Date : 25/07/2006

VIAF (Virtual International Authority File): Verlinkung der Normdateien der Library of Congress und der Deutschen Nationalbibliothek

Rick Bennett
OCLC Online Computer Library Center
Dublin, Ohio, USA

Christina Hengel-Dittrich
Die Deutsche Bibliothek
Frankfurt am Main, Germany

Edward T. O'Neill
OCLC Online Computer Library Center
Dublin, Ohio, USA

Barbara B. Tillett
Library of Congress
Washington, D.C. USA

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

Abstract

Die Library of Congress, die Deutsche Nationalbibliothek und OCLC Online Computer Library Center entwickeln gemeinsam eine virtuelle internationale Normdatei (Virtual International Authority File, VIAF) für Personennamen, die Normdatensätze der nationalbibliografischen Zentren weltweit miteinander verbindet und frei zugänglich ins Web stellt. Ziel dieses Projektes ist es nachzuweisen, dass die automatische Verknüpfung von Normdatensätzen aus verschiedenen nationalen Normdateien gelingt, und die daraus entstehenden Vorteile zu veranschaulichen. Die Normdateien und die Titeldatenbestände der Library of Congress und der Deutschen Nationalbibliothek wurden genutzt, um einen ersten Basis-VIAF zu schaffen, der über sechs Millionen Namen mit mehr als einer halben Million Links enthält. Ein Hauptaspekt des Projektes war die Entwicklung von automatischen Namensabgleich-Algorithmen, die Informationen sowohl aus den Normdaten als auch aus den zugehörigen Titeldaten auswerten. Es wurde nachgewiesen, dass die Personensätze aus beiden nationalen Normdateien auf der Basis der entwickelten Algorithmen erfolgreich automatisch verlinkt werden konnten; siebzig Prozent der Personennamen, die in beiden Normdateien enthalten waren, wurden mit einer Fehlerquote von unter einem Prozent verlinkt. Strategisch zielt das VIAF-Projekt darauf ab, die normierten Namensformen aus vielen nationalen Normdateien und anderen wichtigen Quellen in einem gemeinsamen Normdaten-Service zusammenzuführen.

Einführung

Mehrere Gruppen der Katalogisierungs-Sektion der International Federation of Library Associations and Institutions (IFLA), haben die Möglichkeiten erkannt, die in einer virtuellen internationalen Normdatei (VIAF) [1] liegen, in einem VIAF, in dem die unterschiedlichen Normdatensätze, die weltweit verteilt in den nationalbibliografischen Zentren dieselben Entitäten repräsentieren, miteinander verknüpft und im Internet verfügbar gemacht werden. Ein solcher VIAF würde eine sinnvolle Erweiterung des Konzeptes weltweiter bibliografischer Kontrolle darstellen und auf der Arbeit der einzelnen nationalbibliografischen Zentren aufbauen. Er würde ein Miteinander unterschiedlicher nationaler oder regionaler Namensformen zulassen und gleichzeitig den Benutzern in aller Welt in ihren unterschiedlichen Bedürfnissen hinsichtlich Sprache, Schrift und Schreibweise entgegen kommen.

Gegenwärtige Ansätze zur Zukunft des Web empfehlen die Nutzung von Ontologien, um das Web für maschinelle und automatische Prozesse intelligenter zu machen. Der VIAF könnte - kombiniert mit anderen kontrollierten Vokabularien und Normdateien anderer Anbieter, wie zum Beispiel Abstract- und Indexierungs-Dienste, Archive, Museen, Verlage etc. - einer der Grundbausteine für ein „semantisches Web“ [2] sein. Bibliotheken haben jetzt die Gelegenheit, einen großen Beitrag für diese Zukunft zu leisten, und sollten helfen, die Vision Realität werden zu lassen. Für den Erfolg der gemeinsamen Vision ist es sehr wichtig, dass der VIAF allen Nutzern weltweit frei zugänglich gemacht wird.

Auch andere Projekte haben die Möglichkeiten untersucht, Personennamen aus Normdateien miteinander zu verknüpfen. Im Projekt LEAF [3] (Linking and Exploring Authority Files) wurde vorgeschlagen, Normdatensätze aus vielen verschiedenen Quellen, wie z.B. Bibliotheken, Archiven, Dokumentations- und Forschungszentren miteinander zu verknüpfen. Die Datensätze haben unterschiedliche Formate, und sie unterscheiden sich erheblich in Detaillierungsgrad und Umfang der inhaltlichen Angaben. Das LEAF-Projekt beabsichtigte, die Datensätze automatisch zu verknüpfen, wenn sie ins System geladen werden. Aufgrund der heterogenen Herkunft der Normdatensätze erwies sich aber der Name, einschließlich Namensvarianten und individualisierende Angaben, als die einzige Vergleichsinformation, die als Grundlage für eine Verlinkung herangezogen werden kann. Da die Personensätze der derzeitigen Teilnehmer häufig keine individualisierenden Angaben enthalten, ist bei den Abgleichen mit einer nicht-Übereinstimmungsquote zu rechnen, die unakzeptabel hoch ist.

