このような判断の手掛りが多ければ多いほど自動化の割合は高まってくる。本実験では、置き換えが必要な漢数字のうちの大部分を自動的に置き換えることができた。

2. 4. 5 外字の変換

一方、必ず全文を付き合わせて確認をしなければならないのが、外字の問題である。

印刷組版情報から入手データを出力する際、文字コード体系が異なるため、いわゆる外字・記号など、JIS コードに割り当てられない文字は表現されない。さらに出力ソフトが数種類あるため、変換後の表示形式も様々である。「＝」のような別の文字で表現されるものもあれば、全く痕跡を残さず抜け落ちてしまうもの、文字コード番号がそのまま現れるものもある。そのため機械的なチェックはできず、雑誌本体と見比べながらの校正作業が発生し、手作業の大きな部分を占めることとなった。

2. 4. 6 画像処理

今回は印刷時に使用された画像データが入手できなかったため、画像データを印刷物から変換し、自動化プログラムの処理をした HTML ファイルに、後処理で組み込む方法を取った。

また、雑誌本体からスキャナによって画像を取り込んだ場合、画像データそのものとは発色や精細度が異なるため、色補正・解像度設定などの処理を施した上で HTML 用画像ファイルに変換する必要がある。そのため、画像データも合わせて入手できれば後処理工程の省力化が可能となる。

2. 4. 7 その他

インターネット提供用の HTML ファイルの作成後、実際にブラウザで表示し、レイアウトやリンクをチェックする作業も自動化が難しい作業である。

また、当初は、初期費用として変換プログラムや、雑誌毎の置換表・レイアウトテンプレートの作成にコストがかかり、後は自動変換プログラムを実行するだけで済むと考えられていたが、入手データから中間ファイルの作成をするためのコストがかかることが判明した。そのため、今後出版社や当館がデータを活用する場合、その作成コストを誰が負担するのか問題になると考えられる。

2. 5 今後の対応

本実験では 2. 4 で述べたような課題が明らかとなった。これらの課題に対処し、今後さらなる自動化を目指すためには、以下の点についての考慮が必要となる。

2. 5. 1 基本となるテンプレートからのカスタマイズ

現時点では、雑誌や記事ごとにレイアウトテンプレートを用意している。さらに多くの雑誌や記事を変換する場合には、基本的なテンプレートを基に、必要に応じて簡単にカスタマイズできるようにすることが必要になる。

2. 5. 2 入手データのフォーマットの統一

現時点では、印刷組版情報→（出力）→入手データ→（手作業）→中間ファイル→（自

動変換)→HTML ファイルという段階を経ており、入手データの形態も印刷組版情報とその出力ソフトにより異なっている。今後は、様々な印刷組版情報からできるだけ標準化されたフォーマットのデータを出力し、入手できること、さらには組版情報に付与された制御情報から、できるだけ手作業による前処理なしで中間ファイルを作成できるようになることが望ましい。

2. 5. 3 文字への対応

文字コードについては、シフト JIS コードだけでなく、より多くの表現力を持つ UNICODE での出力を行うことが対応として挙げられる。また、日本規格協会内の委員会より提案されている「日本語文書の組版指定交換形式」(<http://www.jagat.or.jp/>)や「拡張漢字(符号化文字集合調査研究委員会)」(<http://jcs.aa.tufs.ac.jp/jcs/>)の規格が定まれば、外字など一般のコンピュータ上で表示できない文字が減る可能性がある。

また、編集上、縦組みの雑誌であっても漢数字ではなく算数字を用いるなどの工夫をすることで、手作業の部分が省略できるものがある。このような方法は、出版物の編集過程や従来の出版物の形態に対しても影響を与えるが、「出版物の電子的利用」に対応可能なルール作りを考える一歩ともなる。

2. 5. 4 画像処理

現時点では、画像は全て手作業で処理している。これは電算写植システムが主に文字情報を取り扱うという、システム上の制約が大きいためである。今後は画像処理に関しても自動化の方策を探る必要がある。

画像のうち、写真や絵については JPEG¹などの形式で別ファイルとして持つことでそれほど問題ないと考えられるが、表については、HTML ファイルに組み込むことを考える方がよいと思われる。

