

バインダ文書管理システム

Binder Document Management System

山田 彰 *
Akira Yamada

田甫正治 *
Masaharu Tanbo

* ソリューションビジネス本部 営業・SE 第三グループ 第一システム統括部 第三システム部

イメージスキャナを応用した紙文書電子化ソリューションの一環で文書管理システムを開発した。特に、紙文書を扱うということで、読み取った文書に対しフリーフォーマットで文字認識を行い、さらに全文検索を可能とした。

また、開発の面においても、すべて独自に作り出すのではなく、他社製品で利用できるものは利用し、またフリーで使用できるソフトは積極的に採用した。これにより、安価で拡張性の高いシステムを構築することができた。

The document management system has been developed as part of an electronic document solution using image scanners. Since this solution focuses on paper document conversion to an electronic format, a special effort has been made on the free format character recognition, also a full document search has been enabled.

As for development, rather than developing our own software anew, the approach was to utilize existing software developed by other companies, as well as to positively take advantage of free software. This allowed us to achieve the implementation of an inexpensive highly expandable system.

1 まえがき

「ペーパーレス」が叫ばれて久しい。しかし、オフィスから紙がなくなる様子はなく逆に紙の出荷量は年1～2%程度増えている¹⁾。紙文書は頻繁に使っているうちは取り扱いやすく非常に便利であるが、参照頻度が減り記録でしかなくなると、場所を取りまた検索性も悪く「お荷物」になってしまう。このような紙文書にかかわる悩みの解決に役立つのが、当社の強みであるイメージスキャナを活かした紙文書電子化ソリューションである。筆者らは、SEの立場から紙文書電子化システム構築の核となるようなデータベース中心の文書管理ソフトを開発した。

伝達のための文書と記録を残すための文書である。前者は通知書類や回覧などである。電子メールなどもこれに入る。これに対して後者はいわゆる「棄てるに棄てられない」という文書で契約書や伝票、各種記録などである。

情報伝達文書の役割は、如何に迅速に、広範囲に、見やすく内容を伝えられるかである。そのために、楽²ライブラリ²⁾、^{注1)}のようなグループで文書を共有するためのシステムが用意されたり、ポータルサイトを設けて共有文書閲覧場所を設定したりされてきている。

他方、記録文書は増える一方のため量との戦いになる。したがって、記録文書を電子化して保管スペースを大幅に削減するために簡単に電子文書として登録できる事と、効率よく探し出せることが最重要課題になる。この

2 開発の背景と狙い

(1) 文書保管の考え方

オフィスの文書は大きく2種類に区分できる。情報

注1) 楽²ライブラリは、オフィスの書棚をPC上に再現したデジタルバインダで、紙文書のバインダへの登録や文書閲覧時のパラバめくり、付箋紙への書き込み、マーキングなどの操作が実際の紙運用に近い形でPC上に実現したソフトウェアである。

ような事ができる仕組みは、必然的にデータベースを中心としたシステムとなり、SE が活躍する場である。そこで、業務改善を目的とし文書管理をベースとするシステム構築ができる中核プロダクトとしてバイнда文書管理システムを開発した。

なお、情報伝達文書もその役目を終えた時点で記録文書に変わることがある。例えばグループ内で共同作業するためにいろいろな赤入れ、書き足し、マーク付けを行った文書は、作業中は考え方を共有するために使われるが、共同作業終了後は、どのような経過で最終的な作業成果に至ったかを示す重要な文書となる。したがって、グループ共有の場から記録文書の書庫に移すことができる事が望ましい。

(2) 紙文書電子化作業の進め方

ペーパーレスが叫ばれ始めた頃は、まったく紙にすることなく文書を運用することが考えられたが、やはり現実的ではなく、現状では紙に印刷したものを再度イメージスキャナなどで読み込んで電子化するということが考えられている。通常、これは関連作業終了後の整理として行われるのだが、この作業はどうしても後ろ向きのため滞りがちになる。そこで、筆者らは紙文書電子化の作業の進め方を次のように考えた。

