解説 3

IBM InfoSphere™ Identity Insight Solutions
ー脅威・不正に対抗するインテリジェント・ソリューションー

顧客データ、従業員（人事）データ、取引先データなど、企業の情報システムが「人」に関する情報を取り扱う場面は数多く存在します。特定のシステムの目的が何であれ、人のデータを取り扱うソリューションの基本的な機能として、誰に関する情報を取り扱っているのか、対象の人物を正確に特定可能であることが求められます。従来のソリューションでは、人物の特定はデータベースやテキスト処理システムの検索エンジンを用いたあいまい検索や、いわゆる「名寄せ」と呼ばれる同一人物に関する２個のレコードを１個に統合する処理を用いて実現されてきました。本稿では、ブラックリスト照合を主な題材として、従来の方法の問題点を指摘し、IBM InfoSphere Identity Insight と IBM InfoSphere Global Name Recognition による新しいソリューションが、人物特定に対して高い精度を実現し、従来の方法では知り得なかった自明でない人物間の関係を発見できることを紹介します。

Article 3

IBM InfoSphere Identity Insight Solutions
- Intelligent Solution for Fighting Threat and Fraud -

Businesses today implement many types of solutions that manage information about "people," including customers, employees, and suppliers. No matter what the specific objective is, the primary goal of these solutions is to understand who exactly you are dealing with. A conventional solution typically uses the search engine of a general-purpose database system or a data cleansing software that finds and merges duplicated records about the same people. In this article, considering a watch list filtering as the target application domain, we firstly review problems with the conventional solutions, and then discuss how a new solution using IBM InfoSphere Identity Insight and IBM InfoSphere Global Name Recognition brings a new level of accuracy to the concept of identity recognition and discovers non-obvious relationships that conventional technologies cannot conceive of.

❶ はじめに

情報システムが取り扱うデータの中には、「人」に関する情報がたくさん存在します。企業が持つ顧客情報、従業員情報、取引先情報などはその代表例であり、役所が扱うデータの大半は、人（市民）に関するデータであるといえます。情報システムにはそれぞれ異なる目的がありますが、人を取り扱うシステムの共通特性として、人物を正確に特定することが求められています。このことはシステムの目的として当たり前のことではありますが、技術的にすべて解決済みかというと、そうではないというのが現状です。

例えば、社会問題となっている年金記録問題［1］は、被年金保険者の特定がうまくできていないために生じていると考えられます。犯罪やテロの温床となる資金洗浄（money laundering）の防止や、反社会勢力との取引を排除することは喫緊の社会的要請となっており、さまざまな対策［2］［3］［4］が進められている一方、振り込め詐欺のような身元を偽ることによる犯罪は後を絶ちません。これらを防ぐためには確実な人物特定が必要となります。一般に善良な市民が行う本人確認では、本人であることがうまく特定できるよう協力的な情報が提示されると期待できます。しかしブラックリストに載っている人物が、明らかにそれと特定されるような情報を提示するとは到底考えられず、そうした悪意のある相手に裏をかかれないような仕組みが必要となります。また、ブラックリストに載っている本人情報はもとより、そのような人物と何らかの関係が疑われる人物には十分な注意が必要です。つまり、性善説に基づく単純な一致・不一致の二者択一だけでは、人の情報を扱うシステムとして不十分なのです。

ブラックリストとの照合に加えて、人に関する不完全で断片的な情報から身元を特定し、人と人の関係を発見することが重要なビジネス領域には、以下のようなものが考えられます。

