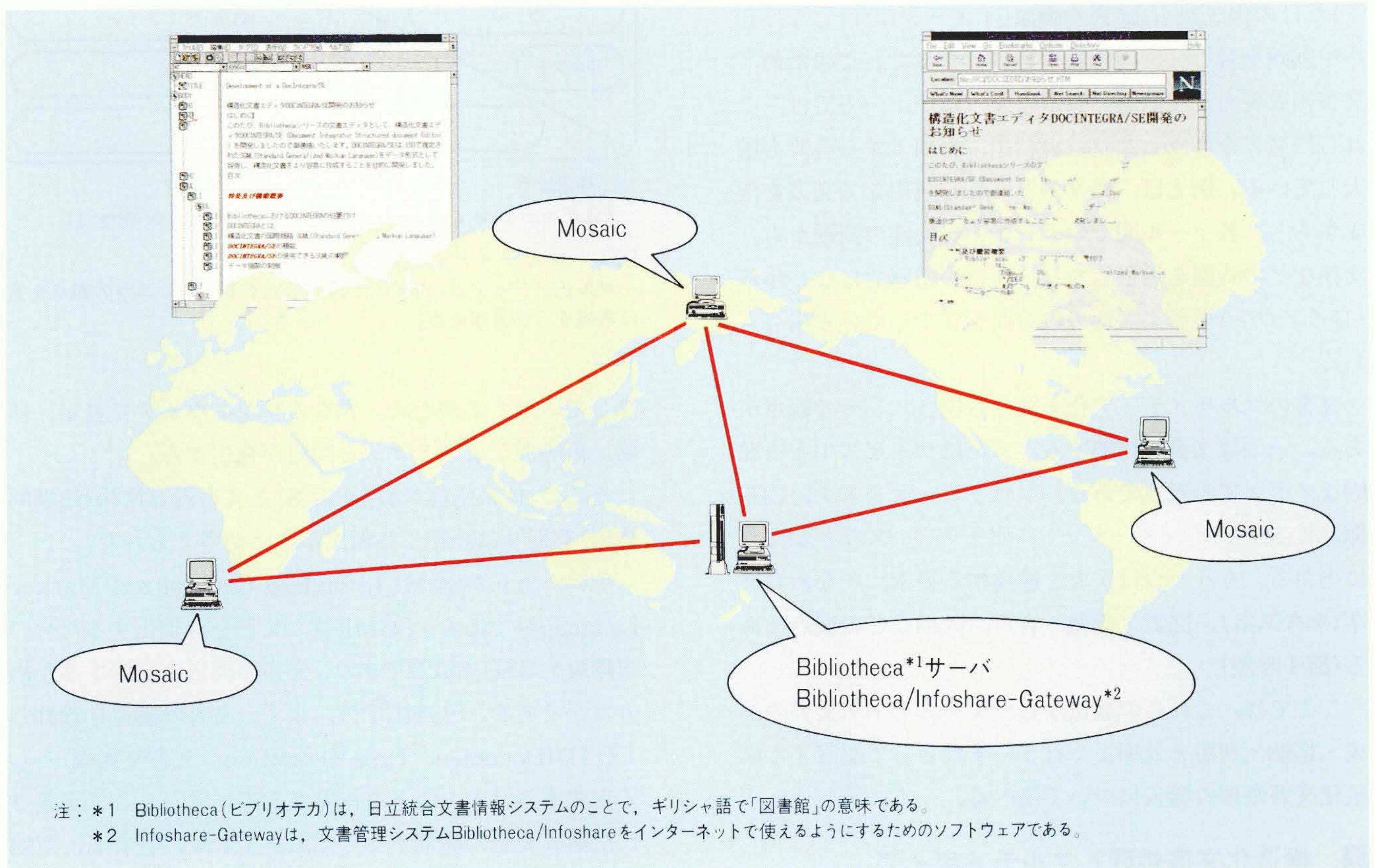


構造化文書処理の適用によるマルチメディア文書の作成

Multimedia Document Creation by Structured Document Processing

山崎紀之* Noriyuki Yamasaki

星 幸雄* Yukio Hoshi



Mosaicの出現によるインターネット利用者の増大

この1年間で、インターネットによる情報受発信が急激に増大した。これは、それまでの文字(テキスト)だけの情報から、イメージを主とした動画や音声などのマルチメディア文書に比較的簡単にアクセスできるブラウザMosaicの出現が大きく影響している。また、アクセスを容易にするために、テキストデータをHTML(Hypertext Markup Language)によって共通化したことも大きな要因である。

