NDP Marketing News

「ビッグデータ」時代の「リトルデータ」活用法を顧客と共に考える

2014年9月号（毎月10日発行）

発行㈱日本ダイレクトプロモーション
東京都中央区日本橋茅場町1-10-8
本社TEL03-5651-0611 Fax03-5651-0600
大阪TEL06-6222-6511 Fax06-6222-6512

「ペイオフ」に見る「名寄せ」と「データ・クレンジング」の関係
実際に業務に携わった技術者が、具体例を織り込んで書いた解説に学ぶ

2013年5月、行政手続で個人を識別するための番号の利用等に関する法律（マイナンバー法）が国会で成立し、2016年（平成28年）1月から番号の利用が開始される運びとなっている
個人がひとつの番号しか持たなければ、なにかの折に「名寄せ」をしなくてすむようになる
この「名寄せ」という言葉が一般に知られたのは2005年に解禁した銀行のペイオフ対応だった

ペイオフの名寄せ作業の現場から

金融機関は自信を持っていたが、実際にやると「名寄せ」という行為はとてつもない難物だった

2005年に解禁した銀行のペイオフ対応では，金融庁は金融機関の破綻に備え，金融機関ごとに口座を持っているお客様の資産を特定するために，同一金融機関で同一顧客の口座情報をまとめて，資産状況を報告することを義務化した。

当初各金融機関からは，十分な準備ができているとの報告があったが，氏名・住所・電話番号をキーにした集計が思った以上に難航した。
それは，情報が微妙な表記の差異を含んだまま登録されているケースが多数存在することが原因だった。要するに、コンピューターの処理には使えないデータだったのだ。

データク・レンジングとはデータを“ごしごし”こすってきれいにする行為

では，使えないデータを使えるデータにするにはどうすれば良いか。その作業は「データ・クレンジング（Data Cleansing）」と呼ばれる。
つまり，データを“ごしごし”こすってきれいなデータにするということ。
例えば，「03(9999)9999」「0399999999」といった電話番号のデータを「03-9999-9999」の形式に補正することもデータ・クレンジング。
まとめると，データ・クレンジングとは，システムが想定している正しいデータに修正すること。

データ・クレンジングをした結果データ間の関連性を引き出す行為が「名寄せ」

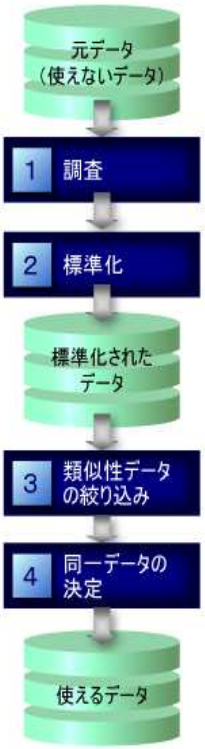
ところで，データ・クレンジングに似た言葉として「名寄せ」というものがある。
名寄せはデータ・クレンジングをした結果，データ間の関連性を導き出す行為。
金融機関のペイオフ対応の例で言うと，同一顧客を導き出す行為が名寄せとなる。
重複データを特定するという観点では，名寄せもデータ・クレンジングの一環。しかし，データ・クレンジングなしに名寄せは実現できない。

◆名寄せ技術1「調査」

まず，どの項目をキーとして判断していくかを決める必要がある。一般的に「生年月日」は空白データが多数あることと、偶然の一致例も多く，キーとして有効ではない。次に「電話番号」は，全く同じ番号であれば同一人物である確率は高い。

◆名寄せ技術2「標準化」

標準化は，データのばらつきを解消するために各データを標準的なデータに変換する作業。



各データを他のデータと突き合わせることができるような形式に変換することがポイント。
例えば次の例では「源義経」と「源義経（全角空）」「源 義経（半角空）」を統一する必要がある。

◆名寄せ技術3「類似データの絞り込み」

標準化が済むと，標準化されたデータを突き合わせる作業を行う。名寄せにおける突き合わせとは，同一データかどうかの比較のこと。完全に一致するデータは問題はないが，ある程度一致するデータを同一データととらえるかどうかが，名寄せの場合のポイントとなる（完全一致を試みるのであれば，名寄せという考え方は必要なく，一般的な付き合せ処理となる）。

◆名寄せ技術4「同一データの決定」

同一データの決定では，パターン化や数値化されて絞り込まれたデータが，どの程度一致すれば同一人物と判断するかを決める作業を行う。どのデータも値によってその傾向は異なるので，どのパターンならOKかという明確な基準はない。

※本誌1・2・4頁はhttp://itpro.nikkeibp.co.jp/article/COLUMN/20070613/274697/を参照させて頂いた。

完璧が求められた「ペイオフ対応」は「名寄せ」の試練だった
時間との戦いの中で、最後はデータ・クレンジングのプロの出番となった

データベース・マーケティングには、データ・クレンジングが不可欠との認識は
今日では、相当数の顧客データを有する企業では当たり前となっている
それには「ペイオフ」の教訓と、次頁で取り上げる年金問題が大きく影響した