- **金融犯罪対策**：すでに述べたブラックリスト照合とは別の対策として、トランザクション・モニタリングと呼ばれる取引の頻度、量、パターンなどの監視が導入され始めています。しかし、本来このトランザクション・モニタリングが監視する対象は、個々の口座ではなく、取引を行う主体である人や組織であるべきです。それを実現するためには身元の特定や、人と人の関係を正しく認識する必要があります。
- **高付加価値の顧客関係管理**：一般に高い利益を生む顧客層と、あまり利益につながらない顧客層が存在します。しかし、一見管理コストが高く利益の上がらないように見える顧客でも、優良顧客の親せきや友人であれば安易に取引を打ち切るべきではないでしょう。高利益が期待できる顧客と深い関連を持つ顧客は、やはり高い利益をもたらす可能性が高いからです。そのような関連を事前に見いだすことができれば効率のよい営業活動が可能です。
- **従業員選考**：カジノのディーラー（従業員）とハイローラー（顧客）や、購買担当者と取引先などの間に密接な関係があると不適切な取引を行うリスクが高まります。企業が巻き込まれる犯罪の多くは、内部犯行であるともいわれています。このため、企業の利害関係者と従業員や採用応募者との間の関係を調べておくことには大きな意味があります。

本稿では、人に関する高度な情報処理を行うことを目的として、2章で従来の「名寄せ」処理の問題点を論じます。また3章でInfoSphere Identity Insight Solutionsが従来方式の課題を解決することを説明し、4章ではInfoSphere Global Name Recognitionによる名前に特化した分析・検索技術について説明します。

❷ 名寄せプロセスとその課題

「名寄せ」とは、金融機関などに同一顧客が複数の口座を保有することになった場合に、人定情報（人物を特定するための情報）と呼ばれる「氏名」、「住所」、「生年月日」などの顧客属性を用いて口座所有者の同一性を確認し、一致する場合には同一顧客の複数口座として一元管理する手続きのことを指します。InfoSphere QualityStage™ [5] [8] はデータの分析、標準化、品質調整、確率的マッチング、データの取捨選択といった名寄せに必要なデータ・クレンジングのプロセスを開発・実行するためのソフトウェア製品であり、国内外で幅広い実績を有しています。

名寄せでは、情報システム内の人物を表すレコードを保持し、それらの一致・不一致によって人物を特定します。QualityStageを用いた標準的な名寄せのプロセスを**図1**に示します.

「調査・分析」では、元データに含まれるパターンや単語の出現頻度を分析し、その後行う「標準化」や「マッチング」に必要なルールや重み付けを決定します。

「標準化」では、次の「マッチング」を効果的に行うため、データの意味を考慮して一貫した形式に変換を行います。例えば次の2つの住所は同じ場所を表していますが、文字列としては大きく異なるため、このままではシステムには一致すると判断できません。

・東京都港区A町3丁目2番12号Xビル205号
・港区A町3-2-12-205

そこで住所の意味を認識した上で、表記揺れがなくなるようにデータを標準化します。電話番号や生年月日、そのほかの属性についても同様です。

「マッチング」では、標準化されたデータ項目を比較して、同一顧客を表している可能性があるレコードの組を作ります。その組ごとに一致の度合いを類似度スコアとして計算し、その値の大小によって名寄せを行うか否かの判断を行います（**図2**）。グレーゾーンでは、自動的な判断をあきらめて、人間の判断を仰ぐようなプロセスが採用されることもあります。

「サバイバーシップ」では、「マッチング」で同一であ

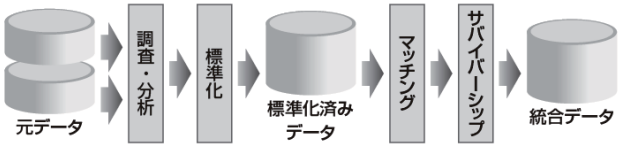

図1. 名寄せプロセス

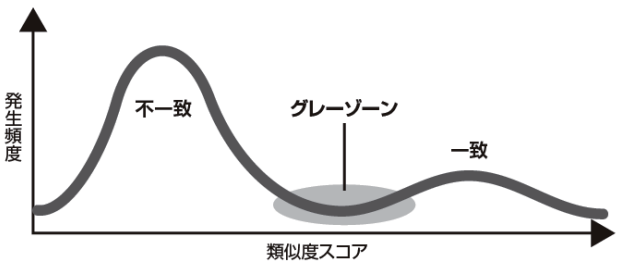

図2. スコアの分布

ると判断された複数のレコードを１つに統合します。このとき、一致しないデータ項目が存在する場合には、統合したレコードに何を残すか（survive させるか）を決定します。