近年、企業内ではほぼ1人1台のワープロ(ワードプロセッサ)やパソコン(パーソナルコンピュータ)が普及し、文書の電子化が進んできた。その電子化も当初の文字中心の文書作成およびその清書目的から、イメージ、音声、動画などを含むマルチメディア文書での表現力向上および理解を支援する目的に変化してきた。これらを格納または提供する物理的なメディアもフロッピーディスク、CD-ROM、通信、インターネット、データベースなどさまざまである。したがって、一つの電子化文書を種々のメディア上で利用できるようにしておくことが重要であると思

われる。また、オフィスワークの大部分を占める文書作成の効率向上を図ることは、現状での大きな課題の一つである。

これらを解決する手段の一つとして、構造化文書処理の導入を提案したい。マルチメディア文書を構成する各要素をオブジェクトとしてとらえ、別々に作成されたものを組み合わせて文書を構成するものである。これによって、文書の作成効率を向上させるとともに、インターネットなどでの情報発信や各種メディアへの同一データ形式での格納などができるようになる。

* 日立製作所 ソフトウェア開発本部

1 はじめに

ワープロやパソコンの低価格化により、オフィスでは1人1台の情報機器が普及してきた。機器やソフトウェアの高機能化によって電子化文書は、当初の文字(テキスト)だけの作成から、図表や画像(イメージ)だけでなく音声や動画を含む文書の作成も容易になった。このため、文書の表現力が圧倒的に豊富になったが、一方では、これらの文書を作成したり、再利用、再加工する時間も増大している。例えば、種々のツールを利用して文書を作成すると、各ツール間でのデータの交換性の問題から、変換などの時間を浪費したり、文書の内容ではなく体裁(レイアウト)を整えるために時間を費やしている場合もある。