- a) 作業は一気に進める（極力時間を掛けない様にする）。
- b) 第三者にやってもらう（例えばアルバイトの作業者など）。

c) 分類や整理方法に凝らない（分類や整理方法の検討を省いて、エネルギーを文書登録だけに向けられるようにする）。

これらにより、紙文書を電子化し登録するシステムでは以下のことを基本とした。

① 管理形態は、一般のオフィスと同様に「バイндаの中に文書がある」とした。

これにより、文書の管理単位、名前のつけ方、くくり方など全く現状の紙文書の管理と同じようにできるため、文書内容を知らない人でも登録作業を行うことができる。また、整理方法を考える必要もない。

② イメージスキャナによる読み込み作業と管理情報入力作業を分ける。

紙文書を電子化し登録する際、文書の検索や保存期限管理などに使用するため、文書名や登録日などの管理情報の付加が必要となる。一気に作業を進めるためには、読み込み作業は読み込み作業でまとめて行い、管理情報の入力作業は入力作業でまとめて行えるようにしないとリズムが出ない。特にあってはならないのは、読み込み作業後にコンピュータの内部処理のため次の作業を待たされることである。そこで、図 - 1 に示すように読み込み作業と入力作業を全く別フェーズで行えるようにした。

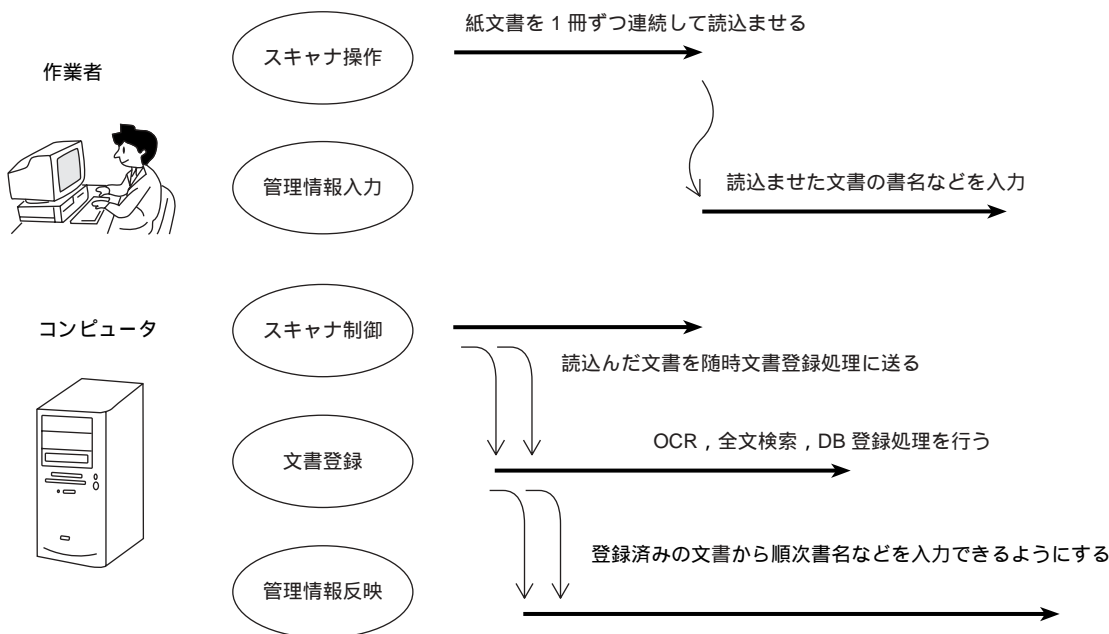


図 1 登録作業フロー図 (Fig.1-Registration operation flowchart)

(3) 紙電子化文書と電子文書との同一管理

もちろん文書は紙から電子化されたものばかりではない。最初から電子的に作成され流通してきたものもある。これらも管理する必要がある。ここで重要なことは「文書を生まれ育ちにより区別しない」事である。具体的には、同じように登録でき、同じように管理情報を持ち、同じように検索ができることである。

この実現方法として、システム内ではどの文書もすべて同じ書式（例えば PDF）にしてしまう方法と、本来の書式をそのままにして検索機能などを複数書式に対応させる方法があるが、バインダ文書管理システムではこの両方ができるようにした。

