漢字名寄せ索引システム

当社では，昨年700万人の顧客(契約者，被保険者）を漢字姓名により名寄せした契約情報データ・ベースを完成し，同時に漢字ディスプレイ装置を用いて顧客の契約情報を検索する“漢字名寄せ索引システム”を開発，実施した．

このシステムは，当社が取り組んできた漢字情報処理システムの集大成ともいえるもので，システムの完成により，顧客からの照会などに対し，迅速・正確に対応することが可能となり，いっそうのサービス向上がはかられた．従来，この種の業務は契約1件ごとにカードを作成し，多数の専門職員が手作業で検索していたが，ディスプレイ端末の導入により担当者は顧客の姓名と生年月日を入力するだけで契約情報をディスプレイ画面上に即座に映して見ることができるようになった．

その結果，従来50名ほどいた検索担当の職員が縮減され，また約 500平方メートルを占めていたカード類の保管も不要となり，事務室のスペースの軽減もはかれることとなった．

以下このシステムの開発の背景，システムの概要を紹介し，併せて開発上の諸問題をあげ，多少の考察を加えたい．

1. 漢字システム採用の動機

当社では，従来保険契約の管理を証券番号による契約1件1件の管理方式により行なってきたが，顧客へのサービス向上，販売支援の強化，事務効率化の観点から，今後顧客単位による管理方式（顧客が複数の保険に加入していても，それをひとまとめにして顧客ごとに管理する）へ移行することが強く要請された．そのためには，数百万人の既契約者を名寄せすることが前提となり，つぎの理由から漢字姓名による名寄せがもっともすぐれていると考え，漢字システムの採用に踏みきった．

○当時，姓名住所ファイルはカタカナによるファイルをもっていたが，収録対象が一部の特定の顧客のみであった．

○カタカナ名寄せの場合，生年月日を加えても使用頻度の高い姓名では，名寄せ率は95%程度しか期待できない．したがってサブキイとして，住所等の他の要素を加える必要がある．

○コンピュータ処理で通常使用しているフリガナ名寄せの場合，a同字別読み（例 河野…カワノ，コウノ），b別字同読み（サカイ…坂井，堺，酒井），c表現のバラツキ（大野…オオノ，オウノ，オーノ），d新旧かなづかい等，により名寄せの精度に問題が残る．

○漢字姓名の場合，当社で一番頻度の高い姓名である「鈴木博さん」のケースで約 700人程度なので，生年月日をキイ項目に加えれば，ほぼ完璧な名寄せが期待できる．

2. 漢字名寄せ索引システムの概要

1) 機器構成

機器構成は，丸の内本社に，漢字ディスプレイ装置2式，ディスプレイ表示部英数カナキイボード付26台，同漢字キイボード付6台，新宿の本社別館に漢字ディスプレイ表示部英数カナキイボード付14台，同漢字キイボード付2台をそれぞれ設置した．各ディスプレイ装置は丸の内本社にあるＩＢＭシステム370モデル158と7200ＢＰＳの専用