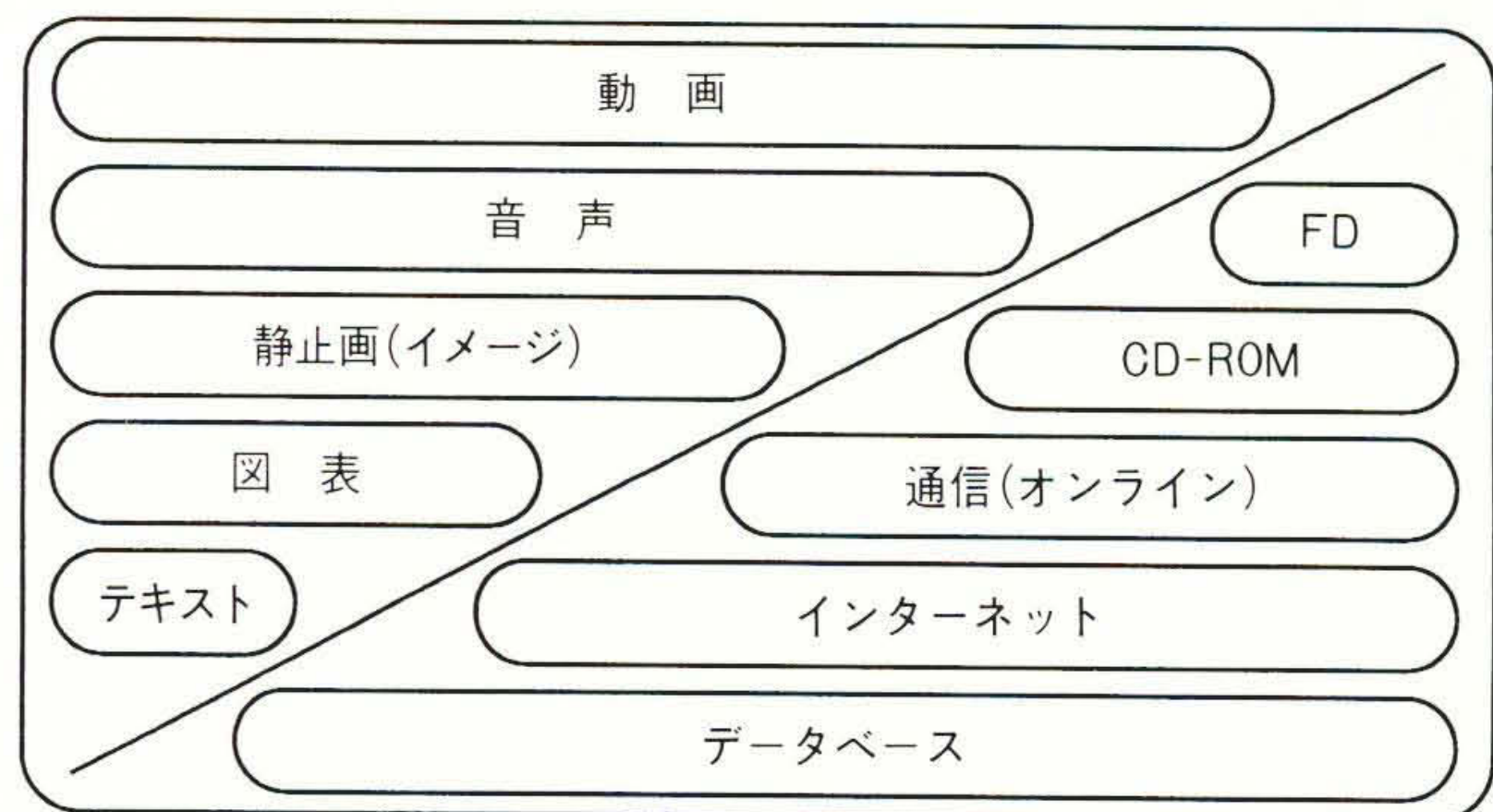
文書のマルチメディア化を論じる場合、二つの観点がある。一つは文書を格納する、またはサービスする物理的なメディアの観点である。フロッピーディスク、CD-ROM、通信、インターネット、データベースなどがこれに当たる。もう一つは文書を構成する要素、すなわち文字(テキスト)、図表、画像、音声、動画などの観点である(図1参照)。

ここでは、これらの観点からマルチメディア文書の作成・格納・利用を効率よく行う一手段として提案する構造化文書処理の導入について述べる。

2 構造化文書処理とマルチメディア

2.1 SGMLによるアプローチ

従来、一般のワープロで文書を作成する場合、文書データの内部形式が、使用するワープロソフトウェアに依



注：略語説明

FD (Floppy Disc), CD-ROM (Compact Disc Read-Only Memory)