Das InterPartyProject [4] ist ein EU-finanziertes Demonstrationsprojekt, in dem Normdateien zwischen verschiedenen Organisationen verlinkt werden, hauptsächlich um das Digital-Rights-Management zu unterstützen. Das im Projekt entwickelte InterParty-System sieht vor, dass über einen gemeinsamen Zugang auf alle im System beteiligten Datenbanken zugegriffen werden kann, d.h. es bietet zu allererst einen zentralisierten Suchdienst. Da die Links zwischen den Namen in jeder der Datenbanken intellektuell verifiziert werden müssen, kann der Bearbeiter, der die Identifizierung vornimmt, auch den Link eingeben. Die einmal hergestellten Links können dann automatisch nachgenutzt werden. Abhängig von der Organisation, die die Links hergestellt hat, sind sie als verlässlich zu betrachten. Das Setzen eines Links durch eine Partei muss durch andere Parteien innerhalb des Systems nicht akzeptiert werden. Im Projekt sind

algorithmische Abgleichverfahren zugelassen, es gibt aber keine Vorgaben für anzuwendende Techniken oder Anforderungen, welche Daten zur Verlinkung notwendig sind.

Das VIAF-Projekt

Während des IFLA World Library and Information Congress 2003 in Berlin beschlossen die Library of Congress (LC), die Deutsche Nationalbibliothek (DNB) und das OCLC Online Computer Library Center (OCLC), eine virtuelle internationalen Normdatei (VIAF) für Personennamen [5] zu entwickeln. Die Ziele des VIAF-Projektes liegen darin, nachzuweisen, dass die automatische Verknüpfung von Normdatensätzen unterschiedlicher nationaler Normdateien ein gangbarer Weg ist, und die Vorteile eines VIAF zu demonstrieren. Im VIAF-Projekt werden die Normdateien der Library of Congress und der Deutschen Nationalbibliothek für Namen in einer gemeinsamen virtuellen Namen-Datei miteinander verknüpfen. OCLC entwickelt die Software, die benötigt wird, um die Personennamensätze der beiden Normdateien abzugleichen.
Strategisch zielt das VIAF-Projekt darauf ab, die normierten Namensformen aus vielen nationalen Normdateien und anderen wichtigen Quellen in einem gemeinsamen weltweiten Normdaten-Service für Personen, Körperschaften, Kongressen, Geografika, etc. zusammenzuführen.

Das VIAF-Projekt besteht aus fünf Phasen:

1. Erstellen „Erweiterter Normdatensätze“ sowohl aus der Personennamendatei (PND) als auch aus dem Library of Congress Name Authority File (LCNAF). Dazu gehört die Identifikation der zutreffenden Normdatensätze, die in die „Erweiterten Normdatensätze“ integriert werden; ebenso die Festlegungen für gegebenenfalls erforderliche Bearbeitungsschritte bei der Übernahme der Normdateien und Titeldatenbanken.

2. Entwicklung von Matching-Algorithmen sowie Abgleich der „Erweiterten Normdatensätze“ aus PND und LCNAF, um die Ausgangsversion des VIAF zu herzustellen. Phase 2 und Phase 1 verliefen als iterativer Prozess: Zwischenergebnisse aus dem Matchingprozess ergaben Hinweise auf zusätzliche, für den Abgleich wichtige Datenelemente, die aus Titeldaten extrahiert und in die „Erweiterten Normdatensätze“ integriert wurden, um im Matchingprozess die Trefferzahlen zu verbessern.

3. Aufbau eines OAI-Servers (Open Archive Initiative Server) [6], um den Zugang zum VIAF zu ermöglichen.

4. Aufbau des Datenhaltungs- und Update-Systems. Zur laufenden Datenhaltung der VIAF-Datenbank ist die Einbeziehung aller Neuansetzungen und Datenänderungen in Norm- und Titeldaten der beteiligten Institutionen notwendig. Das Datenhaltungs- und Update-Verfahren wird sich an den OAI-Protokollen orientieren.

5. Einrichtung einer Benutzerschnittstelle im Web, um den Zugang zum VIAF zu ermöglichen. Längerfristig werden die VIAF-Datenbank und die Benutzerschnittstelle

Unicode sowie Mehrsprachlichkeit und Mehrschriftlichkeit unterstützen. Semantic-Web-Instrumente werden in Zukunft die Suche unterstützen. So werden Suchfragen an die Datenbank, in denen zum Beispiel ausgehend von einer LC-Namensform nach dem zugehörigen PND-Namen gesucht wird, über einfache HTML-Links gestellt werden können.