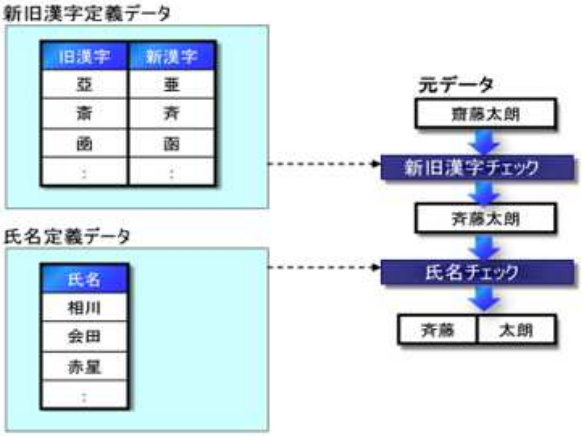
日本語の多様さが、「名寄せ」を一層難しくしている
●氏名の標準化（これらは架空のデータであり，実際のデータではありません）

住所に比べると、氏名データの標準化はもう少し単純で，スペース記号や改行文字を除去した後，姓，名の切り分けをすれば良い。
しかし，調査の段階で「新漢字，旧漢字を混在して用いているものがある」というばらつきを見つけた。

この場合は，新旧漢字の対応表を用いて，どちらかの文字に統一しておくのが良いだろう。
そして，氏名一覧というものを用意しておけば，右図のような処理で姓と名の切り分けができる。

しかし，それでも完全な標準化が行えるわけではない。
「南総一郎」という氏名があったとすると，「南　総一郎」とするか「南総　一郎」とするかは完全には判断できないからだ。
このようなデータは完全に切り分けが行えないため，例外的に保留するということも考慮しておくと良いかもしれない。
この例外処理も入れておけば，図の右側ような結果を得ることができるだろう。

以上のように，データ形式のばらつきを吸収するように，データを切り分けて細かいフィールドに分けていく作業が標準化。標準化を終えたら，標準化されたデータを突き合わせて，同一データを決定する。

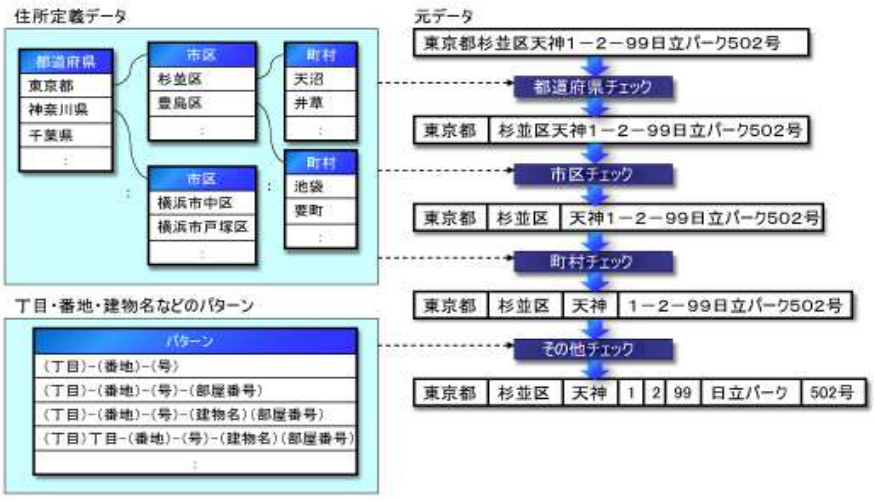


路線検索時に類似駅名で混乱するように、地名の判定も難しい
●住所データの標準化（これらは架空のデータであり，実際のデータではありません）

ここで，住所定義を使わなくても，単純にXX県，YY市の文字列を切りだせば良いのでは？と考える人がいるかもしれない。
しかし，例えば“市”という文字は，「市川市」や「四日市市」などのように，YYの中にも存在する可能性があるので，単純な切り出しはできない。
市区名をあらかじめ定義しておく必要がある。次に，丁目，番地，建物名などのチェックを行う。

この定義は，都道府県，市区，町村などの定義ファイルを用意するのではなく，表記パターンを用意しておくのが良いだろう。
ハイフン（－）で区切るパターンや，XX丁目YY番地といった記述をするパターンがある

ので，それぞれに基づいたチェックを行えば切り分けができるだろう。
このようにデータを切り分けて，標準フィールドに値を埋めていくのが，住所データの標準化だ。



日本保険機構のサイトから 年金の「名寄せ」プロセスを学ぶ

未統合の記録(5000万件)の 基本名寄せ(第1次)の方法

年金記録のコンピュータ上での突合せ(名寄せ)及びその結果記録が結び付く可能性がある方への「ねんきん特別便」の送付

(1)年金記録のコンピュータ上での突合せ(名寄せ)