ブラックリスト照合と名寄せは、人定情報を使ったマッチングを行うという点でよく似ています。しかしブラックリスト照合を目的とした場合の名寄せには、次のような問題点があります。

- **最新情報だけに頼る危険性**：口座所有者の属性情報を管理するという観点では、最新の正しい情報だけを記録しておけばよいのですが、人物を特定するという観点では、人定情報はどれも可変であるため、最新情報だけの使用で個人が特定できるとは限りません。住所や電話番号はもとより、氏名ですら結婚などで変化します。現在の最新情報だけではなく、過去の蓄積情報も活用する必要があります。
- **誤りの蓄積**：名寄せによる一致・不一致が常に 100% 正しく判断できるとは限りません。手元にあるデータに誤りが含まれていることもあります。いったん１つのレコードにデータが統合されてしまうと、後に入力されたデータが正しい情報であるにもかかわらず、システムの判断と矛盾が生じ、自動的に誤りを訂正することができません。
- **１対１比較の限界**：名寄せでは２レコード間の類似性を比較し、類似度スコアとして評価します。銀行口座のようにデータ項目の均一性が高い場合には大きな問題になりにくいのですが、対面営業、電話、Web など、複数の情報源から集められた異質なデータを組み合わせて人物を特定する必要がある場合は、３個以上のレコードを総合的に評価し、初めて同一人物であると判断できることがあります。**図３**の例では、どの２レコードを取っても同一人物を表しているとは言い切れませんが、４レコードを総合すると同一人物を表している可能性が高いと考えられます。名寄せやデータベース検索では、1 対 1 の比較を行うため、このような断片的なデータから１人の人物を特定することはできません。
- **一致・不一致の二者択一の問題**：顧客情報の統合という観点では、一致するか・一致しないかの２通りの結論しかありませんが、ブラックリスト照合では、それで十分とはいえません。例えば、テロリスト A と同じ住所に住む人物 B と、B に送金を行っている人物 C についての情報が分かっているとして、そこに C と同じ電話連絡先を持つ人物 D が新規顧客として現

