Date : 12/07/2006

バーチャル国際典拠ファイル(VIAF : Virtual International Authority File): ドイツ図書館と米国議会図書館の典拠ファイルを利用して

Rick Bennett
OCLC
ダブリン、オハイオ
アメリカ合衆国

Christina Hengel-Dittrich
ドイツ図書館
フランクフルト·アム·マイン
ドイツ

Edward T. O'Neill
OCLC
ダブリン、オハイオ
アメリカ合衆国

Barbara B. Tillett
米国議会図書館
ワシントン DC
アメリカ合衆国

Translated by: Inahama Minoru
Suzuki Tomoyuki,

Meeting:	**123 Cataloguing**
Simultaneous Interpretation:	Yes

WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL

20-24 August 2006, Seoul, Korea

http://www.ifla.org/IV/ifla72/index.htm

概要

ドイツ図書館（DDB）、米国議会図書館（LC）、 OCLC（Online Computer Library Center）は共同で個人名に関するバーチャル国際典拠ファイル （VIAF）の開発に取り組んでいる。VIAF とは、世界の国または地域の書誌作成機関が作成する典拠レコード間にリンクを張り、インターネット上での無料の利用可能とするシステムである。今回の開発プロジェクトの目標は、異なる国で作成された典拠レコード同士を機械的にマッチングすることの実現可能性をさぐり、その有用性を証明することである。LC 及び DDB の書誌データベース、典拠データベースを利用して VIAF 初期システムを作成した。これには、6 百万件の名称レコードが 50 万強のリンクとともに収録されている。このプロジェクトは、典拠レコードと、それにリンクする書誌レコード、双方から情報を得た基にして名称を機械的にマッチングさせるアルゴリズムを開発することを主に目指した。その結果、異なる国の典拠ファイルからの個人名典拠レコードの同定を行うアルゴリズムが有効であることが、実際に明らかになった。LC、DDB の典拠ファイルに共通して存在する個人名典拠レコードのうち 70%に対して自動マッチングが行われ、エラー率は 1%未満であったのである。長期的には、VIAF プロジェクトは各国の国立図書館および他の主要情報提供機関の典拠レコードを統合し、典拠ファイルを世界的に共有することを目指している。

序

IFLA 目録分科会では、複数のグループがバーチャル国際典拠ファイル（VIAF）という考え方の重要性を認識していた[参考文献 1]。世界の国の書誌作成機関が作成するそれぞれの典拠レコードについて同一の実体を表わすもの同士で相互にリンクを張り、インターネット上で利用可能にするのである。VIAF は国際書誌調整の概念を実質的に拡張し、各国の全国書誌作成機関の目録作業を土台として構築される。VIAF においては、国ごとないし地域ごとに異なる多様な標目形が共存することが許容され、世界中のユーザが言語、文字、スペルの壁を越えて利用することができる。

昨今、機械的・自動的処理に際してウェブをよりインテリジェントにするためのオントロジーの活用が、ウェブの将来に関する提案としてあがっている。VIAF は、抄録・索引サービス、文書館、美術館・博物館、出版者などの情報提供機関が使用する統制語や典拠ファイルと結びついて、セマンティック・ウェブの基礎となる可能性を持っている[参考文献 2]。VIAF の構築は、図書館界がこのセマンティック・ウェブの将来に大きく貢献するチャンスであり、この構想を現実のものとするために力を入れるべきである。このためには、世界中のユーザが自由に利用できる VIAF を作り上げることが重要である。

他にも、個人名典拠レコード間のリンクについて検討をしてきているプロジェクトは存在する。LEAF プロジェクト(Linking and Exploring Authority Files)[参考文献 3]では、図書館、文書館、ドキュメンテーションセンター、 研究機関を含むさまざまな情報提供機関の典拠レコードにリンクをはることを計画した。ここで扱われるレコードは、フォーマットがさまざまであり、コンテンツの種類や量も相当変化に富んでいる。LEAF プロジェクトでは、これらの典拠レコードをシステムに落としこむ際にリンクをはる計画を立てた。典拠レコードの提供機関がさまざまであったため、リンクを確立する上で利用可能な共通の情報としては、名称形（「を見よ参照」、生没年を含む）しかなかった。現在、参加機関が作成する名称典拠レコードには生没年が入力されていないことが多く、LEAF の典拠レコードマッチングの工ラー率は容認しがたいほど高くなるであろうと予想される。

InterParty プロジェクト[参考文献 4]は、EU が資金を拠出し、電子資料の権利管理をサポートすることを主要な目的とし、さまざまな組織が構築する典拠ファイルにリンクをはるデモンストレーション・プロジェクトである。InterParty システムは、システムに参加している複数のデータベースにアクセスするのための共通の検索ポイントを提供する計画

の下に、まず中央集中型検索支援サービスの提供から開始した。データベース間の名称を同定する作業は目視で行われ、同定の判断を下した作業者がリンク生成の入力を行うと、リンクは機械的に利用可能になる。リンクは、それを作成した機関によっては、十分高い信頼性があると言えよう。ある機関が承認したリンクが他の参加機関からは認められないこともありえる。このプロジェクトではアルゴリズムによるマッチングの可能性を考慮してはいるが、リンクを補助するために必要な技術的要件、データ要件を明確にするには至っていない。