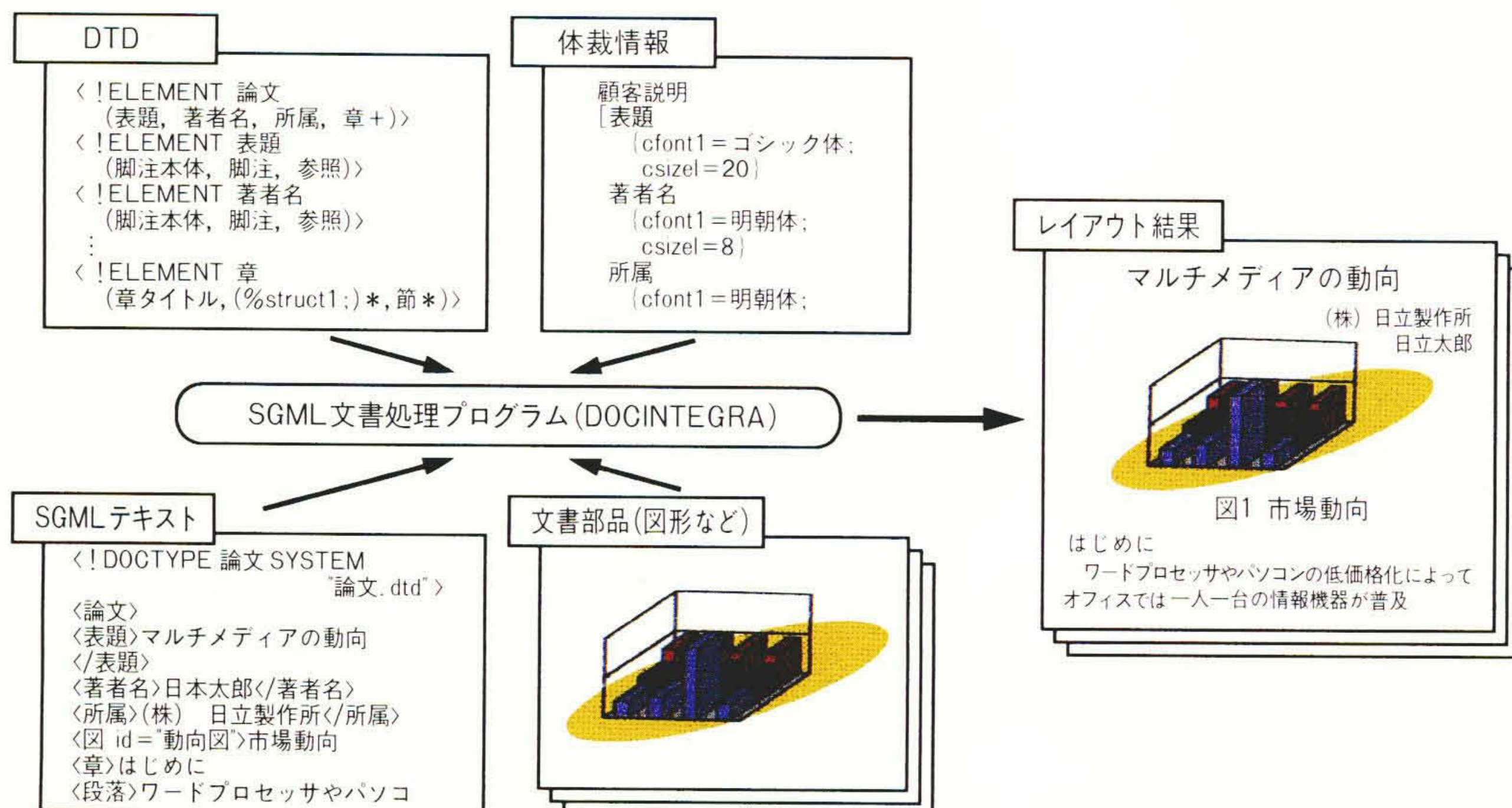
図1 マルチメディアの観点

マルチメディアは、内容とそれを格納する媒体の二つの観点を常に考慮する必要がある。

存していた。このため、異なるソフトウェアで表示、印刷、再処理しようとする問題が発生する。

そこで、体裁情報(物理的情報)と文書内容情報(論理的情報)を別々に管理し処理することが考えられた。

その一方式がSGML(Standard Generalized Markup Language)である。SGMLは、文書を電子化するための国際規約(ISO 8879)であり、文書の構造を記述するための言語である。SGMLでは、まず、文書の構造を設計してDTD(Document Type Definition: 文書型定義)として定義する。テキストを作成する際には、ここで定義された構成要素に識別子として、タイトル、著者名、章題といったタグと呼ぶマークを付ける。SGMLテキストを処理するプログラムでは、このタグを判定して文書の構成要素ごとに体裁などを指定し処理する(図2参照)。このように、文書の論理構造・内容と文書の体裁情報を分



注：略語説明

DOCINTEGRA (Document Integrator; ドックインテグラ)

図2 SGML文書処理

SGMLテキスト中の「表題」といったタグを判定し、体裁情報を参照してテキストおよび文書部品をレイアウトする。SGMLテキスト中のタグはDTDによって構造が決められている。

離して処理することにより、従来のワープロソフトと比較して次に述べる利点を得られる。

- (1) 体裁情報がシステムに依存していないので、紙媒体への印刷、CD-ROM出版など多様なメディアへの変換が容易
- (2) 画像、音声、動画などのマルチメディア情報の管理が容易
- (3) ハイパーテキストの記述が容易
- (4) 文書作成効率向上が容易
- (5) 検索、情報抽出など他のシステムとの連携が容易