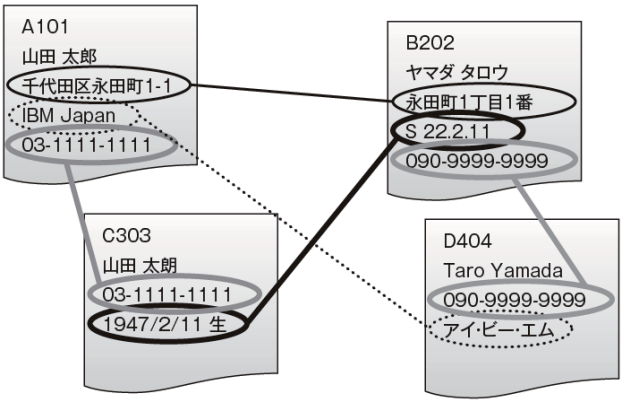

図3. 1対1比較の限界

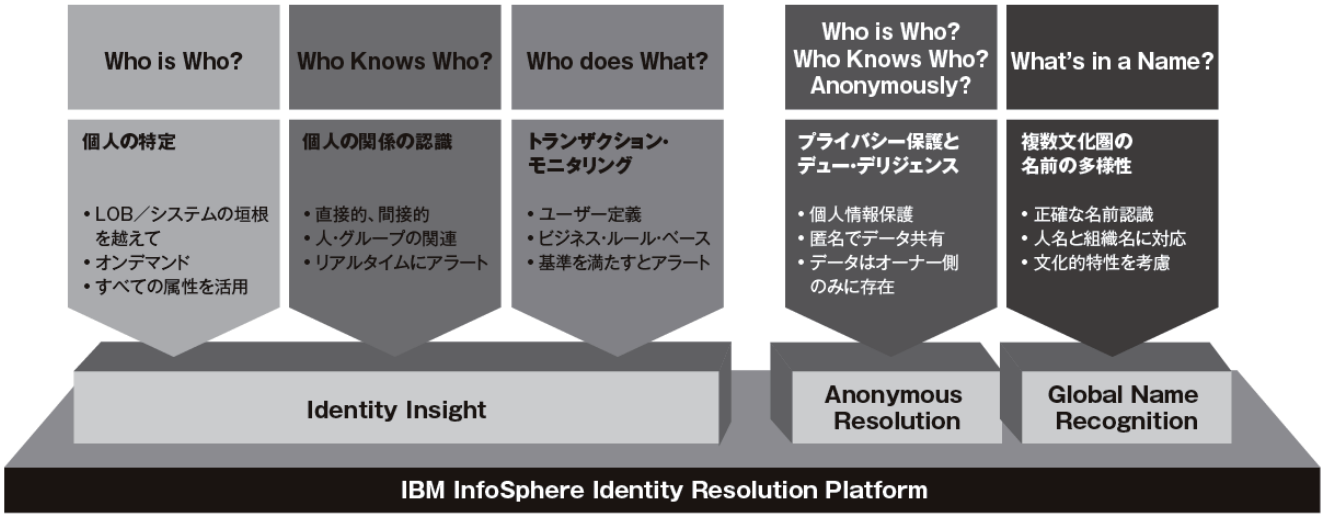

図4. IBM InfoSphere Identity Resolution Platform

れたとします。人物 D はブラックリストに直接記載されている A とは明らかに異なる人物ですが、注意を払う必要があると考えられます（**図 5**）。名寄せや検索による一致・不一致の二者択一ではこのようなリスクに対処することができません。

- **リアルタイム性の欠如**：名寄せは大量のデータを扱うため、しばしば夜間や週末に実行するバッチ処理として運用されます。しかし、送金や口座開設はリアルタイムに行われており、バッチ処理では事後の検査になってしまうため、問題の行為を阻止することができません。

ここまで列記した問題の多くは、データベースやテキスト検索エンジンを用いた（あいまい）検索によるソリューションにもそのまま当てはまります。

❸ InfoSphere Identity Insight

InfoSphere Identity Resolution Platform（**図 4**）は、Identity Insight、Anonymous Resolution、Global Name Recognition の 3 製品からなる、身元解決のためのプラットフォームです［6］。本章では人の同一性（Who is Who）、人と人の関係（Who Knows Who）、人の行動（Who does What）を認識し、そこからビジネス上の価値や危険性を発見してアクションを取ることを可能にする InfoSphere Identity Insight（以下、Identity Insight）を紹介します。

この製品は次のような特徴を持っており、2 章で述べた問題を解決します。

- **コンテキストの蓄積**：Identity Insight はデータが更新されても過去の情報を切り捨てずに蓄積し、過去から現在（もし利用可能なら未来）の情報を使用して人物を特定するためのコンテキストを生成します。
- **誤り訂正**：蓄積されたコンテキストに基づいて総合的に判断を行います。仮に以前に行った判断が間違っていたとしても、新しい情報が追加された段階で自動的に誤りの訂正が可能です。
- **関連性の発見**：Identity Insight は同一人物を認識するだけではなく、別人であっても共通の属性を持つような人物を関係者として認識することができます。
- **複数レコードのマッチング**：蓄積されたコンテキストの中で、3 個以上のレコードを使って初めて分かるような同一性・関連性を発見することができます。
- **ルールに基づくアクション**：新規顧客がブラックリストに登録されている人物と同一である場合はもちろん、関連があるとされる場合にも注意が必要です。しかし、通常の顧客の間に関連が発見されても特に問題はありません。事前にルールを定義しておくことにより、興味のある発見があったときにだけアラートが発せられます。
- **リアルタイム**：データが投入されると即座に処理が行われるので、直ちに必要なアクションを取ることができます。