VIAF プロジェクト

2003 年に開催された IFLA ベルリン大会において、ドイツ図書館(DDB)と米国議会図書館(LC)、OCLC(OCLC Online Computer Library Center)は、バーチャル国際個人名典拠ファイルを開発することを合意した[参考文献 5]。この VIAF プロジェクトは、異なる国の典拠ファイルの典拠レコードを機械的にリンクすることの実行可能性を検証し、VIAF の考え方が有益であることを明らかにすることを目標としている。LC と DDB の典拠ファイル間にリンクを張り、バーチャル名称典拠ファイルに流し込む。 OCLC は 2 つの個人名典拠ファイル間の個人名典拠レコードをマッチングさせるソフトウェアの開発を担当している。VIAF プロジェクトの長期的な展望は、各国の国立図書館やその他の典拠作成機関等が構築する典拠ファイル間にリンクを張り巡らし、個人名、団体名、会議名、地名等を含めて世界規模の共有典拠サービスを展開することである。

VIAF プロジェクトは、次の 5 段階からなっている。

1. ドイツ図書館個人名典拠データ(Personennormdatei: PND) と LC 典拠レコードを統合して、「拡張典拠」レコードを作成する。これには、拡張典拠レコードとして適切なレコードを選択することと、受け取るファイルをどのように処理するかを決めることが含まれる。

2. マッチングのアルゴリズムを開発し、PND と LC の拡張典拠ファイルを突合して VIAF の初期バージョンを形成する。このプロセスは第 1 段階と並行して進められた。相互マッチングの結果から、どのような追加的情報を抽出し拡張典拠レコードに含めれば、マッチング処理が改善できるかがはっきりした。

3. VIAF へのアクセスを提供するために Open Archive Initiative (OAI)[参考文献 6] サーバを構築する。

4. VIAF データベースを継続的に維持管理するためには、参加機関それぞれの典拠レコード、書誌レコードの新規作成、更新が必要である。このような更新・維持管理システムの構築には、情報の更新に適した OAI を利用したプロトコルを使用する。

5. VIAF 典拠レコードにアクセスするためのユーザインタフェースはオープンウェブ上 で利用可能なものとする。最終的には、データベース、インタフェースともに Unicode、多言語、多文字種を扱えるようにする。例えば、LC の名称から検索して対応する PND の名称を HTML リンクで提供するようリクエストするなど、データベースへの直接のアクセスは、セマンティックウェブ機能をサポートする形で利用できる。

当プロジェクトは、まず、LC 名称典拠(Library of Congress Name Authority File: LCNAF)とドイツ図書館個人名典拠データ(Personennormdatei: PND)間にリンクを生成することによって、VIAF が実現可能であることを証明することに焦点を合わせた。2005 年 12 月 31 日時点で、LCNAF は 420 万件の個人名典拠レコードを保有し、LC は総計 930 万件の書誌レコードを作成、頒布している。

2005 年秋時点において、PND には 260 万件の個人名典拠レコードが収載されている。PND 典拠ファイルは、DDB 書誌レコードおよびバイエルン国立図書館 (Bibliotheksverbund Bayern: BVB)書誌レコードに使用されており、2 つの書誌ファイルをあわせて、総計 1500 万件の書誌レコードが PND 典拠レコードとリンクしている。

名称マッチングにおける問題点

まず、VIAF は人名に関する独-英および英-独辞書として機能する。例えば、アメリカ人の利用者が J. P. De Valk（LC における確立形）で検索を行うと、この名前は Johannes P.DE Valk（DDB における確立形）に自動的に「翻訳」される。このケースのように、異なる国際目録作成機関が、同一著者を異なった形の名称として確立することや、逆に、別人である複数の著者を表すために同じ形の名称を用いることは、めずらしくない。J. P. De Valk が、DDB によって、まったく別の著者に対する標目として確立される可能性もありえるのだ。

個人名は、同一人物について異なる形をとることもあれば、別々の人物について同じ形をとることもあり、異なる典拠ファイル間で信頼性の高いマッチングを行うことは容易ではない。2 つの典拠ファイルが対象とする範囲は明らかに食い違っており、双方のファイルに共通して存在する個人名はごくわずかである。従って、マッチングの信頼性を確かなものとするためには、名称それ自体以外の情報も用いなければならない。個人名典拠レコードにおいては、しばしば当該人物の生年および/ないし没年が示されている。通常の場合、似たような名前の人物を区別するには、生没年を組み合わせれば十分だからである。

補助的な情報なしに典拠レコードを同定することの難易度を確認するため、LC と DDB の典拠ファイルに共通する名称をサンプルとして抽出した。これらの典拠レコードの組について、目視でレビューを行い、同一人物であるかどうかを判断した。レビューの結果、こうした個人名の組のうち、10 パーセントは別人同士であることが判明した。このように、名称の確立形だけしか用いずにマッチングを行った場合、エラー率が高く、とても容認できるものではない。しかし、形が似ているからといって、むりやり類似の名称をマッチングすれば、エラー率はさらに高まってしまう。こうした安易な方法をとれば、少なからず存在する、2 つのファイル間において異なる形で確立されている名称形同士のマッチングにも失敗してしまう。

名称マッチングに関する問題に対する解決策

マッチングの可能性がある個人名同士が、本当にマッチングできるのかどうかを確認するためには、さらに追加の情報が必要なことは明白である。例えば、Diane Glynn について、以下の LC の典拠情報を考えてみたい。

