図書構造化入力システム「情報ファクトリ」の提案

2T－7

大門秀章　石田和生　神谷俊之　國枝和雄
NEC 関西C＆C研究所

1.はじめに

最近、世界中の図書館で大量の蔵書を電子化し、多くの人々が手近なコンピュータから自由に閲覧できるような仕組み、すなわち電子図書館を構築しようとする動きが活発になっている[1][2]。電子図書館の構築には様々な技術の開発が必要であるが、大量の蔵書（雑誌、図書、論文、新聞などこれまで所蔵されてきた本）を電子化する入力の問題は、本質的な問題として捉える必要がある。

筆者らは電子図書館構築の要素技術として大量蔵書の電子化に主眼を置き、紙ベースの製本された本を対象に、大量の既存文書を遡及入力するためのシステム「情報ファクトリ」を提案する。

2.図書入力システムに要求される機能

従来、オフィス文書や画像の電子化は、スキャナや文字認識装置等、幾つかの処理装置を併用して行われてきた。しかし、大量蔵書を対象とした場合、入力作業の随所で発生する手間を極力減少させる必要があるため、従来のシステムをそのまま利用することは難しい。大規模図書入力システムには以下のような機能が要求される。

・スキャナ入力からデータベース蓄積までを包括した統合システム

入力から蓄積まで一連の作業を行える統合システムは、各処理間のデータ変換などの手間を減少させるだけでなく、全体をひとつのユーザインタフェースで操作することによって、作業時間と作業労力の両方を改善することが可能となる。

・作業負担を軽減させるユーザインタフェース

作業状況を常に画面上に表示し、作業手順をナビゲートするようなユーザインタフェースが必要。入力作業を行うのは、図書館員やアルバイト等のコンピュータに不慣れな人が多いと考えられるため、簡単に操作できるユーザインタフェースが必要である。

・誰でも一定品質の入力を行えるようにする

入力者の経験等により、入力品質にばらつきが出ないよう、誰が操作しても一定品質が得られるような仕組みを作ることが必要になる。各処理部での設定パラメータを適切なものに自動的、あるいは簡易的に設定する機能が必要である。

・入力作業の分散/協調作業を支援する

図書の入力は図書館を中心に全国的に作業を分担をして行うことになるため、どこでどの本を電子化しているかを把握しておく必要がある。多地点での分散入力、多人数による協調入力を支援するための機能が必要である。

3.「情報ファクトリ」の試作

筆者らは2.の要求を満たすことを目的とした図書構造化入力システム「情報ファクトリ」の試作を行っている。ただし、現段階では分散/協調入力の機能は含んでいない。

3.1 システム構成

試作システムの構成を図1に示す。システムはパソコン上で動作する。

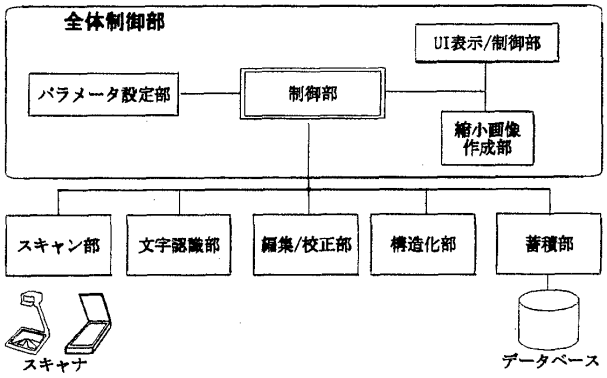

図1　システム構成

試作中の図書入力システムは以下の手順でデータの入力・蓄積を行う。

1. **スキャナ入力**[3]　図書をスキャナで画像として取り込む。文字認識を行うのため320～400dpiで撮像する。
2. **文章レイアウト解析**[4]　文章レイアウトを解析し、文章、図表領域の切り出しと、ページレイアウト情報の抽出を行う。
3. **文字認識**[4]　文章領域に対して文字認識を行い、目次、本文のテキスト変換を行う。

Proposal of the books digitizing system "Information Factory"
Hideaki DAIMON, Kazuo ISHIDA,
Toshiyuki KAMIYA, Kazuo KUNIEDA
Kansai C&C Research Laboratories,NEC Corporation

4. **誤認識文字編集/校正**　誤認識した文字、及びレイアウトの編集と校正を行う。
5. **文章構造化**[5]　文章構造を解析し、構造化テキスト(SGML テキスト)に変換する[1]。
6. **蓄積**[6]　構造化テキスト及び書誌情報をデータベースに蓄積する。