3 システムの概要

(1) 機能概要と特長

バインダ文書管理システムは表 - 1 に示すように、文書やバインダの登録から検索、廃棄および削除に至るまでの文書のライフサイクル全体を管理する。また、パスワードの更新やファイル格納場所の設定など補助作業もすべて同じ画面上でできるようにした。

これらの機能の中でも以下の点を特に工夫した。

1) 操作の短手数化

- a) ほとんどの項目にデフォルト値を設定し入力量を抑えた。

表 1 機能概要

分類	機能
登録	イメージスキャナによる紙文書の読み込み
	電子文書の登録（1冊ずつ/フォルダ内全て）
	バインダの登録
	文書登録時のメール通知
検索	全文検索（選択バインダ内文書 / 部署内文書 / 全て）
	管理情報検索（登録日、登録者、書名など）
	保存期限切れ文書
	廃棄済み文書
処理	文書の置き換え
	廃棄 / 削除
	エクスポート
設定	部署 / 分類登録
	パスワード管理
	ファイル格納場所指定

- b) イメージスキャナ読み込みで文書の分割やまとめ読みを有効にし、紙のセットを効率良く行えるようにした。

- c) 文字認識の機能とイメージスキャナの機能により、複数サイズ紙混在時の自動認識、縦横混在時の自動回転を可能にした。

2) 全文検索

- a) 文書のタイプ（PDF, text など）によらず同じ全文検索エンジンで検索できるようにした。
- b) 管理情報検索と連携して、対象を絞って検索ができるようにした。

3) シンプルな文書管理体系

- a) 文書の管理体系を、部署 - 分類 - バインダ - 文書の 4 段階に限定した。
- b) バインダ単位でできることを充実し、文書を束ねて扱えるようにした。

4) ライフサイクル管理

- a) 保存期間を設定し、期間が過ぎたバインダをワンタッチで一覧表示できるようにした。
- b) 文書の処分については、検索や表示の対象から外すだけの「廃棄」と物理的に無くす「削除」を用意した。
- c) 他媒体への移動、他システムへの移行のためのエクスポート機能も用意した。

5) 実効的なセキュリティ

- a) 個人認証はせず、部署を認証するようにした。これにより、組織変更にも柔軟に対応できる。
- b) アクセスレベルは二つに限定した。しかも文書単位ではなくバインダ単位とし、アクセスレベルの設定を容易にした。
- c) 他部署へは分類単位で参照のみ許可できるようにした。

(2) プラットフォーム

1) クライアントは Web ベースで導入が容易

システムが稼動する環境は Windows^{注2)}のクライアント/サーバ方式としたが、クライアントは Web ベースとした。これは、参照だけなら既存のパソコンに標準で搭載されているブラウザで行える、価格を抑えるためプログラムモジュール配布を避けたい、との理由による。しかし、イメージスキャナの制御は Web ではできないため VisualBasic (VB) で開発

注2) Windows は、米国 Microsoft Corporation の米国およびその他の国における登録商標である。

した。また、弊社製 A4 版イメージスキャナの ScanSnap に対応するためのモジュールも VB で開発した。

2) Windows の発展性を考慮したサーバアプリ開発環境採用

サーバアプリはマイクロソフトの .NetFrameWork の上に構築した。Windows のバージョンアップによる開発アプリへの影響がないという点を重視したことによる。また、開発においても今後様々なコンポーネントやツールが提供される期待もある。

3) 柔軟なサーバ構成

データベースや OCR などの機能はサーバに置くようにしたが、サーバの構成は柔軟にできるようにした。図 - 2 に示すようにすべての機能を 1 台のコンピュータに集める構成から、各機能をサーバに分散する構成まで可能である。OCR サーバを分離させることで CPU 負荷を軽減することができる。

また、特定の部署だけが大量の紙文書登録をする場合を考慮して OCR サーバを複数設置できるようにした。

4) ネットワーク負荷を配慮したファイルサーバ設定
 ファイルサーバは、文書の実体を格納する場所だが、部署毎に設定できるようにした。これは、WAN を含む大規模なネットワークで運用する場合、データベースサーバは 1 箇所において検索するものの、容量が大きいファイルは手元に置いて WAN を通さないで表示させ、WAN での検索性能維持と、ネットワーク負荷軽減を図るためである。