○「5，000万件」の未統合記録と約1億人(受給者約3，000万人・加入者約7，000万人)の方の基礎年金番号で管理されている記録とのコンピュータ上での突合せを行い、その結果記録が結び付く可能性がある方々1，030万人に対して、3月末までに「ねんきん特別便」を送付しました。

○コンピュータ上の突合せは、氏名、性別及び生年月日の3つの情報等を用いて1次・2次にわたって実施しました。

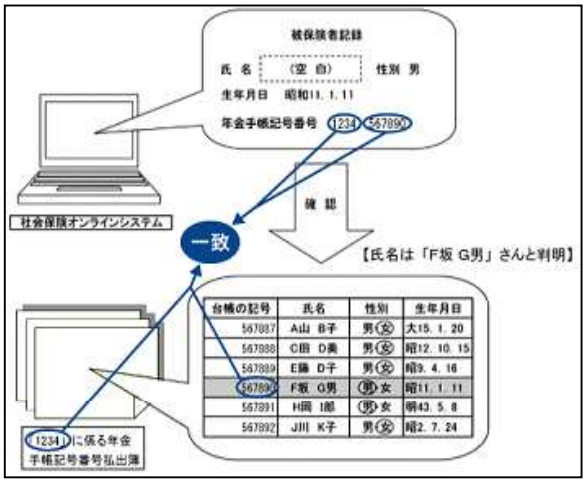
第2次名寄せの方法

○なお、名寄せに先立って、氏名等が収録されていない記録の調査を行ったところ、5，000万件の年金記録の中にそのような記録が約524万件あることがわかりました。

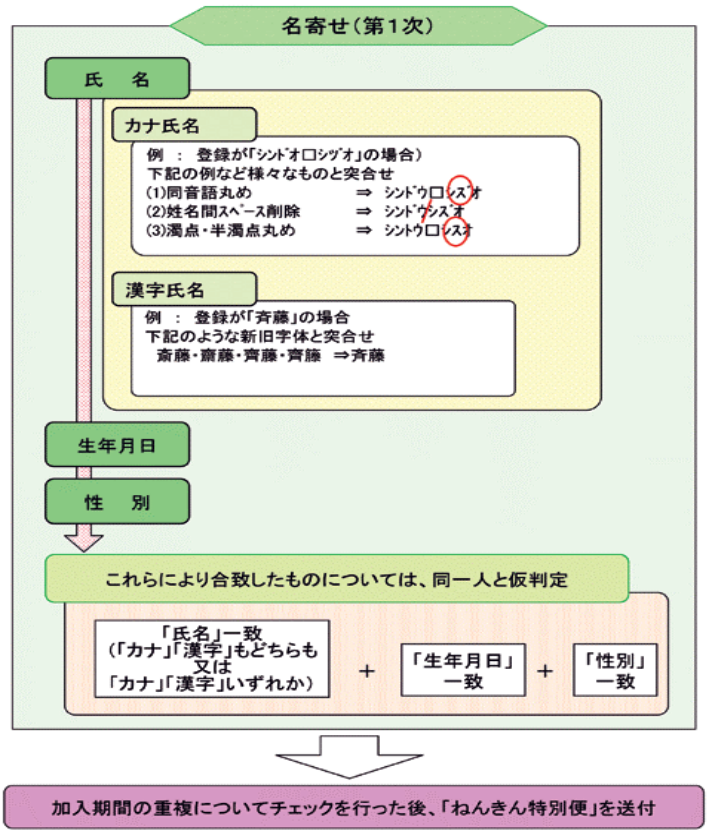
こうした記録については、サンプル調査の結果を踏まえ、平成19年9月7日から平成20年1月10日にかけて、記録に記載されている年金手帳記号番号を手掛りに、年金手帳記号番号払出簿等を参照して補正作業を実施した結果、約99%に当たる記録の補正が完了しました。