100 10 $a Glynn, Diane, $d 1946-
400 10 $a O'Connor, Diane, $d 1946- $w nna
670 $a Country western dancing, 1994: $b CIP t.p. (Diane Glynn) pub. info. (an avid country w. dancer & co-author of How to make your man more sensitive)

ここでは、直接に用いることができる唯一の情報は、名前と生年である。フィールド 670（標目の情報源）には機械処理によって抽出されたと思われるタイトル 2 つが含まれている。ただし、実際問題としては、フィールド 670 からのタイトル抽出は、あまり信頼性が高くない。

当たり前のことであるが、書誌レコードは、ある個人名に関する追加的情報の情報源となる。書誌レコードは当該人物の著作に関する特性を提供し、その特性によってその個人を、類似の名称を持つ別の個人から区別することが可能になる。１件の書誌レコードには、以下のような情報が記録されている。

100 1 $a Glynn, Diane, $d 1946- -
245 10 $a How to make your man more sensitive / $c by Diane and
Dick O'Connor.
700 1 $a O'Connor, Dick, $d 1938- $e joint author –

書誌レコードには 2 種類の追加的情報が含まれている。書誌レコードには、通常、タイトルなどのように著作を特定する情報と、ISBN などのように体現形を特定する情報とが含まれている。タイトルによるマッチングは、名称マッチングを最終的に確定させる働きをする。

共著者がいる書誌レコードも、また、追加的情報の提供源になる。共著者の情報は、特定のタイトルによるマッチングができない場合に、著者をマッチングする上での手助けとなる。共著者である Dick O’ Connor は、この種の情報の一例である。Dick O’ Connor は複数の著作において Diane Glynn の共著者となっており、このことは、複数の典拠ファイル間で名称マッチングを行う際の大きなサポートとなる。同一の著作が両国の書誌データベースに存在しているが、一方が翻訳書であった場合、タイトルによる機械的なマッチングは厳しいが、共著者は両データベースに類似の形で存在する可能性が高く、マッチングを確定することができる。

このようにして、当該名称が基本記入、副出記入、件名として含まれるすべての利用可能な書誌レコードを加工し、「抽出典拠レコード」と呼ばれる中間成果物的なレコードを作成する。次いで、これらの抽出典拠レコードをオリジナルの典拠レコードと組み合わせ、拡張典拠レコードを作成する。拡張典拠レコードには、書誌レコードに由来する名称と結びついた追加的情報が含まれているので、典拠レコードをそのまま用いた場合よりも、より信頼性の高いマッチング処理をサポートすることが可能となる。

名称マッチングの確定

2 つの国の典拠ファイルの名称を単純に比較するのは、同一の個人を見つけ出す上で合理的な方法ある。しかし、名称の形は必ずしも同じではない場合もあり、同一人物を同定する妨げとなる。こうした個人名のマッチングを機械的に確定するために、ここで

は、(1)名称同士に互換性があるか、(2)マッチングを確定する十分な補完的情報があるかという方向から考えてゆくことにする。

名称同士に互換性があるとは、複数の名称形に互いを別人であるととして排除するに足るだけの違いがないということである。John A.Smith と John Allen.Smith のように、完全さの度合いという点で名称が異なる場合がある。これらの名称は、A が Allen の代わりとなりうることから、互換性がある。しかしながら、John A.Smith と John B.Smith では、ミドルネームのイニシャルが不一致であるため、互換性はない。互換性を考える場合は、名称の典拠形と名称の参照形とを考慮に入れる必要がある。

名称間に互換性があると判断したら、それらの名称についての補完的情報を使い、マッチングを確定させる。書誌ファイルには、実は異なるにもかかわらずよく似たタイトルや名称が、多く存在する。それでも、もしタイトルと名称の組み合わせが双方の書誌ファイルにおいて類似していれば、その名称が同一人物を表している可能性は非常に高いということになる。この基本的な考え方を押し進めて、書誌レコードからそのほかタイプの情報を収集する。

日付に関する情報が一致した場合は、それだけで、強い相関関係があると見なす。日付が 2 年以上異なっている場合は、名称には互換性がないと見なし、マッチングを拒否する。日付の相違が 1 年以内である場合、マッチングは許容される。VIAF 開発の過程では、日付の一部に小さな食い違いが見受けられるのは比較的よくあることで、日付に多少の相違はあっても、日付以外の補完的情報はマッチングを確定するに足るに十分であった。

2 つの拡張典拠レコードの比較する作業において、マッチングさせるそれぞれの要素をマッチングポイントと呼ぶ。マッチングポイントは「強い・普通・弱い」の 3 段階に区分した。互換性のある名称形の同定には、強いレベルのマッチングポイントが 1 つ見たされていれば、同一人物と判定できる。強いレベルのマッチングポイントとは、タイトル、ISBN、生没年、共著者名である。生年のみの場合は、同一人物、同名異人の判断基準には十分ではなかったため、情報が生年のみの場合、普通レベルのマッチングポイントとして扱う。普通のレベルのマッチングポイントには、その個人の著作活動に関する情報、例えば、よく利用する出版者、取り扱う主題領域、役割 (挿画家、作曲家など) が含まれる。多数の著者の作品を取り扱う大きな出版者であれば、その中にはよく似た名前の別人がいることも少なくないだろうが、それでも普通レベルのマッチングポイントをいくつか組み合わせれば、同定を行うには十分と言えよう。弱いレベルのマッ