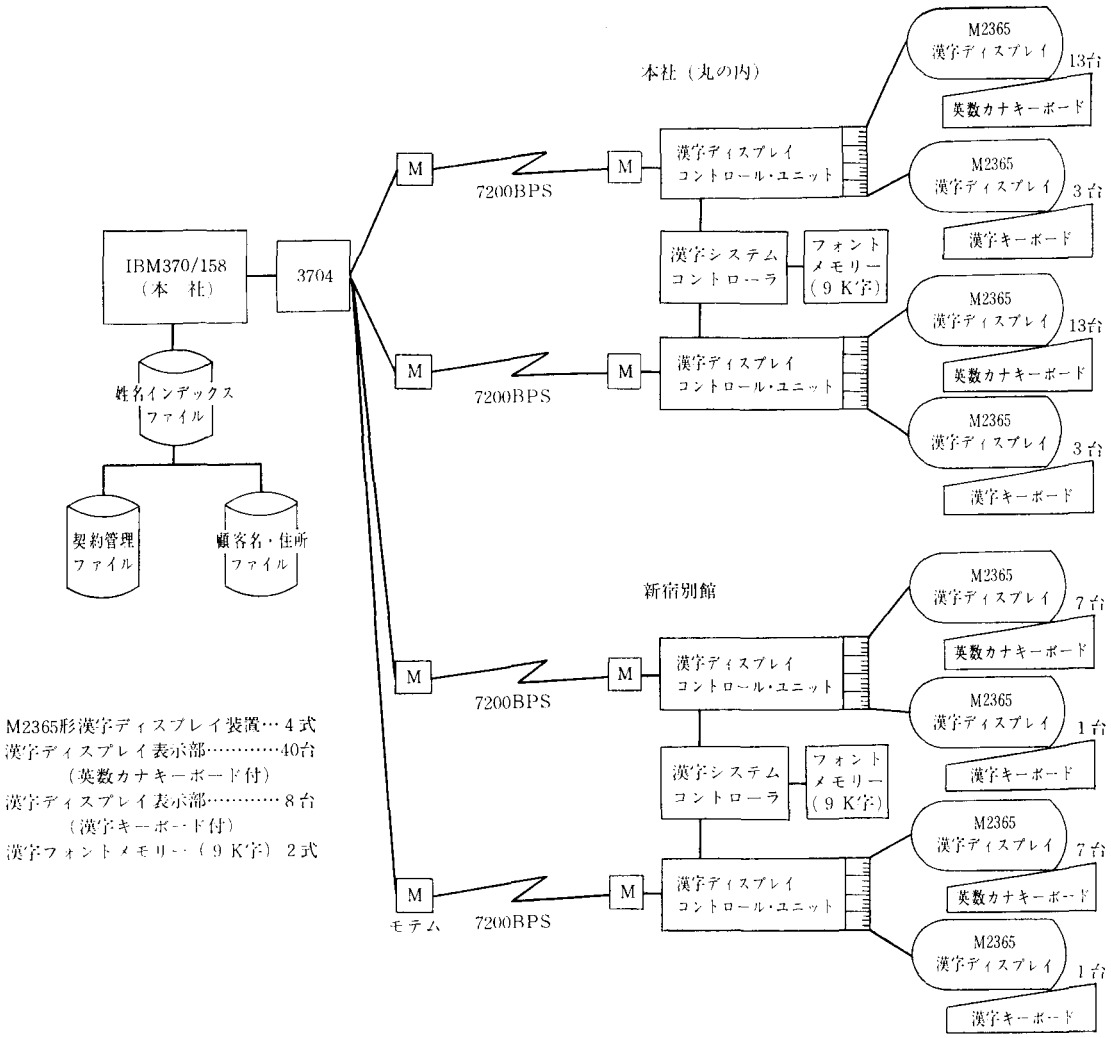


図 1　漢字名寄せ索引システム機器構成図

回線で結ばれている．通信方式は半二重のコンテンション方式で，回線制御はＢＴＡＭ―ＢＳＣを採用した．（図 1 参照）

2）　漢字ディスプレイ装置

今回導入した漢字ディスプレイ装置（三菱電機 M2365）は，３色のカラー表示ができるので，顧客情報等が漢字でかつカラーでディスプレイ画面上に映し出されるので，大変見やすく好評を博している．ただ，一画面の表示文字数が漢字 384文字，ＨＳＰ 768文字のため，英数カナのディスプレイ装置と比べて表示文字数に制約があるのが欠点である．