(3) 他社ソフトやフリーソフトの積極的利用

本システムでは表 - 2 の他社ソフト、フリーソフトを使用している。

これらのソフトを採用した最大の理由は原価低減である。OCR ソフトを除いて全て無償のソフトである。中でも namazu は、国立国会図書館、国土地理院、(独立行政法人) 産業技術総合研究所、(独立行政法人) 理化学研究所、(社団法人) 情報処理学会、各地の大学や自治体および企業で実際に広く使われているオープンソフトで実績は十分ある。また多くの技術者が関わっており品質や将来性についても問題はないと考えた。

OCR 機能は当社でも開発しているが、サーバとして

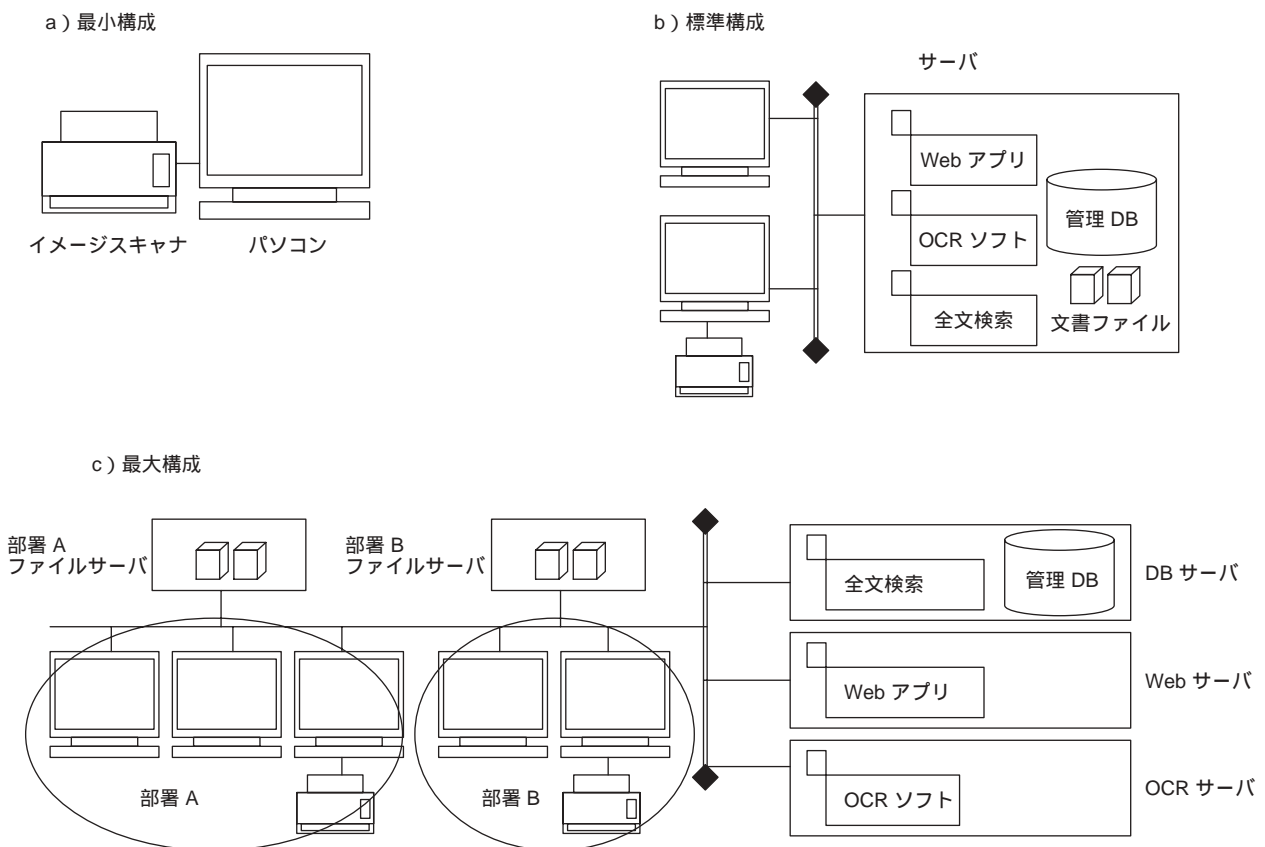


図 2 サーバ構成 (Fig.2-Server configurations)