チングポイントは、他の方法では同定し切れないときにのみ有効であり、言語、取り扱う主題領域、出版国がこれに当たる。

複数のマッチングポイントを組み合わせて使用するために、それぞれの要素のマッチングに対して得点を与えることにした。ISBN のような番号の場合は、正確なマッチングが行われなければ、マッチングとは認めず、マッチングすれば 1 得点、しなければ 0 得点が与えられる。タイトルのようなテキスト形式のデータのマッチングは、テキストの類似度によって 0 点から 1 点の間で得点が加えられる。テキストの類似度判定には、スコア付与手法に基づくトリグラム方式を採用する。それぞれの得点はマッチングポイントのレベルごとに重み付けをし、修正した後に、加算する。総得点がテスト結果に基づいて設定した基準値を上回れば、マッチングを確定する。実際のマッチングアルゴリズムは、多数のレコードでテストを行った結果を受けて、3 段階の得点付与について調整を行われている。今後、別の典拠ファイルがシステムに投入され、経験を重ねながら、さらなる調整を行うことになるであろう。

拡張典拠レコードの作成

上記の手法は、PND、LC の名称典拠レコードの拡張典拠レコードを作成するのに使用した。LC 典拠ファイルを拡張するために典拠レコードを加工するべく LC 書誌ファイルの処理を行った。DDB と BVB 書誌ファイルは PND 典拠レコードを拡張するために処理にかけられた。拡張典拠レコードに関する情報の流れを Figure 1 に簡単に図示した。

LC の拡張典拠ファイルでは、420 万件の典拠レコード中 380 万件 (90%) が拡張処理の対象となった。うち、書誌レコード (計 740 万タイトルが利用された) からの情報を使い拡張されたのは 260 万件 (60%) だけであった。残りのレコードは、当該典拠レコードのフィールド 670 (標目の情報源) から抽出された 410 万タイトルによって拡張されている。マッチングを行うためには、タイトルは何よりも重要な拡張要素である。このことについては結果報告のところでも振り返る。

PND の拡張典拠ファイルでは、260 万件の典拠レコード中 240 万件 (90%) に何らかの拡張がなされた。うち、書誌レコードからの拡張が行われたのは 200 万件 (80%) のみである。残りの 40 万件の典拠レコードは、その典拠レコードそのものから抽出されたタイトルによって、拡張が行われた。

マッチング手法のテスト

VIAF プロジェクト関係機関は、マッチング処理システムの開発にあたって、処理結果の精密な検証、意見交換を行ってきた。例えば、シリーズタイトルは初めマッチングに使用されていたのだが、不正確なマッチングを作成する頻度が高いことが判明した。検証を重ねるたびに、修正を加え、正しいマッチング数の増加や誤ったマッチング数の減少に結びついている。この期間に、マッチングの得点について精度の高い基準値を設定し、得点手法のアルゴリズムを開発した。ここでは、最終確定テストについて説明する。

マッチング処理の正確性、有効性を確認するために、DDB、LC 双方のベテランの典拠作業カタロガーがマッチングのための名称のサンプルを検証した。最初のサンプル検証の目的は、2 つの典拠ファイルの間でどれくらいの名称が共通して存在しているのかを割り出し、その共通する名称のペアのうちどの程度までがマッチング処理により同定可能かを調査することである。2 番目のサンプルは、システムエラーや修正できる欠陥を洗い出し、全体のエラー率を推定するために検証を行った。

1 番目のサンプルには、まず、391 件の PND 典拠レコード無作為抽出で選び出した。このレコードをマッチングさせるために、LC 典拠ファイルに、機械的、人力の両方の手段で検索をかけた。サンプル中の PND 典拠レコードは、姓が同形の LC 典拠レコードとペアを構成し、74,000 組が検査の対象となり、マッチングのためのアルゴリズムにより 79 組の PND/LC 典拠レコードが自動でマッチングされた。

PND 典拠レコード 391 件を人の手により再検索した結果、さらに追加で 35 件のレコードが LC 典拠レコードと対応することが分かった。しかし、これは姓の形が異なるマッチングであるか、マッチングアルゴリズムでは同定できないケースであった。79 件の自動マッチングは、この人の手を介して行われた検証作業で、正しいマッチングであることが確認された。この PND サンプルを使った調査から、PND の個人名のうち約 30%には LC 典拠レコードが存在し、この共通に存在する名称のうち 70%はアルゴリズムを使用してマッチングできるという見込みが立った。これを言い換えれば、2 つの典拠ファイルは 80 万件の名称を共通して持ち、そのうち 55 万件が機械的マッチング処理で同定できるということである。

この結果検証は、名称の組み合わせを作成する処理方法を改善するためにも用いられた。姓だけを使用すると、マッチング 1 件を生成するために、ほぼ 1,000 件の組み合わ

せをマッチング処理全工程にかけなければならない。目視で検証したテスト結果から、姓、名、限られた範囲での生没年情報を使い名称同定の粗分けをする方針にを転換した。この単純なインデックス方法でもマッチングの 95%は特定できるし、1 件のマッチング生成に対して 4 件の組み合わせを検証するだけでよいことが分かった。インデックスは単純でも必要十分であり有効に機能するし、多少の工夫を行えば結果はさらに改善される。