その他の特長はつぎのとおり．

○画面上に縦書き，横書きが可能．

○画面，行，文字単位の消去，削除，挿入などの編集が可能．

○漢字データの入力を容易にするため画面上に罫線表示，数字エリアの指定が可能．

○操作ミスを守るために，表示画面に保護領域を設定．

○ライトペンによる画面上のデータ修正および漢字入力が簡単．

3）　漢字キイボード

漢字キイボードは，メインボードに2,800文字，

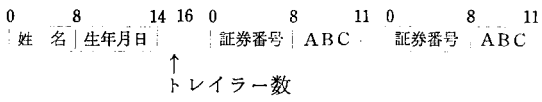
サブボードに1,000文字合計3,800文字を収録することができる．

当社ではメインボードは使用頻度の高い文字を収録することとし，50音順でかつ当社が長年使用してきたフリガナ方式の音訓併用型の文字配列とした．サブボードは使用頻度の低い文字を入れ，その配列は部首別画数順とした．それはメインボード上の漢字は読みやすい字が多いがサブボード上の字は一般職員にはほとんど読めない字が多いので字形から索引できるようにしたものである．当社独自のこの漢字キイボードの採用により，いままで手作業で索引していた職員はすぐこの端末操作に慣れ，一般の職員でも２～３日の訓練で検索できるようになった．（図 2 参照）

4） 名寄せファイル

当社の契約情報データ・ベースは，契約要項等の数字コードを記録した契約管理ファイルと顧客の姓名・住所等を漢字コード化した姓名住所ファイルから成り立っている．今回完成した名寄せファイルは「姓名インデックス・ファイル」と称し，ＤＡＳＤのスペースの節約および検索のスピードアップの観点から，契約１件１件の要項をもった既存の契約管理ファイルおよび姓名，住所ファイルへリンクするためのインデックス機能のみとしている．キイ項目は姓名，生年月日のみで，データとしては契約管理ファイルおよび姓名，住所ファイルのキイである証券番号とその属性をトレイラーとして収録している．顧客が数件保険に加入している場合は，証券番号等のトレイラーを複数もつことになるので，これに対応するためレコードの長さは可変長としている．