表 2 利用している他社ソフトとフリーソフト

機能名	ソフト名称
データベースエンジン	Microsoft MSDE 2000
データベースアクセス	Microsoft MDAC
全文検索エンジン	namazu
全文検索用キーワード辞書	kakasi
namazu の API	ActivePerl
namazu の PDF 対応モジュール	Xpdf
OCR ソフト	PScanServe (ハイパーギア社)
	PDF OCR (xelo 社)

動作するソフトを選び採用した。通常の OCR ソフトは名刺管理ソフトに代表されるように個人利用の対話操作を必要とするものが多い。これに対して、本システムでは、読んだイメージファイルを投げつければ勝手に（操作することなく）OCR して PDF ファイルを作ってくれる必要がある。さらに OCR した結果を別のテキストファイルに出力するようでは該当文書を探した後にその文書内の該当部分を探すのに役立たない。これを可能にするためには PDF 文書の読み込みイメージの上に透明テキストを配置するようにしなければならない。

このような要件を満たすソフトとして評価したのが上記 2 社の製品である。

OCR ソフトには常に認識精度向上の課題が付いて回る。本システムではフリーレイアウトの文書を認識する必要があるので更に難しい。いろいろな会社がノウハウを集め日々性能向上に努める分野なので、最適又は最高のものを採用できるよう着脱が容易にできるような組み込み方にした。

当社は伝統的に独自技術を提供することをよしとし、フリーソフトの採用は「障害などに対し責任を果たせない」という理由で拒否してきた。しかし、オープンソースの流行が示すように世の中は安価なものをうまく利用する方向に変わってきている。本システムはこの点では当社の伝統的な考え方にも挑戦したものである。

(4) ScanSnap への対応

ScanSnap はアプリケーションからの制御インタフェースに Twain インタフェースではなく独自インタフェースを採用しており、またアプリケーション開発ツールの SDK も公開されていない。操作はボタンを押すだけなのでアプリ側ではいつ読み込みが行われたのかわからない。そのため、ScanSnap がファイルを保存するフ

ォルダを監視し、ファイルが発生するとこれを OCR ソフトのほうに送るといった処理を行っている。

なお、ScanSnap では 1 枚目が読み込まれた時点でファイルが生成され 2 枚目以降はこのファイルに追加されるようになっている。したがってファイルが発生してすぐに移動すると残り分を入れる場所がなくエラーとなるので、一定時間容量が変化しないことを確認して移動するようにしている。

(5) namazu の性能維持

従来から全文検索ソフトでは「検索は早い登録には時間が掛かる」というケースが多かった。これは検索のための Index 作成に時間が掛かるためである。Index の量が本体の 2 倍というソフトもある。このようなことから namazu の性能維持については考慮を払った。

namazu は非常にシンプルなコマンド体系になっており、index の更新もファイル 1 点 1 点を指定するのではなくフォルダ単位で行うものが用意されているだけである。このような特性を考えて、本システムは以下のようなアーキテクチャとした。

- 1) Index は部署単位に作成する。
- 2) 文書ファイルを置くフォルダの収納数の限界を設定する。

基本的に検索は部署内の文書に行われることが多い。したがってこの時に最高性能を出す必要があるので、部署設定機能で指定された格納場所フォルダの直下に Index フォルダを作成した。

また、文書ファイルは部署設定での格納場所フォルダに直に置かないでサブフォルダを設けその下に置くようにした。このサブフォルダは収納するファイル数がシステム設定で決められた量を超えると新しく作成される。これにより、新しく文書が登録されても Index 変更のために namazu が調べるファイル数は限定されるようになり性能の維持は図れると考えた。