Identity Insight の仕組みを**図 6** に図示します。Identity Insight は、1 つまたは複数の情報源から人定情報の供給を受け、Entity データベースに解決された身元情報を蓄積するとともに、注意を払うべき事象に対するアラートをリアルタイムに出力します。

「**標準化**」は名寄せの場合と同様に、データの意味を考慮して入力されたデータを一貫した形式に変換します。

「**データ補強**」では、入力データを精査するために外部データの参照を行います。例えば入力された住所データからその緯度・経度を求めることで、住所の見か

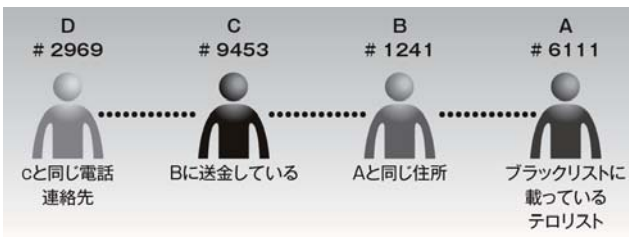

図5. 関連の連鎖

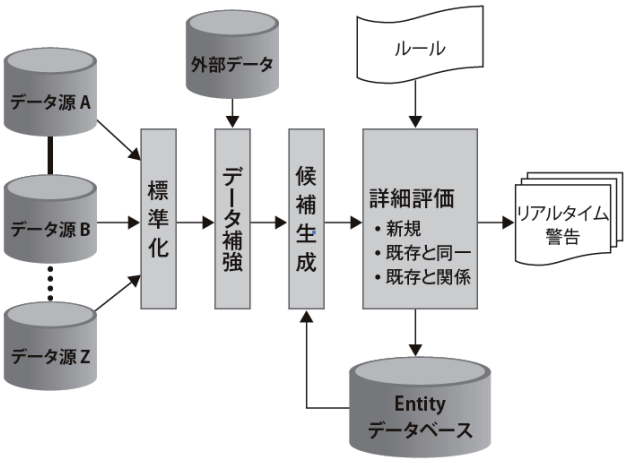

図6. Identity Insightによる身元解決プロセス

けの類似度だけではなく、地理的な近さを評価できるようになります。また漢字の名前に対して読み仮名を振ることで、カナやローマ字のデータとのマッチングが可能になります。

「候補生成」では、Entity データベースに蓄積されたコンテキストを参照して、新しく入力された人定情報と同一または関連の可能性がある人物の情報を候補として列挙し、次の「詳細評価」に送ります。

「詳細評価」では、候補をルールに照らして 1 つずつ評価し、入力された人定情報が新規の人物を表しているのか、既存の人物と同一なのか、さらにそれが既存の人物と関係があるのか判断します。その結果は Entity データベースに格納され、その結果が事前に定義されたルールに抵触するとアラートが出力されます。

Identity Insight の動作例

図 7 に示す例を用いて Identity Insight の動作を見ていきましょう。初めに、①既存顧客リストと②ブラックリストが存在したとします。現時点ではそれらの間には何の関連もありません。そこへ③新規顧客が追加されると Identity Insight は、まず③新規顧客の住所が①既存顧客のものと一致することを発見します（図中の太線 a）。しかし一般には、顧客同士が密接に関連していてもなんら問題はなく、これ自体は特に注目に値しません。続いて、③新規顧客の電話番号が②ブラックリストのものと一致することを発見します（図中の点線 b）。新規顧客とブラックリストに載っている人物は別人であっても、同じ電話番号を持つような関係があるなら注意を払う必要があるのでアラートが出力されます。さらに、①既存顧客が③新規顧客を通じて②ブラックリストと関係していることに気が付きます（図中の実線 c）。