さらに、本実験で使われた画像のうち、ある程度の大きさを持つ写真画像については、比較的に見やすくなっているが、小さな表やグラフは見づらいものが多い。こうしたものは、元のレイアウトに関わらず、拡大表示ができるような仕組みが望ましい。

2. 5. 5 ファイルフォーマット

本実験では、HTML をインターネット上で提供するフォーマットとして採用した。HTML の場合、データ側の環境を問わず同じ表現が可能であること、閲覧ソフトの入手が簡単であること、ネットワーク上でもローカルでも利用できることが、さまざまな利

¹ Joint Photographic Experts Group ジェイペグ。画像圧縮技術の標準化を推進している ISO と ITU-T の合同組織、およびそこで標準化された静止画の圧縮方式とフォーマット。高圧縮で知られる非可逆圧縮が、広く使われている。

用者への提供を想定する上で、大きなメリットであると考えたためである。

この他のファイルフォーマットとして、データの構造化や表示能力に特徴のある SGML²や、出版物の表現を図表の位置関係も含めて1つのファイル内に格納できる PDF³ も考えられるが、本実験は出版社が自社のホームページ上で提供することを想定し、また、テキストデータの構造化についても、自動化の部分をできるだけ多くするために最小限に止めたいということから、HTML フォーマットのみに限定した実験とした。

今後はそれぞれの目的にあわせて、フォーマットを使い分けることも考える必要がある。欧米の例では、インターネット上の出版物の配布・売買に PDF が普及しつつある。ただし、元のデータは SGML ファイルを保持しているところが多いようである。すなわち、データ入力・編集には SGML を使い、頒布用に PDF ファイルを作成しているという状況のようである。

3. 出版データの現状

2. において述べてきたように、出版データの活用において、いくつかの課題が明らかになった。これらは、技術的な制約だけではなく、出版データの特性や、出版社と印刷会社の関係など、社会的・制度的制限から生じる課題でもある。

技術協力調整班では、出版社と印刷会社に対し、出版データの現状についてヒヤリング調査を行った。以下にヒヤリング調査によって明らかになった出版の現状と課題を示す。(注：ヒヤリング調査は2社に対して行ったので、ここに書かれたことが、すべてが他社に当てはまるわけではない)

3. 1 電算写植とデスクトップパブリッシング (DTP⁴)

今回実験で用いたデータは、電算写植システムで作成されたものである。印刷業界では、現在電算写植システムと、DTP が主に用いられている。ヒヤリングを行った印刷会社では、活字から電算写植システムへの移行を 1983 年頃から始め 1986 年頃には完全に完了し、電算写植による出版が確立した。その後 1990 年代半ばより DTP による出版が台頭してきた。

DTP は文字と画像をまとめて処理できる特徴があるため、ビジュアル中心のもの、絵本などの児童書の編集においては既に主流となりつつある。文字のみのものについては、

² Standard Generalized Markup Language 汎用マークつき言語。ISO 規格にもなっているドキュメントの構造化に重点を置いた言語。

³ Portable Document Format Adobe Systems 社が開発した、PostScript ベースのドキュメントファイルフォーマット。

⁴ 編集やレイアウトの部分でコンピュータを使って出版物の編集を行うこと。

以下に述べるような理由で、電算写植が主流である。

3. 1. 1 文字

扱える文字数では、DTP は約 7,000、電算写植は約 14,000～15,000 と差があるが、DTP の文字数でも特別な場合を除き、ほとんどはこれで賄うことができるため、それほど問題ないとされる。

ただし、DTP、電算写植にかかわらず、電子データとして入校される原稿を編集システムに取り込む際、同じ文字でも、原稿作成に使用されたアプリケーションによって対応するコードが異なったり、外字の問題などがあって、どの文字が何に変換されるか分からないという問題がある。そのため、必ず全文校正が必要となる。この他、コンピュータ業界と組版業界でコミュニケーションがうまくいっていないことも問題である。