(6) 性能

各処理の実行性能測定を 1 万件～10 万件の文書を対象に 10 段階に行った。結果を表-3 に示す。

一覧表示は全体文書数に依存せず大体 1～2 秒で表示される。管理情報検索は文書数に比例して処理時間が増えていくが、これは検索方法を入力した文字列に、データ中のどこかで一致する文字列を検索する中間一致としているためである。そのためインデックスを作ることができないのでどうしても遅くなる。全文検索の性能はやはり文書数には比例せず 1～5 秒で行われている。

一方、登録については、A4 の原稿 1 枚をイメージ

表 3 検索時間測定値

全体の文書数	文書一覧	管理情報検索 (文書)	全文検索
1 万件	00'92	03'51	01'24
2 万件	01'86	05'58	02'46
3 万件	03'74	07'75	03'44
4 万件	02'25	09'84	03'40
5 万件	01'93	11'77	03'43
6 万件	01'87	14'19	05'39
7 万件	01'81	15'53	04'41
8 万件	02'42	17'10	02'94
9 万件	02'26	19'90	03'68
10 万件	02'36	21'68	05'35

注) 単位は秒, 表示対象件数は 10 件

スキャナで読み込み, OCR 処理, 全文検索用 INDEX 作成, MSDE へのレコード追加までで 10 ~ 20 秒が掛かっている。OCR 処理だけ測定したところ約 3 秒であった。なお, 先のトータル時間には OCR や MSDE 登録で発生する監視時間 (両方とも 1 秒) も含まれる。登録時間については当初予想したより短かったがやはりじっと待ってられる時間ではなかった。非同期処理にして読み込み作業とその後の処理を分けたアーキテクチャが活きた。

4 導入事例

(1) 某社品質管理システム部門

この会社では品質記録の文書を紙で保管してきた。ところが急遽保管場所のスペースを空ける必要に迫られ本システムを導入した。かなりの量の文書の読み込みを行っており, スキャナ読み込み作業が終了して 2 ~ 3 時間経たないと登録が完了しないほどである。

なお, ここでは通常の文字主体の文書の他に写真のデータもある。後者は文字認識して回転を掛ける事ができない (誤認識して余計な回転をしてしまうことがある)。そこで, 各種設定の変更を BATCH ファイル化しモノクロ文書とかカラー写真と名前を付けたアイコンにし

た。これをダブルクリックすることで文書種別に対応して適宜設定を変えられるようにした。

(2) 某社技術情報管理部門

この会社は, 業界情報や設計情報を作業者全体で共有するために本システムを導入した。業界動向については他社動向や競合製品の情報 (紙ベース) をイメージスキャナで読み込み誰もが同時に参照できるようにした。また, 設計情報は製造に使用する材料の情報 (購入先のカatalog) をイメージスキャナから読み込み本システムに登録している。これにより小数部しかないカatalogの取り合いを防止している。

5 今後の展開

現在, 本システムを使用する商談が数多く進められているが, やはり本システムをそのまま OA 的に利用するだけでなく他システムと連携して業務システムの一環とするケースが多い。これを可能にするために「システムとしての入出力インタフェースを整える」事が課題である。これはオペレータが操作することなく大量の文書を登録したり, 大量の文書情報を外部に出力する機能である。これにより他システムで生成したデータと文書を融合してバッチ的に登録したり, また, 他システムに対し文書検索した結果の全データを提供できたりする。楽 2 ライブラリとの連携もこのようなインタフェースを設定して実現することを考えている。

6 むすび

本システムは北陸支店 SE により開発されたものであり, 自分達の得意技の具体化, ビジネス展開の橋頭堡となる事を目的に進めた。その甲斐あって文書管理に関する様々な商談が発生している。顧客の抱える種々の難問に立ち向かいながら技術を磨き, 更に洗練されたソリューションにしていくつもりである。

参考文献

- 1) 日本製紙連合会: 2004 年 紙・板紙内需見通し (2004.1.20). http://www.jpa.gr.jp/rengokai/shiryo/data/naijyu_04.pdf
- 2) 井波ほか: 楽²ライブラリ, *PFU Tech.Rev.*, 15, 1, pp.25-32 (2004).