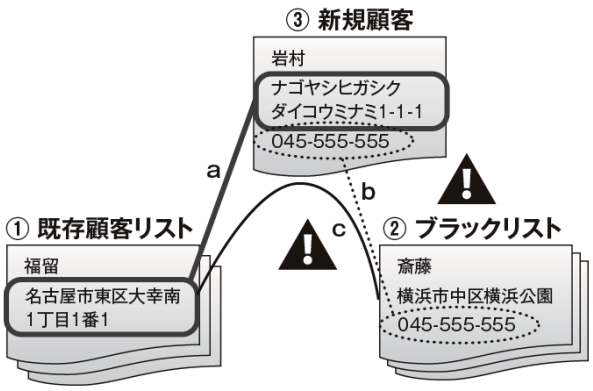

図7. Relationship Resolutionの例

これも注意を払うべき事象として、直ちにアラートが発せられます。

2010 年 3 月現在、Identity Insight は製品自体としては日本語データを直接サポートしていませんが、筆者が所属する大和ソフトウェア開発研究所では Identity Insight で日本語データを扱うための拡張モジュールを開発しており、ソリューション・コアとしてお客様に提供しています。

Anonymous Resolution

図 4 の Identity Resolution Platform の構成要素である Anonymous Resolution は、複数の組織の間で相手の組織に自組織が持つ個人情報を明かさずに、本章で述べた Identity Insight の機能を実現する製品です。本稿では詳しく述べませんが、人定情報はすべて一方向ハッシュ関数を用いて匿名化します。このため他組織の持つ身分情報を具体的に知ることはできませんが、自組織の持つ身元情報との一致や関連の有無を知ることはできます。

❹ InfoSphere Global Name Recognition

人定情報の中でも「名前」は最も重要なデータ項目です。グローバルなビジネスを展開する際には、世界中の人名を取り扱う必要がありますが、名前は単なる短い文字列ではなく、それがよって立つ文化的背景ごとに異なる構造や意味を有しており、それらを正しく認識しなければ、正しく処理することができません。国際的な人名の比較・照合に関する主な課題を見ていきましょう。

通常において、一般の単語には正しい表記法が定められていますが、人名にはどの記述が正しいかを論じる基準がありません。例えば、“bear”（熊）と “bare”（裸の）の発音は同じですが、一般構文において、どちらを使えばよいか迷うことはありません。もし取り違えていたら単純な間違いであるといえます。しかし “Smith” と “Smythe” は共に名前として間違っているとはいえず、どちらを用いるべきかを知るためには、本人の身分証を確認するしかありません。

パスポートや国際的なクレジットカードの名前表記には、ラテン・アルファベットが使用されています。しかし日本人の名前は漢字、ひらがな、カタカナで、ロシア語の名

前はキリル文字で、ギリシャ語はギリシャ文字、またペルシャ語ではアラビア文字で、それぞれ記述されていますので、それらの文字からラテン・アルファベットへの変換が必要となります。このように文字種を変換することを翻字（transliteration）、あるいは転写（transcription）と呼びます。翻字にはどうしても揺らぎやノイズが生じます。例えば「小野」と「大野」はどちらも “Ono” と書くことができ、区別がつかなくなります。“Чайковский”（チャイコフスキー）は “Tchaikovsky” と書かれることが多いようですが “Chaykovsky” とつづるかもしれません。“محمد”（ムハンマド）は “Muhammad”、“Mohamed”、“Mehmet” など多くの種類の翻字が使われています。

近代の日本人名は姓と名の 2 つの部分からなりますが、海外では地域によって名前の仕組みや考え方が根本的に異なります。ミドルネームを用いることがあるのはよく知られていますし、 現サウジアラビア国王の “Abdullah bin Abdulaziz”（アブドゥッラー・ビン・アブドゥルアズィーズ）の “bin Abdulaziz” は Abdulaziz の子孫を意味するように、先祖の名前を使用する文化圏もあります。ニックネーム、イニシャル、略称を用いることもあります。