補正のためさらに調査を継続すべき記録については、現在、被保険者名簿等を基に丹念に調査を実施しています。

2013年9月現在　解明された記録2983万件、解明できなかった記録2112万件

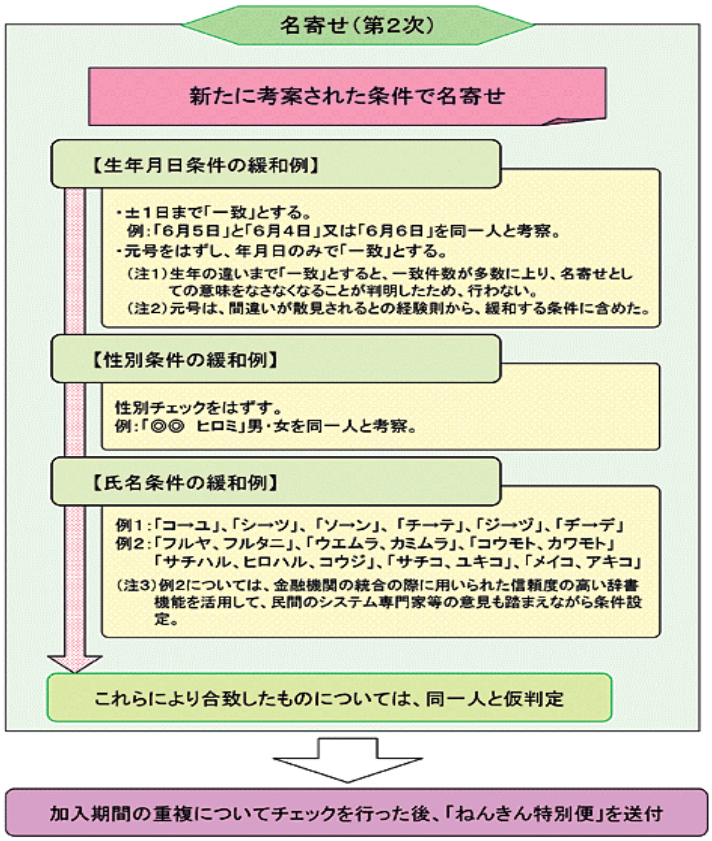


基礎年金番号に未統合の記録(5000万件)の基本名寄せの方法



加入期間の重複についてチェックを行った後、「ねんきん特別便」を送付

基礎年金番号に未統合の記録(5000万件)の第2次名寄せの方法



加入期間の重複についてチェックを行った後、「ねんきん特別便」を送付

■日本保険機構　http://www.nenkin.go.jp/n/www/service/detail.jsp?id=986　参照

データ・クレンジング調査　BtoB顧客データ整備のポイント

企業の顧客リスト(BtoB)は個人顧客リスト(BtoC)よりはるかに項目数が多い

一般に企業、とくにキーマンや窓口担当者の名簿管理は、複雑であり
精緻に設計されたシステムが必要とされています。なぜなら
企業名簿管理には左下囲みのような変更項目が目白押しだからです

企業名簿管理上、欠かせない変更事項とは

●担当者の人事異動による変更

●キーマンの人事異動による変更

●担当部門の組織変更

●部門名称の変更

●関連部門の追加と担当部門との役割変更

●会社所在地変更による一斉変更

●企業統合、所在地の統合による変更

●顧客企業のITインフラ変更による連絡先メールアドレスの変更

●自社の担当者と部門の変更

難易度が高い、顧客データベースの整備

実際の顧客の実情は、左記が複雑に絡み合い、最低でも半年度単位で変化していきます。
基本となる社名については、登録名、正式名、通り名、略称、商標トレードマーク(trademark)、サービスマーク(service mark)などがあります。

具体例としては、JR東日本、JRE、東日本旅客鉄道株式会社、Viewカード、あるいはNTTdocomo、NTTドコモ、株式会社エヌ・ティ・ティ・ドコモ、そして旧社名であるエヌ・ティ・ティ移動通信網株式会社など。さらに株式会社と㈱の表記違いもあります。
住所においてもカナ住所、旧住所。電話番号では
(　)くくりの有無、ハイフン入り、桁数間違い、国際番号などが入り乱れています。

なお、企業の顧客データ管理には、「経理システム」「代理店管理システム」「Web会員システム」「コールセンターシステム」「会員カードシステム」など多数の顧客データ管理システムがそれぞれに稼働しており、連携がむずかしい場合もあります。

したがって、営業支援システム導入とBtoBプロモーション実施の手始めには、顧客名簿データのデータ/クレンジングや名寄せが必要となってきます。

NDP『展示会来場者【顧客化】マニュアル』より

名寄せ技術1「調査」(本誌1頁)の続編　BtoCのポイント

◆名寄せ技術1「調査」(1頁記述内容)
まず，どの項目をキーとして判断していくかを決める必要がある。一般的に「生年月日」は空白データが多数あることと、偶然の一致例も多く，キーとして有効ではない。次に「電話番号」は，全く同じ番号であれば同一人物である確率は高い。

しかし，同一人物でも複数の電話番号を使っている場合は全く別の番号になる。少数だが空白データも見られるため，優先順位は低くしたほうが良いかもしれない。「住所」と「氏名」については，空白データはないが入力内容にばらつきが見られた。まず氏名データには次のようなばらつきがあった。
姓と名の間のスペース記号が様々
新漢字，旧漢字が混在している

また，住所データには主に次のような表記のばらつきがあった。
都道府県を省略して記述しているものがある
建屋表記(マンション，ビル名称)の有無が混在している
大字(おおあざ)の表記を省略しているものがある
丁目/番地の記述の有無が混在している
英数字が全角/半角の記述方法が混在している
カタカナ表記と漢字表記が混在している

このばらつきが解消できるようならキーとして有効となりそう。このように，調査の段階では，有効そうなデータ項目とデータ値を見て，キーとする項目を決めていく。データ値については，後の作業となる標準化への適性？があるかどうかを見ていく必要がある。

データの傾向が見られるサンプル・データをうまく抽出し，どの項目をキーとしてその後の名寄せをすれば良いかを決定していくことが調査の段階(サンプル・データの抽出方法も，調査としては大切な作業。データの傾向をつかむためには，いかに精度よくサンプル・データを抽出するかが重要だからです)。