3. 1. 2 ページの表現力

ページ表現については、日本では活字時代から培ってきた組版に関するルールがある。英文では主なルールが 2 種類ほどあるだけだが、日本語においては、それぞれの出版社において独特のルールを持つ。電算写植では、ソフトでそれぞれの出版社のルールに対応できる。一方、DTP はソフトが外国で作られたものであるため、縦書きの日本語にはなじまない上にそれぞれのルールに対応するようカスタマイズしなければならない。そのため、現段階では、文字が主体のものは電算写植の方が有効であるとされている。

3. 1. 3 編集と出版の工程

電算写植では、印刷会社が原稿（打ち出し原稿と FD）から組版を作り上げていくが、DTP ではソフトさえあれば、出版社において、版下まですべてパソコンで作成し、最後の印刷工程だけを印刷会社に持ち込むことができる。このため、DTP を採用する場合、出版社と印刷会社の作業分担が変わってきている。ただし、その分担の仕方としては、編集部内で DTP 編集を行う出版社も一部にはあるが、多くの出版社ではデザイナーや他の会社に委託する傾向にあり、むしろ印刷会社以外への分業化が進んでいるというべきである。

3. 1. 4 DTP の可能性

以上のように、これまでは電算写植による出版が主流であったが、DTP による出版は去年から飛躍的に増えてきている。DTP においてもこれまで電算写植で培ってきた組版ルールなどのノウハウを蓄積していけば、文字組みだけのものも容易に扱えるようになると思われる。さらに、コストの面においても、電算写植の場合、初期投資（1 億円以上）と文字の使用料（数百万円／月）等の維持費がかかり、DTP の場合、初期投資にソフトの導入費用がかかる程度であることを考えれば、今後は電算写植から DTP に移行し

ていくと思われる。そのため、今後は電算写植システムだけではなく DTP によるデータも出版データとして活用することを視野に入れる必要がある。

3. 2 データの保管

3. 2. 1 出版社

出版社は、これまで、出版データの活用に対して、それほど意識していなかったようである。例えば、文芸関係書を主に発行する出版社では、「我々は出版社なので、電子化データなどは持たない。著者から預かった原稿を印刷所に送り、印刷所から戻った原稿を著者に返すのが、出版データに関わる部分だ。」と述べている。ただ、少数の出版社ではあるが、「自分たちが出した出版物のデータ保全を行っており、当社では電子化データを CD-R に保存している。これにより、いつでも出版物の再現を行うことが可能になっている。」というように、データの保存に関して積極的な姿勢を示すところもある。

これは出版社がその商品を印刷物としてとらえているか、あるいは情報そのものとしてとらえているかという視点によるものと思われる。現状では多くの出版社が、紙の出版物を維持することを、電子化された情報そのものを商品として流通させるよりも優先しているように見受けられる。

3. 2. 2 印刷会社

印刷会社においては、電算写植、DTP とともにデータを保存している。電算写植では、現在は、初校、再校などのデータも残しているため、フロッピーディスク (1.4MB/枚) で、段ボール (小) 1 箱くらいになるものもある。DTP の場合はデータ容量がさらに大きくなり、絵本 1 冊で 2 ギがほど必要となる場合がある。これは MO (230MB/枚) に換算すると 10 枚程度が必要となる計算である。

一口に保存といってもそれだけの量を管理しなければならない。また、ファイル管理も大切である。データは保存されているはずなのに、ファイル管理がしっかり行われていないと、探し出せないこともある。

印刷には、平面オフセット印刷の場合、印刷組版データをフィルムに転写し、さらにアルミの印刷版 (PS 版という) に転写したものを原稿として印刷するという工程をとっている。印刷会社では、このフィルムを版として保存している。例えば図書の印刷で 2 刷で訂正がある場合は、フィルムをストリップ処理 (フィルムをはがして文字を張り込む) などしてフィルムに最新の情報を保存する。ページ全体の修正などで組版データ・DTP データからもう一度フィルムを作り直すこともあるが、稀である。

保存に関する業界内のルールや法的規制はないが、ヒヤリングを行った印刷会社では雑誌は半年以上、図書は数年以上保存しており、保存の期間は出版社との個々の契約で決まるとのことであった。