ファイル・レイアウト



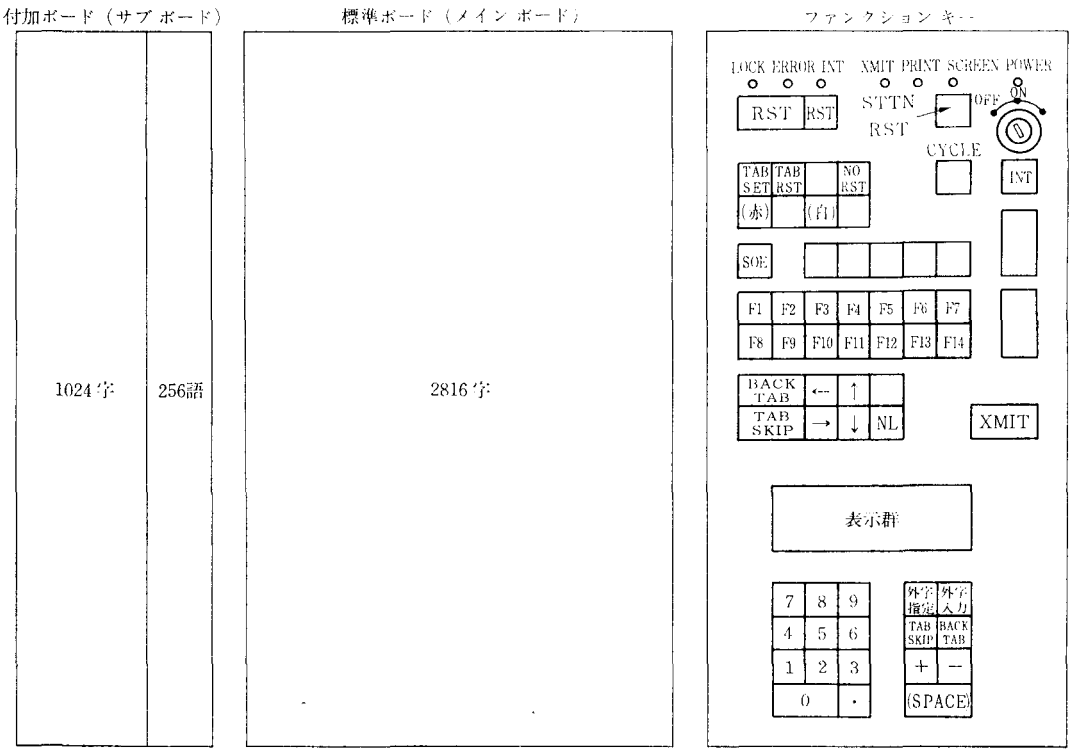


図 2 漢字キイボード盤

姓名…姓と名はつづけて左詰めにしている．
文字は1文字2バイトで表現し，16バイトあるので，ほとんどの姓名がカバーされている．

生年月日…西暦で表示（6バイト）．

トレイラー数…キイ項目につづくトレイラー数を表示（2バイト）．

A…キイ項目の姓名の顧客が被保険者であるか契約者であるかを表示（1バイト）．

B…ファイルの区分を表示（1バイト）．

C…現在のところ未使用（1バイト）．

ファイル編成は，姓名と生年月日のフルキイはもちろんのこと，姓名のみ，姓名＋年，姓名＋年月というような「階層別索引」ができるＶＳＡＭ (VIRTUAL STORAGE ACCESS METHOD) ファイルを採用している．このＶＳＡＭファイルの「階層別索引」機能により，生年月日が不明でも同じ姓名の集団を全部抽出し，その中から該当の顧客情報を検索することができる．

5) コードおよび字種

当社では漢字コード体系としてＥＸＯＫおよびＥＸＥＫコードを採用している．ＥＸＯＫ (EXPANDED ORIGINAL KANJI CODE)は，原始データ作成時における漢字コードであり，ＥＸＥＫ (EXPANDED EDITED KANJI CODE) はＥＸＯＫコードを内部処理に適したコードに変換したものである．このＥＸＥＫコードをフォント・メモリーに収録して使用しており，現在の収録字種および字数はつぎのとおりである．

漢字	明朝体	7,646字
	ゴシック体	508字
記号		428字
ひらがな		180字
カタカナ		178字
数字		30字
英字		156字
計		9,126字

3. 開発上の諸問題

1) 姓名レングス

漢字姓名名寄せのインデックスファイルを作成するにあたり，契約情報ファイルで使用している姓名のレングス（32バイト）では検索のパフォーマンスに問題があるので，姓名インデックスファイルで使用する姓名レングスの短縮を検討した．そこで既契約 630万件の漢字姓名の文字数を調査し，姓名レングスを決めることとした．調査の結果姓名合わせて最大8文字で漢字姓名が表示できることが判明し，姓名レングスを16バイトに決定した．（表1参照）

2) 漢字キイボードの字種選定

ディスプレイ端末の漢字キイボードの文字配列は，メーカー指定の音順による文字配列であるので，人名，地名を主体に使用する当社の要求と合致しなかった．そこで前述のとおり独自の文字配

表 1 姓名の長さ調べ (6,302,864件)

姓	件数	全体比	累計比
1 字	220,452	3.50%	3.50%
2 字	5,822,727	92.38	95.88
3 字	257,827	4.09	99.97
4 字	1,356	0.02	99.99
5 字以上	502	0.01	100

名	件数	全体比	累計比
1 字	998,212	15.84%	15.84%
2 字	4,579,027	72.65	88.49
3 字	718,725	11.40	99.89
4 字	6,263	0.10	99.99
5 字以上	637	0.01	100

姓と名の結合	件数	全体比	累計比
2 字	27,926	0.44%	0.44%
3 字	1,087,335	17.25	17.69
4 字	4,304,263	68.29	85.98
5 字	849,469	13.48	99.46
6 字	32,605	0.52	99.98
7 字	846	0.01	99.99
8 字以上	420	0.01	100