2.2 構造化文書処理システムDOCINTEGRAでの実現

DOCINTEGRA(ドックインテグラ)は、SGMLに基づく文書処理プログラムである。SGML文書処理モデルでの文書作成・編集処理、ブック構成登録処理、自動レイアウト処理、および文書表示・印刷処理に対応した製品群で構成している(図3参照)。この製品では、SGML採用による利点を発展させて、文書の部品化、ハイパーテキスト化などの機能を実現している。

(1) 文書の部品化

文書の基本部分は、ワープロソフトに代えてDOCINTEGRA/SE(Structured Document Editor)という構造化文書エディタを用いて作成する。65ページの図に示すように、DTDで定義された文書構造が左側に表示され、これに従って、執筆者は、書体やレイアウトを意識せずに内容だけを作成し、必要な図・画像なども別途部品として作成していく。

DOCINTEGRA/BB(Book Browser)が行うブック構成登録処理は、図・画像などのデータとSGMLテキスト中のタグとの対応づけを行う処理である。図・画像などのデータはMicrosoft^{※1)} Windows^{※2)}でのOLE(Object Linking and Embedding)機能を使用して対応づけている。OLE機能とは、例えばEXCEL^{※3)}で作成した表を

WORD^{※3)}の中にリンクしたり、埋め込んだりする機能である。

(2) ハイパーテキスト機能

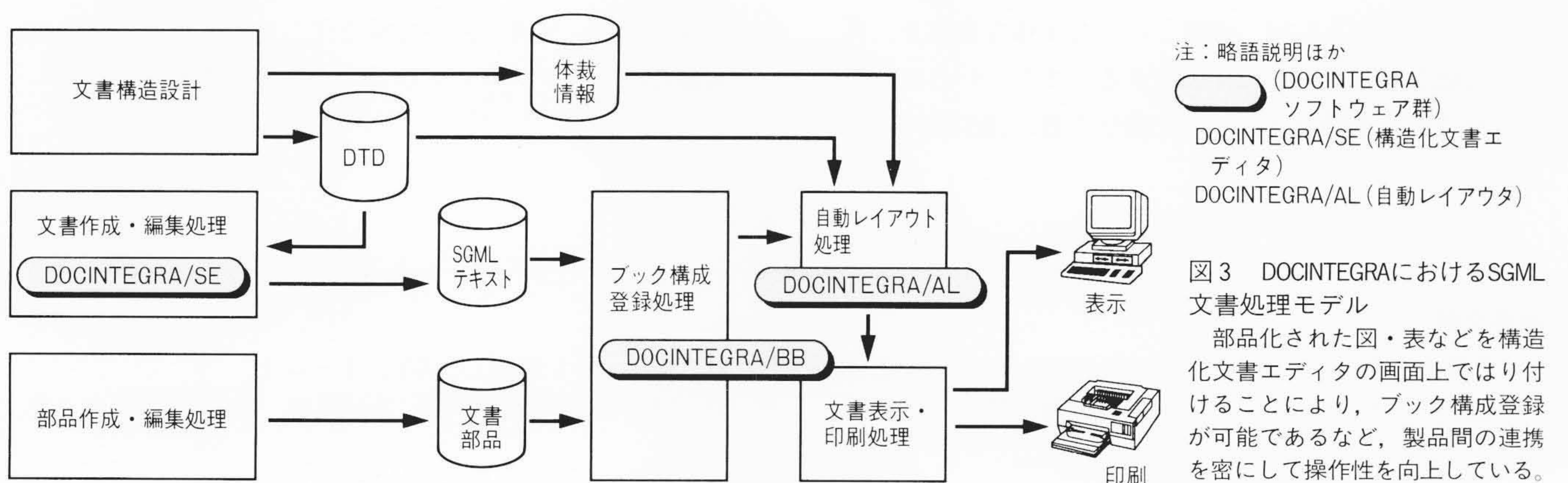
電子化された文書のメリットの一つに、ハイパーテキストと呼ぶ関連情報間のリンク機能がある。文書作成・編集処理でタグを利用して関連づけたりリンク情報は、文書の編集集中に関連部分へのジャンプ(リンクフォロー)を可能にするだけでなく、レイアウト後の文書表示画面でのリンクフォローも可能にしている。また、レイアウト後の文書表示画面上では任意の位置にリンクを付加する機能を実現している。このリンクでは文書データとは別のファイルにリンク情報を保持するため、CD-ROMなどの読み出し専用のデータに対してもリンクづけができる。