3. 3 出版データの所有権

3. 2で述べたように、現状では、出版社と印刷会社では役割分担がされており、出版の企画・内容に関しては出版社に権利があり、組版に関しては印刷会社にあると考えられる。ごく一般的な契約をしている場合（組版やデータの権利の所在を明記していない場合）、組版に関する権利は印刷会社にあるとされる。出版社がデータを引き上げるときには有料にしたり、先に出版社とデータを引き渡すことを条件にする契約を交わすこともある。しかし、これは、印刷会社が組版について無条件の権利を持つわけではない。印刷会社は、出版を企画した出版社の依頼によらなければ、印刷や、第三者に譲渡することはできない。そのため、例えば、今後図書館など第三者が出版データを入手するとすれば、直接印刷会社から入手するのではなく、出版社の許可を得て、出版社経由で入手するしかないと思われる。

3. 4 出版流通への影響

出版データの活用は、今後の出版流通にとっても大きな影響を与える。出版流通業においては、電子書籍など電子出版物の可能性を模索しているところであり、流通コストや在庫圧力に対する手段として、出版データの活用は効果的な問題解決に発展する可能性がある。例えば、雑誌のバックナンバーや、入手が困難になった過去の出版物を電子的な形態で容易に手に入れられるようにし、過去に出版された各種の印刷物が再編集や別のメディアによって利用できるようにすることが想定される。

しかし、急激な変化は既存の出版流通の仕組みを変えることとなり、対応が難しい出版社が出てくる可能性なども否定できない。また、既に印刷組版情報をHTML化したり、DTPデータからSGML・HTML・PDF等に変換するソフトはあるにもかかわらず、大規模な利用に至っていない現状もある。この点に関しては、先にあげた流通の問題、データの所有の問題と関連して慎重に見極めをする必要がある。

また、本実験がそのまま生かされるわけではないが、電子化されたデータが商品になり得るという視点が生まれ、今までのように大量部数を印刷、販売して収益を得る形態から、オンデマンド出版など、必要なときに少量部数の印刷をしても十分コストの回収ができるような販売戦略をとる可能性もある。

4. 国立国会図書館における課題

本実験は、出版社によるデータ提供の視点で変換を行ったものである。今後、出版データを電子図書館の『蔵書』として取り扱う際には、多種類のデータを大量に取り扱うようになる。そのため、出版社がそれぞれ独自に変換する場合とは別の課題が想定される。

4. 1 技術的課題

4. 1. 1 汎用性の高い変換プログラム

出版組版情報から出力されるデータには、複数の種類がある。当館では、多種類のデータを取り扱うことが想定されるので、多くのフォーマットに対応できる変換プログラムで、カスタマイズの工程ができるだけ少なくなるものが望ましい。

4. 1. 2 簡便なレイアウト

本実験では、見栄えの良さを重視したため、それぞれの雑誌・記事に対応する形にレイアウトテンプレートを作成したが、取り扱う資料によっては、元の紙媒体のスタイルにこだわらずに、一つのレイアウトテンプレートに統一したほうが、統一感を与え好ましい場合もある。またその方が手間やコストがかからないという利点もある。

4. 1. 3 検索と管理

本実験では取り上げていないが、当館で多量のデータを扱う場合、検索・管理に対する配慮が必要になってくる。今回は目次情報から記事を選択する方法がとられたが、当館で活用するためには、タイトル、著者、その他キーワード等による検索を可能とすることが必要条件となる。また、テキストデータであるメリットを生かし、全文検索を可能とすることも、データによっては必要である。

そのためには、書誌的な事項をデータに取り込む仕組み、データから検索に必要な情報を抽出する仕組みを作り上げることが求められる。今回手作業で挿入したヘッダー部分をメタデータとして取り扱うことができるようなフォーマットにしたり、ファイルの構成等を考慮し、管理ができるようにしたりすることも考慮されるべきである。

また、利用に供するためには、全館の書誌情報のあり方と整合性を持った形で運用されることが望ましい。

4. 1. 4 データ保存に関する課題

電子情報の保存については、本班において別途報告書が作成されているが、同様の配慮をするべきである。

4. 2 制度的課題

4. 2. 1 出版データの位置づけ

既に発行された出版物と、それと同じ出版データから作成する出版物との関係は、出版データの活用主体により異なってくる。出版社が出版データを活用する場合は、その

ままインターネット上で提供するだけでなく、著作者との権利関係を逸脱しない範囲内で、再編集などの処理もでき、タイトルも変えて流通させることが可能であるため、既に発行された出版物とは別の出版物を作成することができる。