表 2　当社既契約における使用文字の頻度ランキング

1.	田	2,459,807	11.	本	820,685
2.	子	2,260,305	12.	木	788,306
3.	藤	1,240,843	13.	村	784,783
4.	一	1,235,618	14.	井	783,007
5.	山	1,150,606	15.	中	771,853
6.	雄	944,512	16.	男	729,530
7.	野	901,219	17.	正	707,633
8.	夫	877,466	18.	小	679,599
9.	川	868,689	19.	三	678,294
10.	郎	823,816	20.	美	667,766

（数字は 4,800万字中の頻度）

列によるキイボードをつくることとした．文字は使用頻度の高い文字からキイボードに入れることとし，姓名住所ファイルで使用している漢字に関し，その使用頻度を調査した．（**表 2**，**図 3** 参照）

3）　名寄せ精度の向上

漢字には異体字というものがあり，意味や発音が同じでも字形が異なるものがある．本字，古字，別体字，俗字などとよばれる，字源が一つでも時間の経過や環境の変化で長い間に別の字形が形成されたものである．ＥＸＥＫコード体系では，文字の字形パターンに対してコードが付与されているので同意同義の文字であっても字形が異なれば別字の扱いとなる．たとえば「会沢広」という人を仮定しよう．異体字として「会」と「會」，「沢」と「澤」，「広」と「廣」があり，これらの文字を書き分けると８通りの組合せができ，漢字コード上は８人の別人となってしまう．この問題を解決するために「異体字の相互変換テーブル」を作成し，この変換テーブルを通すことにより「会」「沢」「広」はそれぞれ「會」「澤」「廣」と同一の文字であるとみなして処理をすることにより，完全な名寄せをすることができた．検索する場合も同様に変換テーブルを通して行なうので索引洩れが起こる心配はなくなった．つぎに類似文字の問題がある．つまり字形が類似しているための書き誤りや読み違いにどう対処するかということである．たとえば「未」と「末」，「千」と「干」，

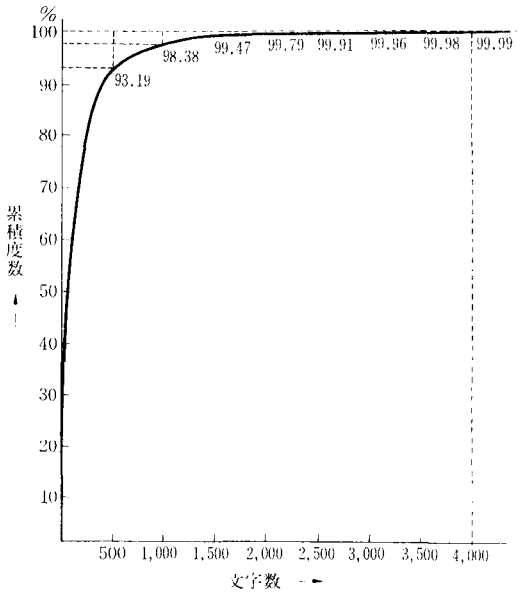


図 3　姓名について使用文字の頻度順文字数の累積度数分布

「右」と「石」などは筆の勢いや用紙の汚れでエラーの発生につながる．これらは字形は似ていてもまったくの別字であるし，類似の客観的範疇の設定が困難であることから変換テーブルには入れていないが，名寄せ精度確保のためにぜひとも解決すべき問題である．

おわりに

本システムの完成により，顧客からの照会に対する応答が迅速になり，顧客サービスは大幅に向上した．また名寄せファイルの活用により新規契約の事務の機械化，セールスマンの販売実績算定処理の機械化等，各種の名寄せ関係の機械化が促進できた．加えて顧客単位の管理が可能となったことにより，顧客動向の把握や市場情報の収集分析など，マーケッティング面への効果も大きい．今後はこのシステムを全国の支社に順次拡げてゆき，顧客サービスの充実と事務の効率化をいっそう推進していきたいと考えている．

おがわ・こういちろう　　1946年生
明治生命保険相互会社　システム部