2 つ目のサンプルは、マッチングのエラー率を推定するために使用した。作業の一環として、サンプルを使い最初に設定した得点の基準値の妥当性を検証し、必要に応じて修正を加えた。得点基準値の性質をから考えれば、基準値すれすれの得点を得て成立したマッチングは、基準値を大きく上回る得点で成立したマッチングよりもエラー率が高いことが予測される。大半のマッチングは基準値をはるかに上回る成績で成立している。最小限の要員で正確なエラー率を算出することを目指して、サンプルを得点に応じて 4 つのサブグループに分けた。マッチングがエラーかどうかを人手をかけて検証し、サブグループごとにエラー率と信頼性を求めた。このサブグループごとの結果を重み付けし、足し合わせてマッチング処理全体に対するエラー率を算出した。マッチングエラー率は 1%未満であった。

サブグループのひとつは、わずかであるが基準値以下の得点のものを対象とするグループである。もし基準値を引き下げ、このグループのマッチングが承認されると、正しいマッチング 3 件に対して 1 件の正しくないマッチングが追加されてしまうことになる。明らかに、今回設定した基準値を引き下げることは認められない。基準値をわずかではあるが上回っているグループでは、25 件のマッチング中に 1 件しか正しくないマッチングが存在しないのである。このサブグループはマッチングのサンプル数が少ないため、全体のエラー率への影響はごく小さく、全体的には大多数のマッチングは正しいのである。以上の結果から、基準値の初期設定レベルは妥当であると判断した。

VIAF 初期システムの構築

LC、PND それぞれの典拠から作成された拡張典拠ファイルは、マッチングアルゴリズムにかけられ、そこから生成するレコードは、マッチングの可否にかかわらず、VIAF レコードとして変換される。この処理については Figure 2 に図示した。VIAF ファイルに変換された 630 万レコードの内訳は、マッチングしたレコードが 50 万件、マッチングしなかった LC 典拠レコードが 370 万件、PND 典拠レコードが 210 万件であった。

これは、人手により行ったテストから予測していた結果に非常に近い。最終的には、このように機械的ではなく確定されたマッチングやそれ以外の知的手段で確認されたマッチングについて、人手を使ってリンクを作成できるシステムにする予定である。典拠レコードには VIAF レコード番号が連番で付与される。

Figure 3 は、VIAF レコードを MARC21 フォーマットで示した例である。VIAF の最大の目的は、ファイル間のリンク生成であり、VIAF レコードには、それぞれの名称が作成機関とともに、フィールド 700 (標目リンク記入) に記入される。唯一の存在としての確立形は存在しないため、フィールド 100 (個人名著者標目) は使用しない。アルゴリズムによりマッチングが行われた場合、標目リンク記入/エントリが 2 つ生成される。マッチングが行われなかった書誌にはフィールド 700 が 1 つしか出現しない。

拡張典拠レコードには補完的情報も含まれ、フィールド 9xx (ローカルフィールド) に収められる。Figure 4 に、拡張典拠レコードにおいて使用されるローカルフィールドを示す。マッチングを単純化するため、テキストの標準化はすべて NACO (Name Authority Cooperative Program of the Program for Cooperative Cataloging) 標準化規則を使って行われた[参考文献 7]。ある特定の用語の出現回数は、サブフィールド$9 に格納される。これは何よりも機械処理のための情報なので、エンドユーザ用の画面には必ずしも表示しなくともよい。新たに他の国の典拠ファイルが投入すると、その典拠レコードは、まず既存の拡張 VIAF レコードと比較され、それが VIAF レコードに変換される時に、新規のマッチングを追加する。マッチングが成立すると、拡張部分の情報も統合される。

ある典拠ファイルの 1 典拠レコードが他の典拠ファイルの複数のレコードとマッチングしてしまうケースは非常に多い。VIAF の目指すところは 1 対 1 のリンクであるために、このような複数のレコードとマッチングした場合は、マッチングと認めなかった。このためアルゴリズムにより生成されたマッチングのうち 7 万件が取り消された。今のところ、複数レコードとのマッチングが生成されてしまう原因は 2 つある。

1 つ目は、PND に同名異人の区別がつけられない典拠レコードが多数含まれることである。このようなレコードは 2 つ以上の LCNAF 典拠レコードにマッチングしてしまう。ドイツでは、ドイツ目録規則である RAK-WB に則り、(同名異人の) 個人名の区別をしていなかった。DDB が典拠ファイルを構築し始めてからはこの規定は適用されなくなり、現在、DDB では同名異人の区別を行っている。しかし、PND にはなお同名異人の区別がつけられないレコードが多く残っている。DDB は、拡張典拠レコードに記録された LC と DDB からのタイトル情報により、可能な限り機械的に、これらの複数マ

ッチングを引き起こした典拠レコードの分別を行う予定である。機械的処理が不可能な場合には、人手をかける。訂正が行われると、結果は定期更新により VIAF に反映され、1 対 1 のリンク作成に至る。