こうした調査結果からペイオフでは，「住所」と「氏名」を名寄せのキー項目として決定することにした。

出典はhttp://itpro.nikkeibp.co.jp/article/COLUMN/20070613/274697/

大企業での活用例が多いデータ・クレンジング・ソリューション
データベース・マーケティグの成功確率を高めるためには導入が望ましいのだが・・・

何といっても高額なのがネックか

データ・クレンジング・ソリューションは、ある程度の汎用化はされているものの、クレンジングの対象となるデータベース自体の構造がさまざまであることから、実際の運用においては一定のカスタマイズが必要となる。
したがって、導入時には自社データベースについての十分な理解をもった人材が不可欠であるが、運用がスタートしてしまえば、クレンジング条件の変更などを行わない限り、ある程度自動的に処理が行えるので、特別な知見やスキルをもった人材は不要である。

導入費用については、最低コストが数百万円規模となっているソリューションが多い。
対象となるデータベースのサイズなどによりコストが変動するので、自社のマーケティング施策全体におけるデータベース・マーケティングの重要度などを鑑みながら、マーケティング予算のうち、どの程度を投入できるかを十分に検証し、身の丈にあったソリューションを選択することが肝要であろう。

また、データ・クレンジングをサービスしている事業者も存在するので、データベースの活用度合いによっては、例えば1年に1回など定期的にこのようなサービスを利用して、データベースの精度維持を図るという手法も有効であろう。

現在、日本国内で提供されている主なデータ・クレンジング・ソリューションとしては、
㈱アグレックスの「トリリアム」、インフォマティカ・ジャパン㈱の「Informatica Data Quality」、日本IBM㈱の「IBM InfoSphere Quality Stage」、ビツニーボウズ・ソフトウェア㈱の「Spectrum(スペクトラム)」などが挙げられる。

金融機関中心に150社導入実績あるシステムも

「ビジネスプロセスアウトソーシング」「ソフトウェアソリューション」「システムインテグレーション」の3本柱で事業を展開するアグレックスが、1999年4月から提供している「トリリアム」は、金融機関を中心に150社以上の導入実績を誇るデータ・クレンジング・ソリューションだ(右欄で詳細紹介)。

マーケティング業務におけるデータ・マッチングをはじめ、法改正対応のデータ整備業務、与信管理業務、企業合併時の顧客統合業務など、さまざまな業務で使用可能であり、データ・クレンジングや名寄せに関する専任技術・営業体制を有しているため、きめ細かいフォローアップを期待できる。

受託処理サービスも提供しているので、その利用により処理精度を確認した上での本格導入も可能だ。

代表的なデータ・クレンジング・ソリューション

提供企業	(株)アグレックス
名称	トリリアム
費用の目安	初期費用:2000万円～ 年間保守費用:310万円～ 受託処理:100万円～
特徴	・国内NO.1の実績(金融機関を中心に150社以上の導入実績) ・さまざまな業務システムで使用可能(統合顧客管理システム、マーケティング業務におけるデータマッチング、法改正対応のデータ整備業務、与信管理業務、企業合併時の顧客統合業務など) ・データ・クレンジング／名寄せに関する専任技術、営業体制を保有 ・高精度なクレンジングと柔軟な名寄せ機能 ①(アグレックス保有の住所・姓名、企業名キーワード辞書を組み込み ②さまざまな項目(20項目)を組み合わせた最大1,000パターンの名寄せ設定が可能 ・明確なロジックに基づくクレンジング／名寄せを提供
提供企業からのコメント	近年の成長企業は、顧客情報の保有件数増加とともに顧客情報周辺の管理・運営方法を随時見直し、顧客との間で質の高いコミュニケーションを行うことを目指しています。 保有している顧客情報は最大の情報資産であり、この活用は企業の命題といえます。 トリリアムは、顧客情報を最大活用するために必要なクレンジング／名寄せ機能を、パッケージソフトウェアと日本における14年間の実績・ノウハウを合わせて提供するものです。
問合せ先	営業統括部　ソリューション営業部 www.agrex.co.jp sys_eigyou@agrex.co.jp 03-6831-8111