しかし、当館が活用する場合は、紙媒体の出版物を電子化する場合と同様と考え、既存の出版物との対応関係を維持すべきである。出版データは、テキスト形式での入手が可能であるため、その一部を再編集したり加工することは容易であるが、当館はそれを行うべきではない。

4. 2. 2 納本制度との関係

冒頭で述べたように、出版データは納本の対象とはならない。そこで、必要なデータを選択的に契約等に基づき購入したり寄贈を受けたりすることが収集の方策となると考えられる。その場合は、どの出版社のどの出版物が、出版データを利用可能な形で所有しているかを知る手だてが必要となってくる。あるいは、電子図書館の『蔵書』整備を進める上で、当館からの積極的な働きかけを検討する必要がある。また、出版データの対象は民間出版社に限らないため、行政機関、学協会、大学、試験研究機関等からのデータ収集の方法としても考慮されるべきである。

4. 2. 3 変換の体制

実際に当館が出版データを入手し活用する場合、いくつかの方法が考えられる。それぞれの方法について、十分検討する必要があるが、現時点で考えられる方法をあげた。

- (1) 出版データは出版社が持ち、当館はゲートウェイとしてナビゲートする方法
- (2) 当館が変換ソフトを配信し、出版社等により変換されたデータを受け入れる方法
- (3) 第三者（企業など）が委託等により変換を一括して請け負う方法
- (4) 出版社から出力データを受け入れ、当館が変換する方法

(1) から (4) の順で当館が主体的に出版データに関わる比率が大きくなり、コスト負担も増大する。(2) の方法でデータを手入手する場合は、変換ソフトの作成費用を除いては当館におけるコストはそれほどかからないが、出版社の負担は大きくなる。

(3) や (4) のように、当館が主体的に変換をする場合は、当館においてコストはかかるが、必要なフォーマットへの変換あるいはデータの作成なども、当館が決定することとなる。一方、出版社も、それぞれ独自の方法でフォーマット化し、任意の媒体で商品としての情報を流通させることができる。当館では出版社の領分を侵すことなく、データを活用する方法を考えるべきである。

コストについては、現状では紙媒体の資料をスキャンなどの方策で電子化するコストと比べて割高になるが、変換プログラムが標準化・高度化し、自動化の割合が増えれば、変換のコストは低下することになると考えられる。

ただし、他の方法での電子化同様、著作権の処理や利用許諾契約等については十分な配慮が払われるべきである。

5. 提言

本実験では、変換の過程で手作業が必要となったため、当初の想定に比べてコストがかかる結果となったが、それでも全てを手作業で行う場合の参考価格に比べれば安価にテキストデータを入手できることが判明した。また、テキストデータは、従来の画像データに比べ容量が小さいこと、文字が読みやすいこと、全文検索などができることなど利点は多い。

本実験により、出版の過程で作成される出版データを、電子図書館の『蔵書』として、また出版社が活用するデータとして生かす可能性が示された。今後は、目的に合わせたフォーマットの使い分けを行うために、今回用いた HTML フォーマットのみならず、SGML・PDF への変換実験を行い比較調査することも必要である。

出版データから生成される情報が商品としての価値を持つようになると、電子図書館の『蔵書』として、出版社が無条件で当館にデータを提供することは難しく、著作権・課金・料金決済の方法・費用負担の問題など、多くの解決すべき課題が生じる。しかし、民間出版物であっても、電子図書館の公共性を鑑みデータの提供をする出版社は少なくないと思われる。あるいは、文化的価値は有するが、商品としては価値の低くなったデータを電子図書館の『蔵書』とするなどの方法をとれば、出版社と当館での出版データの共存を図ることは可能であると考えられる。

一方、非営利出版物は電子図書館での活用が有望な分野である。官庁出版物や各種学術情報などは、これまでに示した方法で利用に供することが民間出版物と比して、比較的容易であると考えられる。