3 インターネットはマルチメディアの世界

インターネットということばが一般的になってきた要因は二つある。その一つは、当初その利用が学術的な用途に制限されていたものが、ビジネス利用も可能になったことである。二つ目は、現在インターネット上の標準的なビューアであるMosaicが開発されたことである。Mosaicは、それまで文字を基本としたデータの送受信から、画像データまでを含むデータを標準で扱えるようにした。また、動画や音声もクライアント(パソコンやワークステーション)側に対応するソフトウェアがあれば、データにアクセスして見たり聞いたりできる。このため、利用者層が大幅に増大した。

Mosaicが処理するデータは、SGMLのサブセットであ

- ※1) Microsoftは、米国Microsoft Corp.の登録商標である。
- ※2) Windowsは、米国Microsoft Corp.の商標である。
- ※3) EXCEL, WORDは、米国Microsoft Corp.の商品名称である。



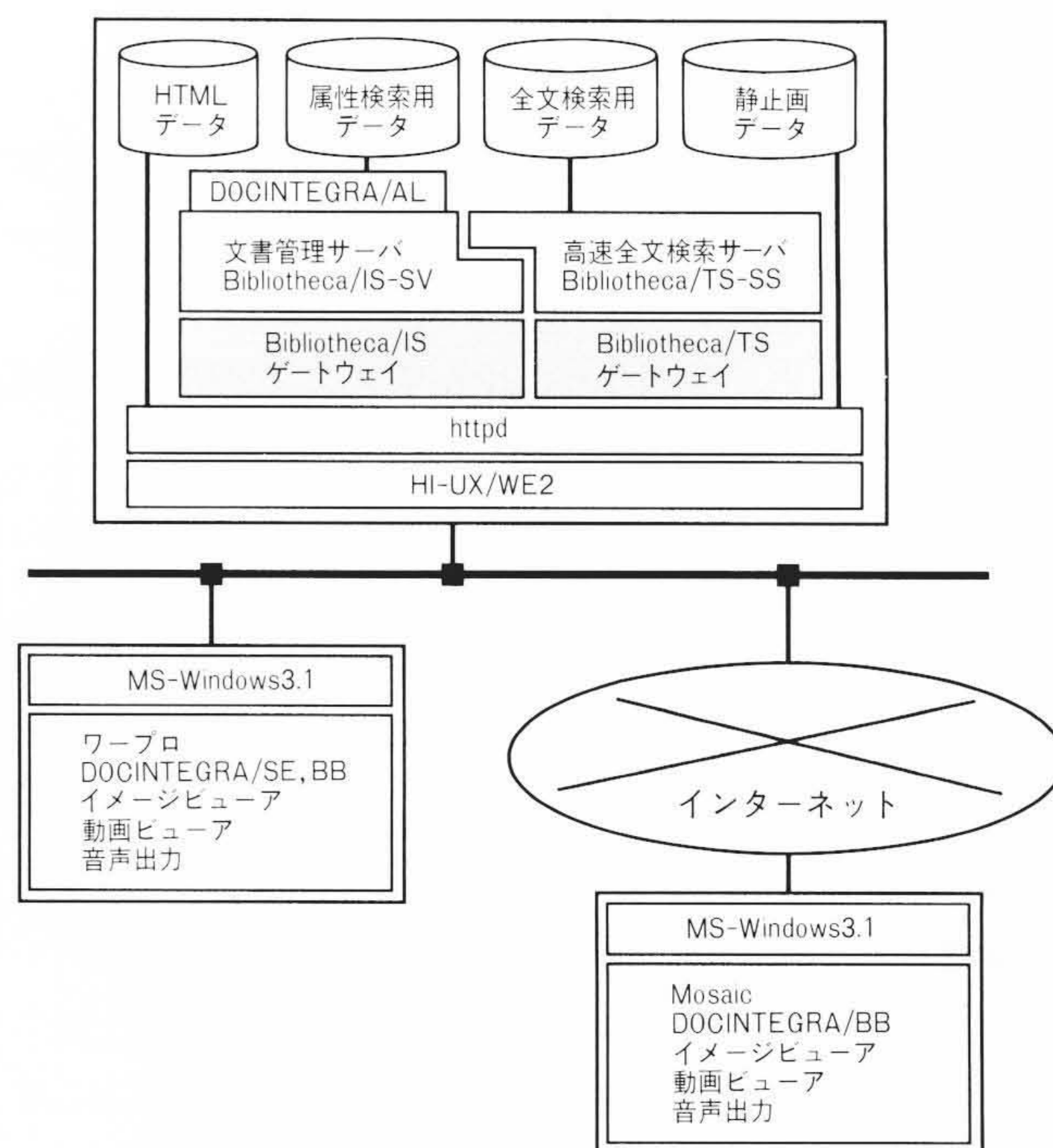
るHTMLで作成されたデータである。HTMLでは、利用できるタグ名が約40種に決められている。MosaicはHTMLデータを入力して、その中のタグに従って標準的なレイアウトを行って画面に表示する。また、電子化文書では必須(す)の機能であるハイパーテキストの機能を持っている。このため、種々のマルチメディアデータを一つの文書中にすべて取り込むのではなく、所在データだけを保持しておく。利用者はそのデータ実体を参照したい場合、リンクの存在が示されている個所(アンカー)をマウスで指示する。Mosaicは、リンクされたデータの存在するサーバと通信し、データを読み込んで表示する。