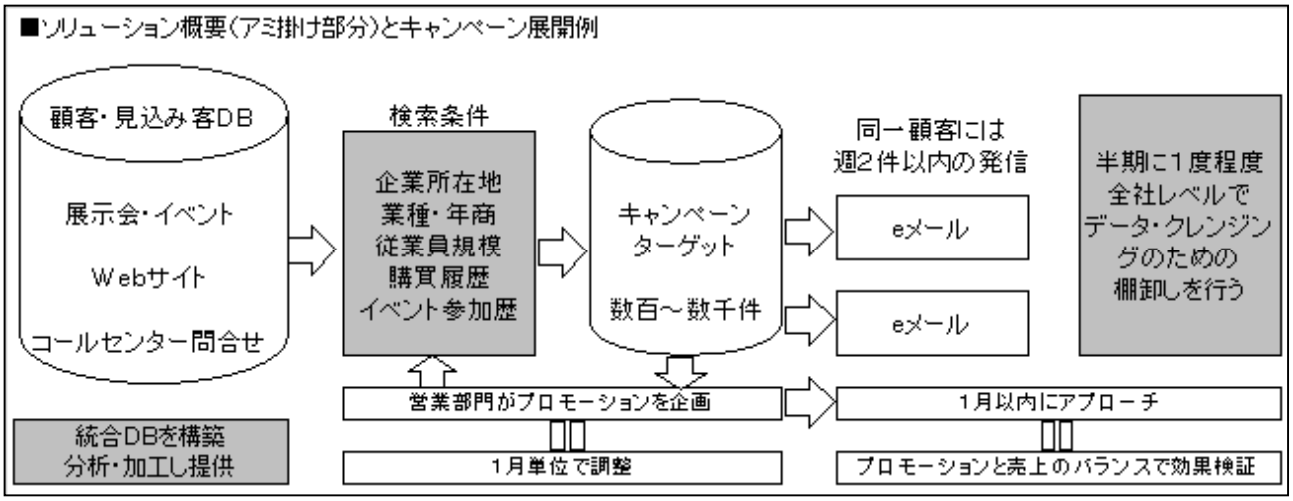
■本資料の出典は『I.M.Press』(2013年8月号P.50～53)

データ・クレンジング・ソリューション 【ライバル比較】
日本IBMと日本マイクロソフトのソリューションはどう違うか

提供企業	日本アイ・ビー・エム㈱
名称	IBM InfoSphere Quality Stage
費用の目安	最小構成の標準価格：400万円～
特徴	・データ・クレンジングに必要な本格的ETL*機能（Data Stage）を包含 ・名寄せなどのマッチングにおいて同一性の確度を数値で算出 ・作成したクレンジング、名寄せモデルをWebサービス化する追加機能を提供 ・大容量のデータを処理する並行処理を追加機能として提供 ・グローバルおよび日本国内での豊富な導入実績 ・分析調査、標準化、マッチング、およびサバイバーシップという、データ品質管理に欠かせない4つのプロセスを網羅 *ETL：Extract Transform Load

提供企業からのコメント	ビッグデータの流れもあり、多様な情報をいかにマーケティング活動に生かすかは、企業戦略の大きなテーマです。 特にCRMの観点では、新規情報と既存情報との整合性を保った連携が必須です。「IBM InfoSphere Quality Stage」は高度なデータ・クレンジング、名寄せ機能を提供するソフトウェア製品です。 コールセンターなどOne to oneマーケティングでのリアルタイムのビジネス・プロセスを支援するWebサービスとの連携も実現します。
問合せ先	ソフトウェア事業　インフォメーション・マネジメント事業部 http：//ww.ibm.com/software/jp/data/infosphere/qualitystage/ 0120-550210

管理＆運用	㈱日本マイクロソフトの顧客データベースの管理手法	効果的運用のポイント
管理システム	**自社ソリューションを使用**	
DB構築目的	マーケティング・キャンペーンのPDCAサイクルが回せるプラットフォームとしての顧客データベースの整備 現場のニーズにわせて個別にデータベースを構築するのではなく、顧客データは基本的にグローバルで統一されたフォームのデータベース上で管理し、データの活用を行いやすいかたちに加工できるようにする	一般的に営業用のデータベースの構築に当たっては現場の使い勝手を優先しがちだが、一定のルールに則って顧客データの集約化を図り、その活用段階で、各現場の意向を最大限に満たしていくという方法で、全社で使える顧客データベースを構築した
データ・クレンジング	プロモーション用に抽出した顧客リストは基本的に1カ月以内にアプローチを行う。プロモーション実施時に判明した要修正点（異動や退職など）は、できる限りその都度修正を行う	半期に1度程度、全社レベルでデータ・クレンジングのための“棚卸し”を行う