2 つ目の原因は、LC 典拠が、同一人が異なる書誌的実体を持つとき（ペンネーム使用など）に、それぞれの実体に対して異なる典拠レコードを作成することに起因する。これは AACR2 の規定によるものであり、PND の同名異人を区別しないことと対照的である。LC の場合、同一人物に対して複数の典拠レコードが作成されてしまうのである。RAK-WB に従った PND では、同一人物の異なる実体すべてに対して 1 つのレコードのみが作成される。同名異人の区別ができない典拠レコードと同様に、これらの「細かくされすぎた」典拠レコードも、十分満足の行く解決法がみつからない問題を提起する。

リンク先の名称は、（LC 典拠から PND 典拠へも、その逆も）自動翻訳され直接利用できる。これは、セマンティックウェブやこのような機能を必要とする統合検索システムの必要条件を満たしている。「から見よ（see from）」参照を整備すれば、人の目にも分かりやすくなる。

参加館典拠ファイルの典拠番号や VIAF 番号も URI の元になる。典拠 URI の積極的サービスに向けての可能性を示唆している。文書、記録/レコード、ウェブ中に登場するどんな URI からでも、ユーザはすべての資料、記録、資源等–URI の形での典拠が紐づく–や典拠レコードそのものにたどり着くことができるようになる。

現行システムについて

国の名称典拠ファイルと書誌データベースはどちらも常に追加･修正が加えられている。変化し続ける 2 以上のデータベースからなるリンクデータベースは、そのリンクを頻繁に見直し、更新しなくてはならない。VIAF 初期システムのロジック、ソフトウェアは、継続的なレコード更新に対応すべく、変更を加えられているところである。新規の書誌レコードまたは典拠レコードが追加されると、既存の拡張典拠レコードに変更が加えられ、データベース間のリンクの見直しが行えるようになる。新規に生成されるリンクも次々あるし、基礎になる源データに変更があり根拠を失ったマッチングは廃棄される。破棄に際しては、過去のマッチングの履歴はおのおののレコードに参照情報として記録される。

VIAF システムでは、将来的に、その情報源となるデータベースの管理者からの OAI によるデータ送付が可能になったときには、それを活用する予定である。今のところ、このテストプロジェクトには、FTP のような既存のファイルアクセス方法を利用してゆく。

大量のデータを一箇所に集中させると、そのデータにどうアクセスするかやそれをどう使用するかに関してさまざまな可能性が考えられるようになる。リンクが張られていることにより、個人名をセマンティックウェブの一環としてエンドユーザの希望するフォーマットへ変換することができるし、次々と書誌データベースを渡りながら、そのデータベースに適した名称で自動検索をするツールを開発することも不可能ではない。目録作成や典拠コントロールのためのツールも、書誌レコード中の個人名の形式を選択するという同様の考え方で開発できる。言うまでもなく、VIAF データベースへの直接検索も可能である。

結論

PND 側では、このプロジェクトへの参加が、すでに実質的なメリットとなって帰ってきている。LC ファイルとの機械的マッチングテストの結果、PND ファイルに目に見えて改善があったのだ。DDB は、拡張レコードのペアに含まれるタイトルのマッチングが、個人名の識別作業にとって実質的な役に立つと期待している。このプロジェクトのために開発したプロセスやアルゴリズムは、他にも種々のアプリケーションに応用可能である。書誌情報へのアクセスを向上させ、参加機関の目録作業を支援するために、個人名のマッチングデータを活用できるようなどんなサービスがあるか詳しい調査を行っているところである。

プロジェクトにより分かったことは、2 つの国の典拠間での機械的なリンク生成に意味があることである。双方のファイルに共通して存在する個人名の典拠レコード中 70% がリンクされ、そのエラー率は 1%未満だった。元の典拠レコードに書誌レコードからの情報を補う手法はマッチングの成功率を大きく上げ、不正確なマッチングを減少させた。典拠レコードを少し加工するだけでマッチングの結果が大幅に向上するのだ。マッチング失敗の原因は、大半が、フィールド 670 (標目の情報源) の解析ミスである。データの構造を追加し名称やタイトルを簡略に記録しないこと、または、出典書誌レコードとの間に明示的なリンクを張ることで、問題解決の一助になるであろう。通常の役割表示や専攻 (作曲家、挿画家、数学者等) もきちんと記録されていれば、また、少なく

とも相互参照に名前のスペルの完全形（イニシャル形ではなく）いれば、自動、手動を問わず、マッチングの幅が広がるのである。

この研究は、典拠コントロールについて、典拠レコードの活用について、ネットワークや相互リンクについて、そして、図書館におけるセマンティックウェブの構築について説得力のある結果を出している。ドイツの図書館や図書館ネットワークで LCNAF のアクセスポイントをもつ書誌レコードを入手したり、そのままで利用しているところにとっては、VIAF は、書誌中の LCNAF のアクセスポイントを PND のアクセスポイントへ変換する、または、VIAF を通じて PND 標目を検索することができるなど、2 つの典拠ファイルをまたがって使えるプラットフォームとして機能することができる。The European Library ポータルのように国家や言語を超えたポータルのなかに VIAF を組み込めば、LCNAF と PND 双方の検索クエリは自ずと統合されるであろう。そして、利用者は、両方のデータベースでその典拠レコードに紐づく書誌レコードに到達できる。

マッチングの手法は定まったので、OAI によるデータ送付を利用して、参加機関から個人名典拠データと書誌データを集め更新するシステムの開発を計画している。このシステムは拡張性のあるデザインで、典拠レコード、書誌レコードの共有を希望する新規参加機関を求めている。VIAF の限界については、今後参加機関の増加を待って結論を出したい。