インターネット上でのマルチメディア情報の提供およびアクセスを容易にするために、文書情報管理システムBibliotheca/IS(Infoshare)および高速全文検索システムBibliotheca/TS(Text Search)を利用できるようにした。これらをMosaicから利用できるようにするために、WWW(World-Wide Web)サーバとBibliotheca/ISおよびTSをつなぐゲートウェイプログラムを開発した(図4参照)。このプログラムは、Mosaic上で入力されたデータをWWWサーバを介して受け取り、これを解析してBibliotheca/ISやTSに蓄積されたデータにアクセスし、結果データをHTMLに編集して、再びWWWサーバを介してMosaicに返し表示させる。

このシステムを利用すれば、すでに蓄積された情報をMosaic用にHTMLに変換しなくても、Bibliotheca/ISのデータベースに登録するだけで世界へ向けて発信できる。これを企業内で利用すると、分散した事業所での情報共有が容易に実現できる。

4 おわりに

文書のマルチメディア化の進展に対し、内容の作成・格納・利用のためには構造化文書が有効であることについて述べた。SGMLを基礎とした各種ツールを活用することで、テキスト、図、画像といった多様な表現を、再利用可能な形式で容易に作成できる。また、いったん作成した文書を多様なメディアに載せて表示、配布できる。



注：略語説明ほか

httpd (Hypertext Transfer Protocol Daemon)

HI-UX/WE2 (Hitachi UNIX*/Workstation Extended 2)

* UNIXは、X/Open Company Limitedがライセンスしている米国ならびに他の国における登録商標である。

図4 インターネット環境でのBibliothecaシステムの利用
Mosaicを利用し、インターネットを介して、Bibliothecaに格納されている文書にアクセスできる。

一方、マルチメディア文書は紙の文書とは異なるという認識が必要である。すなわち、現状の紙の文書、特にマニュアルなどをそのまま電子化しても利用しやすい文書にはならない。マルチメディア文書を参照する場合、紙の文書のようにぱらぱらとめくって全体を一覧するようなことはできないし、同時にあちこちを参照することが難しい。

これをある程度解決する機能がハイパーテキスト機能である。マルチメディア文書を作成する場合、この機能をうまく利用した文書作りが必要である。これを実現するうえでも、文書の内容を構造化してとらえる、構造化文書処理へのアプローチが必要であると考えられる。

参考文献

1) 和歌山, 外: クライアントサーバで実現する統合文書情報システム—Bibliotheca—, 日立評論, 77, 5, 367~370(平7-5)

2) 根岸, 外: SGMLの活用, オーム社

3) MARTIN BRYAN著, 山崎監訳, 福島訳: SGML入門