※参考資料：㈱アイ・エム・プレス『BtoBマーケティング成功事例集』及び、NDP『展示会来場者【顧客化】マニュアル』

展示会来場者情報(名刺)をメールアドレスで名寄せ
「アルプス・システム・インテグレーション㈱」のeメールマーケティング事例

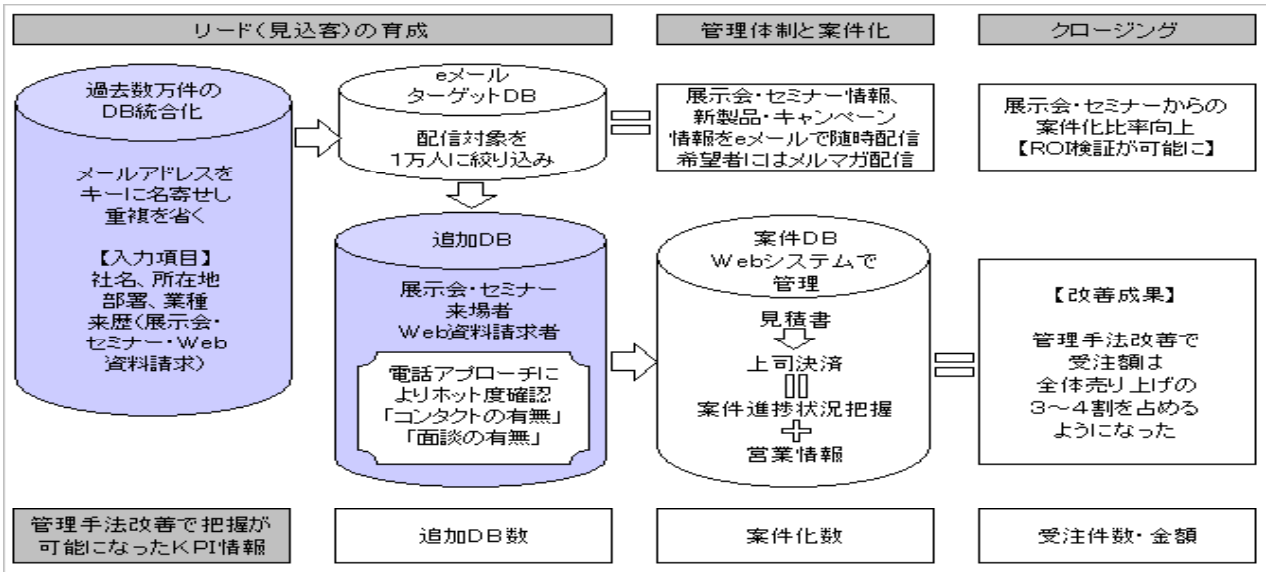
貴重な見込客情報が“使い捨て”状態だった

同社では、以前から上半期と下半期にそれぞれ1回を目処に大規模な展示会出展やセミナー開催を行い、ソリューションのPRやプロモーションを実施してきた。
ところが、展示会ごとに集まった名刺のデータは表計算ソフトに入力し、営業担当者に渡す程度にとどまっていた。しかも、入力項目や順番などはバラバラで、展示会・セミナーごとに異なるといった状況であった。展示会・セミナーの規模や人気度合い、集客状況にもよるが、多い場合では1回で3,000枚近く集まった名刺が“使い捨て”に近い状態となっていたのである。

当然、これらの他にWebサイトを通じた問合せや資料請求も相当数あったが、その扱いも名刺と同じレベルだった。同社が本格的なデータ整備の取り掛かった時点では、対象となる見込み客データ件数は数万件に及んでいた。
これをeメールでアプローチすることを前提に、まずメールアドレスをキーとした名寄せを実施した。重複を省くかたちで社名、所在地、部署、業種、来歴(どの展示会・セミナーへの来場か、Webサイト経由の資料請求者か)などの項目を備えたデータベースを整備したのである。
これらの経緯は下の表に。また、eメールマーケティングの推進システムは下段のフローの通り。

展示会出展 セミナー開催	上半期・下半期に各1回、大規模な展示会に出展orセミナーを開催 多いときで3,000名近くの名刺が集まることも		
管理手法比較	従来の管理手法	改善後の管理手法	改善効果
DB内容	名刺データを入力 展示会・セミナーごとに入力項目、順番が異なった ボリュームは数万件規模	eメールアプローチを前提に、メールアドレスをキーに名寄せし重複を省く 入力項目は「社名」「所在地」「部署」「業種」「来歴(展示会・セミナー・Web請求資料)」	随時なeメールマーケティングが可能となる
DB活用	営業担当者に配布	展示会・セミナー情報、新製品・キャンペーン情報をeメールで随時配信 ※配信対象は1万件に絞込み 希望者には更にテーマ別ファイリング情報をメールマガジンにて配信	eメールマーケティングによりターゲットの絞込みが可能に
課題	見込み客DBが使い捨て状態	DBセグメントにより、ターゲット別の効果的なアプローチ手法の確立	

■eメールマーケティングと名寄せ中心のデータ・クレンジング概要(網掛け部分)