イメージデータを文字認識し、章タイトル、段落等の文章構造を含んだ構造化テキストデータ(SGML 文書)に変換して蓄積する。また、本には視覚的効果を考慮したレイアウトがあるため、レイアウト情報からも検索できるよう、ページのレイアウト情報も合わせて蓄積する。これにより、入力図書のページの再現や、「右上に図のあるページ」といった検索を行うことができる。

3.2ユーザインタフェース

ユーザインタフェースは、簡易入力用と一般入力用の２種類を用意した。簡易入力用インタフェースは、入力から蓄積までを、図書入力ウィザードによる流れに従って行えばよいようデザインした(図２)。作業手順を常に画面の一部に表示し、現在どの作業を行っているかを作業者に伝えている。一般入力用インタフェースでは、アイコンを流れ作業順に並べたアイコンチャートを使用し、次に行うべき作業をナビゲートするとともに、作業状態を常に確認できるよう入力された画像は縮小画像として一覧表示している(図３)。

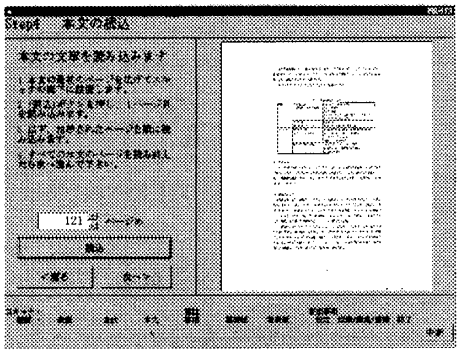

図２　簡易入力用ユーザインタフェース

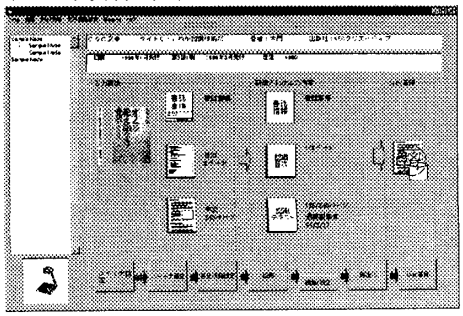

図３　一般入力用ユーザインタフェース

画像入力は表紙から始まり、目次、本文、書誌事項、裏表紙、背表紙の順にページをめくって行う。目次や本文などは見開き両ページ同時読み込みとし、スキャナに本を設置する労力を軽減している。書誌事項(タイトル、著者、ISBN 番号など)の登録を行った後は、レイアウト解析、文字認識、構造化、ＤＢ蓄積を自動的に行っている。

3.3パラメータ設定

入力する本を、その図書の入力上の構造から分類し、分類毎に各処理で使用するパラメータ情報を持たせることにより、同一分類に属する図書は、同じパラメータ設定で入力作業が行えるよう構成した。図書の入力上の構造とは、本のサイズ(A4,B5 等)、紙質(普通紙、光沢紙等)、使用言語(日英)などのことである。例えば、シリーズになっている文庫本や、月刊誌の各号などは、同じ入力上の構成を持っているため、同一分類に分別できる。図書１冊毎ではなく、分類毎に各処理のパラメータを与えることによりパラメータ設定にかかる手間を減少させている。

4．おわりに

大量蔵書を遡及入力するための図書構造化入力システム「情報ファクトリ」の提案を行い、簡易入力用、一般入力用の２種類のユーザインタフェースを実装したシステムを試作した。

簡易入力用インタフェースとパラメータ設定の簡易化により作業効率は向上したが、まだ快適な作業環境とは言えない。画像入力手法の改善、各処理にかかる時間の短縮などのシステム面と、パラメータ設定の自動化、効率の良い編集/校正環境の構築等のインタフェース面の改良が必要である。

今後はシステム及びユーザインタフェースの評価・改良、ネットワークを用いた分散/協調入力作業の実現を行いたい。

参考文献

[1]市山　他,「多様な情報源を対象とする WWW ベース電子図書館システム」,デジタル図書館(第７回),pp32-50,1996
[2]芹沢:「仮想電子図書館と個人情報環境」,デジタル図書館(第 6 回),pp11-21,1996
[3]柏谷,瀬川,辻澤:「平面ミラー回転走査型イメージスキャナ(第２報)」,第７２期通常総会講演会論文集,Vol.Ⅳ,p77,1995
[4]辻,「スプリット検出法による文書画像構造解析」, 信学論, Vol. J74-D-II, No. 4, pp.491-499, 1991.
[5]石田,市山:「既存文書のレイアウト情報付き構造化手法」,第 53 回情処全大,3S-05, 1996
[6]波内:「OODB による SGML 文書データベースの設計」, 情処 DBS 研究会第 109 回予稿集, July 1996.