VIAF プロジェクトは、典拠レコードのマッチングに焦点をしぼったものである。今後 VIAF を維持し、拡張して、実際に使用していくには、長期的な視点でのサービスと維持管理に関する戦略が必要となる。団体名も含める方向でのプロジェクトの拡大、参加機関の追加については議論を重ねていかなければならない。Unicode 文字セットを導入し、システム能力を改善する計画がある。Unicode により非ローマ字の文字種を取り込むことができるようになるが、マッチングアルゴリズムの拡張開発は容易ではないだろう。中・日・ハングルの表意文字が基本になる言語でのケースが特に難問である。

参考文献

1. IFLA Core Activity: IFLA-CDNL Alliance for Bibliographic Standards (ICABS) http://www.ifla.org.sg/VI/7/icabs.htm [May 2006]

2. Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities." Scientific American, May 17, 2001. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 [May 2006]

3. LEAF Project, http://www.leaf-eu.org [May 2006]

4. Project InterParty: From Library Authority Files to E-Commerce, Andrew MacEwan, http://www.haworthpress.com/store/E-Text/View_EText.asp?a=3&fn=J104v39n01_11&i=1%2F2&s=J104&v=39 [May 2006]

5. VIAF: The Virtual International Authority File, http://www.oclc.org/research/projects/viaf [May 2006]

6. Open Archives Initiative - Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.html [May 2006]

7. Hickey, Thomas B., Jenny Toves, and Edward T. O'Neill. "NACO Normalization: A detailed Examination of the Authority File Comparison Rules", Library Resources & Technical Services, Vol. 50, No. 3, p. 18-24. [forthcoming]