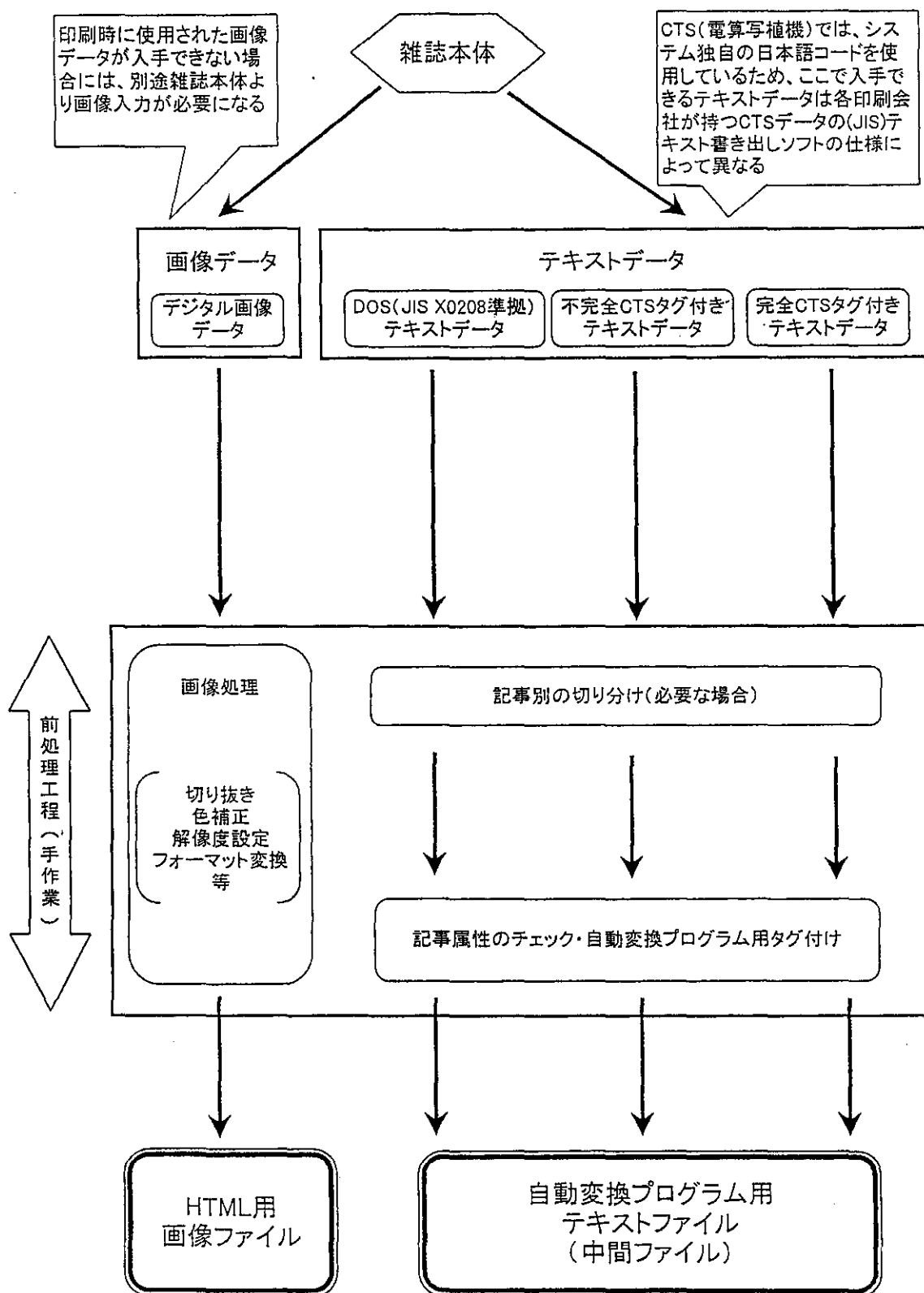
出版データの活用に国立国会図書館が関わっていく場合、単に契約上の問題や権利処理の問題として考えるに止まらず、社会・文化的な図書館の意義や日本の出版流通をどのように進化させていくことが望ましいか、といった広い視野が必要になってくる。当館は広く情報の保存と流通に関わる図書館の役割の再構築として積極的に模索するべきであろう。