※参考資料:㈱アイ・エム・プレス『BtoBマーケティング成功事例集』及び、NDP『展示会来場者【顧客化】マニュアル』

電機関係4社(日立・三菱・東芝・シャープ)　3年での組織・系統変化
50%変わった日立、20%の東芝、変らない2%の三菱、跡形もないシャープ

企業組織は短期間にこんなにも変わってしまう

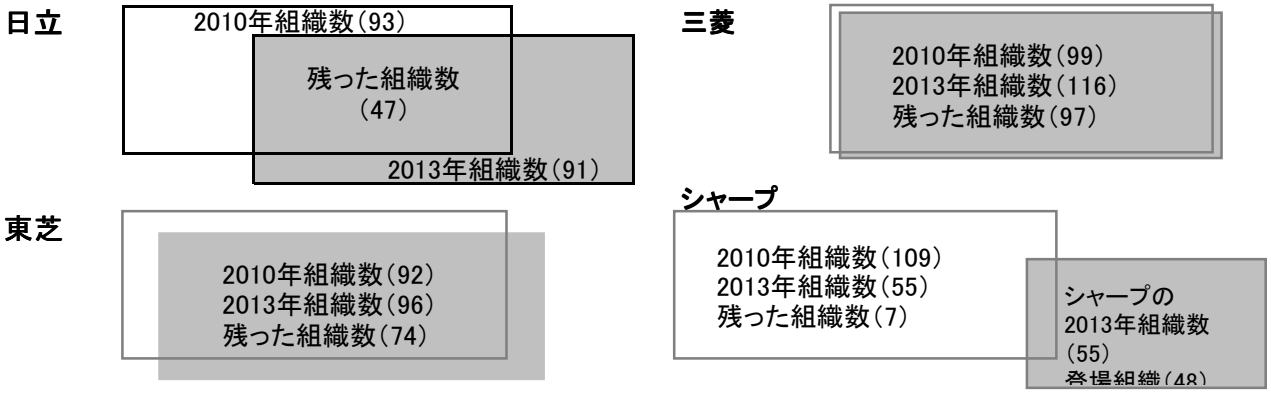
ダイヤモンド社が発行していた『組織図・系統図便覧(全上場会社版)』は2011年版で廃刊となった。
これを引き継いだのがネット上の検索サービスである「ダイヤモンドD-VISION NET」。
この2つの資料を基に、電機大手で組織図・系統図を体系だって掲載している4社に絞り、この3年間(2010年と2013年の比較)の組織(部門名も含む)の名称の変化(○○室が○○部に変更も変化とカウント)を調査した。

なお、2013年データは4～7月の調査であることが判明しているが、2010年については、本誌に調査月日のないものがあり、期間はおおよそ3年間となることをお断りしておく。
結果は右上表の通り。
ただし、組織の変動状況を調べることを目的としたため、名称の変わらない支社・事業所・工場ならびに厚生施設(病院など)はカウントから除外した。

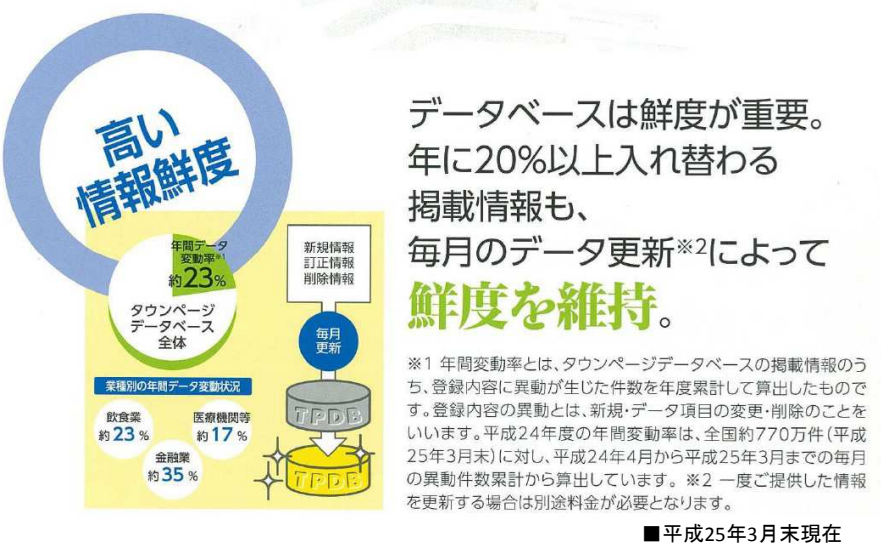
	日立	東芝	三菱	ｼｬｰﾌﾟ
2010年時点	93	92	99	109
2013年に消えた組織	46	18	2	102
退場率	49%	20%	2%	94%
残った組織	47	74	97	7
2013年に登場した組織	44	22	19	48
2013年時点	91	96	116	55
登場率	48%	23%	15%	87%

経営問題でゆれたシャープは、ドラスティックな組織改変があったことが一目瞭然。
近年業績回復の目覚しい日立は、ダイナミックな展開が読み取れる。
東芝の20%は理解の範囲だが、三菱の2%には驚かされる。
なお、三菱で新登場が19あるが、うちの13は各事業分野にコンプライアンス部が創設されたもの。

■電機4社の2010～2013年組織変化イメージ



■『NTTタウンページ データベース資料』から、事業所データの1年での変化率を見る



年23%の変動率の意味

この数値は電話番号のほか、商号、住所変更(同一局番内の移転であれば、電話番号は変わらない)を含んだものでの変動率で、事業所の転廃率ではない。
なお、忘れてはならないのは、この数値が全国平均であるということ。
つまり、都市部では変動率はさらに高くなることが予測される。
それを、金融業の約35%という数値が示唆しているように思わる。