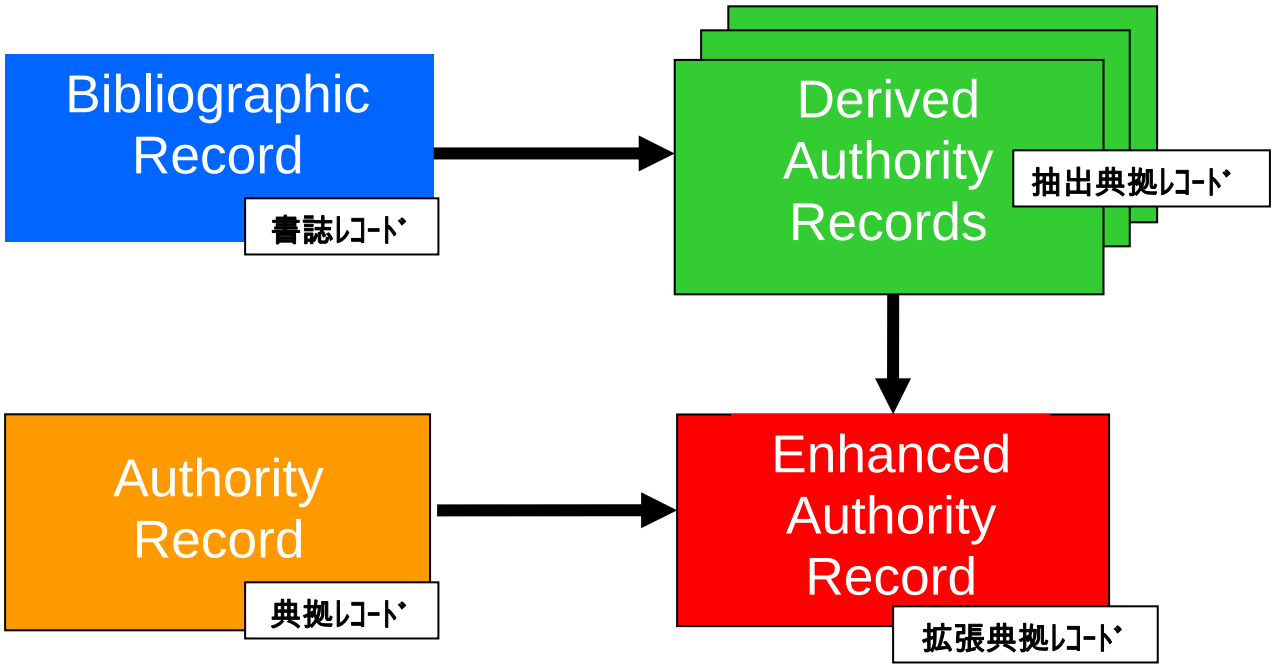

Figure 1. 拡張典拠ﾚｺｰﾄﾞの作成

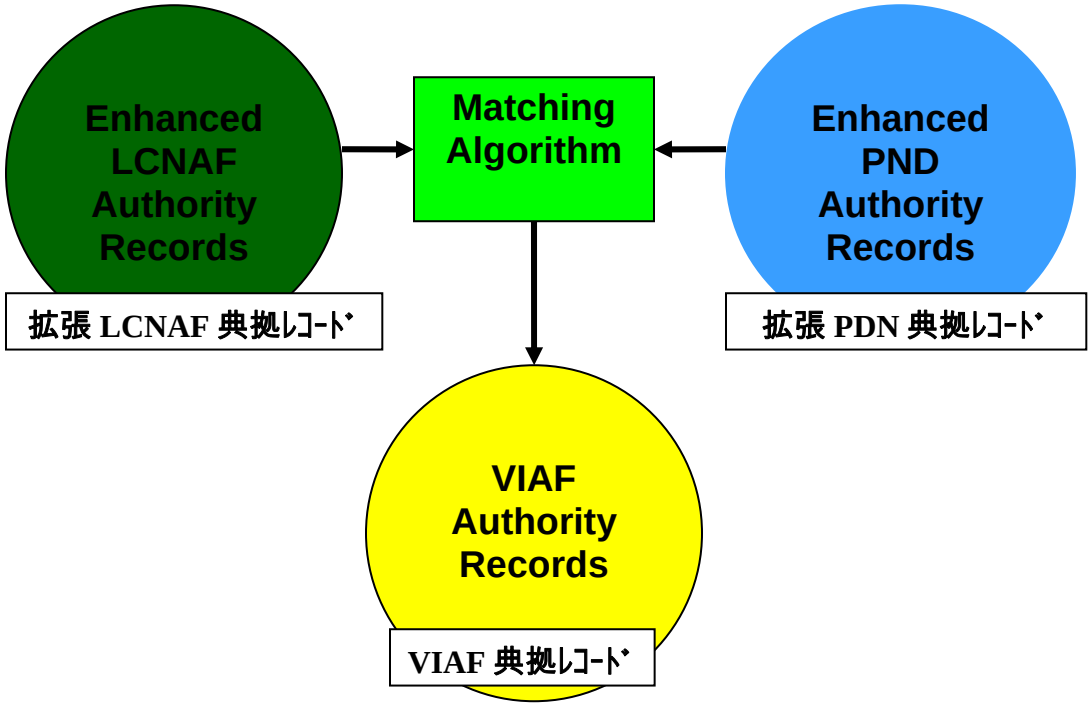

Figure 2. VIAF 典拠ﾚｺｰﾄﾞの作成

000 nz n
001 viaf 30543
005 20050826163535.0
008 050826n||anannabbn |a aaa
040 VIAF $c VIAF
400 10 $w nnaO'Connor, Diane, $d 1946-
700 17 Glynn, Diane, $d 1946- $2 DLC $0 n 94057411
700 17 O'Connor, Diane $2 DDB $0 108982424
901 052512920 $9 1
901 349917275 $9 1
901 350215532 $9 1
903 75014386 $9 1
910 11 how to make your man more sensitive $9 3
910 11 macht eure manner zartlicher $b liebevolle ratschlage fur e neues rollenverhalten $9 1
910 11 macht eure manner zartlicher $b wie e frau ihrem mann helfen kann e verstandnisvoll
$9 1
919 country western dancing, $9 1
920 0-525 $9 1
920 3-499 $9 1
920 3-502 $9 1
921 dutton $9 1
921 rowohlt $9 1
921 scherz $9 1
922 gw $9 2
922 nyu $9 1
940 eng $9 1
940 ger $9 2
942 18 $9 1
943 197x $9 3
944 am $9 3
950 11 oconnor, dick $9 2
950 11 oconnor, dick $d 1938 $9 1
999 1 $b 75014386 //r94 $2 DLC
999 1 $b n 94057411 $2 LoCNA
999 2 $b 780147766 $b 790425319 $2 DDB

Figure 3. VIAF Record

Figure 4
拡張レコードフォーマット

90x	Control numbers		
	901	ISBN	$a Numeric portion of ISBN (no check digit or dashes)
	902	ISSN	$a Numeric portion of ISSN (no check digit or dashes)
	903	LCCN	$a Numeric portion of LCCN (no check digit or dashes)
91x	Title fields		
	910	Title from 245	Subfields a & b
	911	Abbreviated title from 210	Subfields a & b
	913	Uniform title from 130 or 240	Subfields a & b
	914	Translated title from 242	Subfields a & b
	915	Collective uniform title from 243	All subfields
	916	Variant title from 246	Subfields a & b
	917	Authority Record Uniform Title	Extracted from Name/Title authority records, field 100 $t
	919	Title extracted from other text	Various note or similar fields
92x	Publisher fields		
	920	Publisher number	$a Publisher number from ISBN
	921	Publisher name	$a Publisher name from the 260 b or 533 c.
	922	Place of publication	$a Country of publication code from 008
93x	Usage		
	930	Name Usage	$a Form of name found in the statement of responsibility, 245 subfield c
94x	Attributes		
	940	Language	$a Language code from the 008 or 041 subfield a
	941	Author's role	$a Relator code from 700, subfields e and/or 4
	942	NATC Subject	$a NATC survey line number.
	943	Decade of publication	$a Decade of publication
	944	Format	$a Type and bib level (008/06-07)
	945	Conspectus Subject	Custom usage, see PND discussion
95x	Joint Authors		
	950	Personal Authors	Subfields $a, $b, $c, $d, and $q from either the 100 or 700 fields
	951	Corporate Authors	Subfield $a from either the 110 or 710 fields
96x	Name Subjects		
	960	Name as Subject	Sub-fields $a, $b, $c, $d, and $q from the 600 field
	969	Subject usage	Text “Subject” indicating the authority heading was used as a subject, and was extracted from a 600 field
99x	Special Fields		
	999	Associated bibliographic records	$a Total number of records $b Record Control Number $2 Source of Record