図 8 に名前を構文解析した例を示します。このように名前には構造があり、その構造を認識することによって初めて、名前のどの部分が重要でどの部分が補助的で省略可能なのかが理解できます。

以上の課題はすべて、名前が所属する文化を理解しなければ解決できません。**図 4** の Identity Resolution Platform の構成要素である InfoSphere Global Name Recognition（GNR）は、25 年間以上をかけて約 10 億人の人名を研究した成果に基づいて開発されたソフトウェア製品で、この問題を解決するために次のような機能を持っています［7］。

- **名前の分析**：入力された名前が所属する文化圏を推定するとともに、性別、相対頻度が高い国を知ることができます。例えば “Takeshi Fukuda” を入力すると、文化圏は日本、性別は男、国は日本、ブラジル、ペルーなどという答えが得られます。“Naomi Campbell” を入力すると、文化圏はアングロ（英語）、女性、国は日本、バハマ、イギリスなどが返ってきます。このように “Naomi” は日本の名前であると同時に、英語文化圏の名前でもあります。名前を構文解析して**図 8** のような構成要素に分解することができます。さらに、人名、組織名なのかを推定することもできます。また名前のバリアント（類似表記）を生成することもできます。例えば “Robert” のバリアントとして、そのニックネームである “Bob” を認識できます。どのようなバリアントがあり得るかも、名前の文化圏に依存します。
- **2 つの名前の類似度スコアリング**：名前の分析で得られた名前の文化圏に基づいて、2 つの名前の類似度スコアを計算します。例えばアラビア語圏では Qadafi と Kadafi の違いは大きくありませんが、中国では Quan と Kuan は別人です。このように、文化圏によってどのような違いが重要で、どのような違いが重要でないかが異なるため、ここでも名前の文化圏を正しく認識することが重要です。
- **名前の類似検索**：入力された名前と類似度の高い名前を大量のリストから見つけ出し、類似度スコアの高い順に返します。

日本語人名処理の課題

これまで海外の名前を中心に課題を述べてきましたが、日本人名を扱う際にも日本文化固有の課題があります。

- **漢字・読みの多様性**：「さいとう」という姓には、斉藤、斎藤、齊藤、齋藤、西藤、西塔、犀東、柴桃…など多数の漢字表記がありますが、どれが正しいかは身分証明書を確認しなければ分かりません。逆に、