<参考文献>

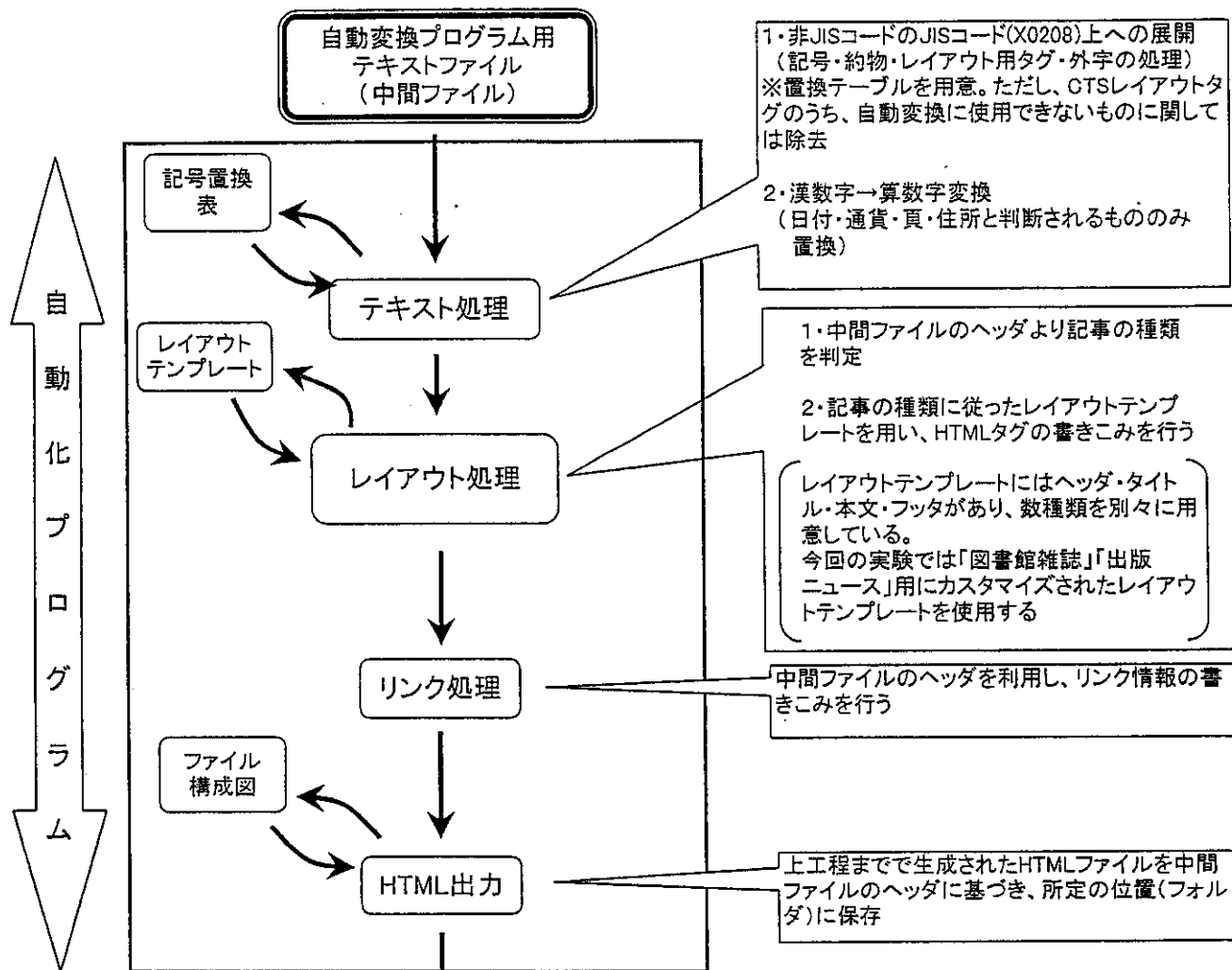
- ・ 田屋裕之 出版データのネットワーク上での提供実験について 『出版ニュース』 99.2/上, p6.-9, 1999
- ・ 川畑和夫 ネットワークでの図書館・電子出版の技術 『出版ニュース』 99.2/中, p.6-9, 1999

<資料> ワークフローおよびプログラム概要

1. 前工程



2. 自動処理



3. 後工程