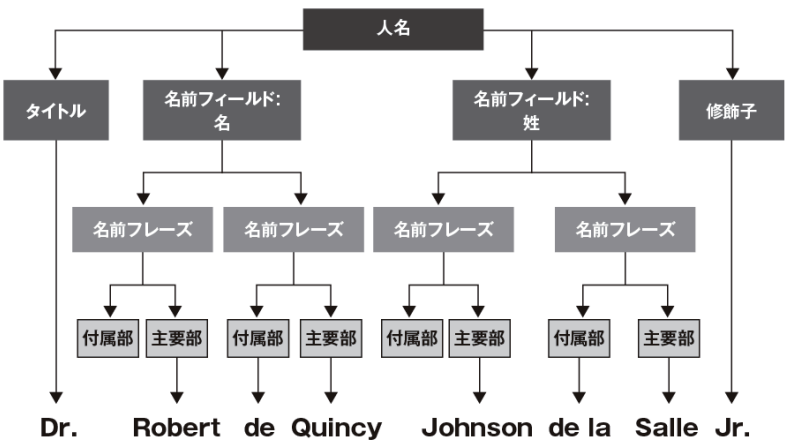

図8. 名前の構文木（例）

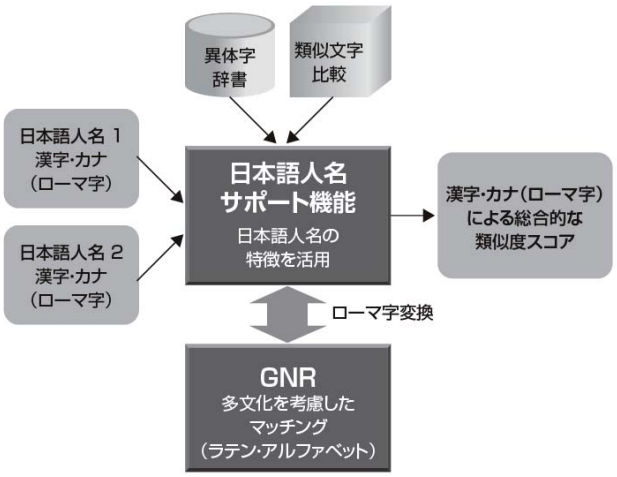

図9. GNR用　日本語拡張モジュール

「河野」は「こうの」とも読めますし「かわの」とも読めます。日本人名の比較は漢字だけ仮名だけではなく、その組み合わせや相互の変換が必要になります。

- **異体字**：身分証明書上の表記では「齊藤」が正しいとしても、簡略のために普段は「斉藤」と書いているかもしれません。この例の「斉」と「齊」は異体字と呼ばれ、同じ意味と発音を持つので交換して使用されることがあります。そのような文字は日本語で用いられるものだけで 1,200 組 2,700 文字以上も存在します。
- **類似文字**：「萩（はぎ）」と「荻（おぎ）」は意味も読みもまったく異なる文字ですが、見かけが似ているので取り違えられてしまうことがよくあります。
- **誤入力などで生じる誤り**：日本語の入力方法から生じる誤りとして、かな・漢字変換での変換誤り、光学文字認識（Optical Character Recognition; OCR）の誤認識などが考えられます。
- **姓・名の区切り**：日本語は、単語をスペースなどの区切り文字で分かち書きしない言語です。このため、姓と名の区切りが明確に与えられない場合があります。

2010 年 3 月現在、GNR はアルファベットで書かれた日本人名はサポートしていますが、漢字、ひらがな、カタカナで書かれた名前を直接サポートしていません。著者が所属する大和ソフトウェア開発研究所では、日本語人名の姓名の分割、漢字名の読み仮名付与、およびローマ字変換、異体字や類似文字を考慮した人名の類似度評価などが可能な GNR の拡張モジュール（**図 9**）を開発し、お客様にソリューション・コアとして提供しています。

［参考文献］

［1］ 社会保険庁：年金記録問題について，http://www.sia.go.jp/top/kaikaku/kiroku/，(2007).
［2］ 警察庁：犯罪収益移転防止管理官 (JAFIC)，年次報告書，(2009).
［3］ 金融庁：金融機関における本人確認について，http://www.fsa.go.jp/policy/honninkakunin/，(2008).
［4］ 全国銀行協会：反社会的勢力介入排除に向けた取組み強化について，http://www.zenginkyo.or.jp/news/2007/07/24152232.html (2007).
［5］ 日本 IBM：InfoSphere QualityStage，http://www.ibm.com/software/jp/data/infosphere/qualitystage/
［6］ IBM Corp: InfoSphere Identity Insight Solutions, http://www.ibm.com/software/data/identity-insight-solutions/
［7］ IBM Corp: InfoSphere Global Name Recognition, http://www.ibm.com/software/data/global-name-recognition/
［8］ 濱野正樹：IBM Information Server 概説，ProVISION，No.52，(2007).

日本アイ・ビー・エム株式会社
大和ソフトウェア開発研究所
InfoSphere 製品開発担当

福田 剛志 **Takeshi Fukuda**

［プロフィール］
1991 年日本 IBM 入社。東京基礎研究所にてオブジェクト指向データベース、データ･マイニング、オートノミック･コンピューティングなどの研究に従事。2004 年より大和ソフトウェア開発研究所にて情報統合関連製品の製品開発を担当し現在に至る。2009 年より IBM ソフトウェアグループマスターインベンター。博士（情報科学）。