Das Projekt konzentriert sich zunächst darauf, die Machbarkeit des VIAF zu demonstrieren, indem die Personennamensätze des Library of Congress Name Authority File (LCNAF) und der Personennamendatei (PND) miteinander verknüpft werden.
Am 31. Dezember 2005 umfasste der LCNAF-File 4,2 Millionen Normdatensätze für Personennamen, und zum selben Zeitpunkt verfügte die LC über einen Datenbestand von 9,3 Mill. bibliografischen Datensätzen, die gemeinsam mit dem LCNAF als Basis für den VIAF zur Verfügung gestellt wurden.
Im Herbst 2005 umfasste die PND 2,6 Millionen Personennamensätze.
Neben den Titeldaten der Deutschen Nationalbiografie wurde auch der Titelbestand des Bibliotheksverbunds Bayern (BVB) im VIAF-Projekt herangezogen, in dessen Datensätzen ebenfalls die PND-Normdaten verwendet sind. In beiden Datenbeständen sind insgesamt 15 Millionen bibliografische Datensätze enthalten, die mit PND-Normdatensätzen verknüpft sind.

Das Matching-Problem bei Namen

Zunächst soll der VIAF als deutsch-englisches und englisch-deutsches Wörterbuch für Personennamen dienen. Zum Beispiel kann für einen amerikanischen Benutzer, der nach **J. P. De Valk** sucht – der von der LC angesetzten Namensform -, der Name automatisch in **Johannes P. De Valk** „übersetzt“ werden – in die von der DNB angesetzten Namensform. Wie in diesem Fall, ist es für die verschiedenen internationalen Katalogisierungsinstitutionen durchaus üblich, die Namen unterschiedlich anzusetzen, oder, umgekehrt, dieselbe Namensform zu benutzen, um unterschiedliche Personen zu repräsentieren. Es ist möglich, dass **J. P. De Valk** in der DNB für die Ansetzung eines völlig anderen Autors verwendet würde.

Personennamen können in verschiedenen Namensformen für dieselbe Person auftreten oder mit derselben Namensform unterschiedliche Personen bezeichnen. Das macht es schwer, zusammengehörige Personennamen aus unterschiedlichen Normdateien zuverlässig zusammenzuführen. Die Ausrichtung der beiden Normdateien unterscheidet sich erheblich. Nur ein kleiner Anteil von Personennamen-Ansetzungen stimmen in beiden Dateien vollkommen überein. Aus diesem Grund müssen andere Angaben als nur der Name genutzt werden, um einen zuverlässigen Abgleich zu gewährleisten. In Personennamensätzen sind oft Geburts- und/oder Todesdaten der Person vorhanden. Die kombinierte Angabe beider Daten ist in der Regel ausreichend, um Personen ähnlichen Namens hinreichend zu unterscheiden.

Um dieses Problem beim Matching von Normdatensätzen ohne zusätzliche individualisierende Angaben nochmals zu verdeutlichen, wurde eine Stichprobe mit übereinstimmenden Namensformen aus den LC- und DNB-Normdateien gezogen und intellektuell überprüft, in wie vielen Fällen diese Normdaten-Paare wirklich dieselbe Person repräsentierten. Die Analyse ergab, dass etwa 10% der Namens-Paare verschiedene Personen meinten. Die Fehlerquote wäre

also inakzeptabel hoch, wenn nur die angesetzten Namensformen in das Matching einbezogen würden. Da in den beiden nationalen Normdateien die angesetzten Namensformen auch nicht immer identisch sind, und die zusammen gefundenen Paare zwar namensähnlich, aber nicht miteinander identisch sind, würde die ausschließliche Verwendung von Namen sogar zu einer noch höheren als der festgestellten Fehlerquote führen. Dieser einfache Ansatz schlägt natürlich erst recht fehl, wenn die zahlreichen Namen, die unterschiedlich angesetzt sind, abgeglichen werden sollen.

Die Lösung für das Matching von Namen

Um mögliche Übereinstimmungen zwischen verglichenen Personen(namen) abzusichern oder zurückzuzuweisen, sind zweifelsohne zusätzliche übereinstimmende Angaben notwendig. Nehmen wir zum Beispiel die folgenden individualisierenden Angaben im LC-Personensatz für Diane Glynn:

```
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)
```

Die einzigen unmittelbar für einen Abgleich nutzbaren Daten sind die Namensformen und das Geburtsdatum. In Feld 670 (Quelle) sind zwei Titel angegeben, die unter Umständen maschinell extrahiert werden könnten. In der Praxis sind allerdings nur wenige Titel aus Feld 670 zuverlässig extrahierbar.
Offensichtlich sind aber Titelsätze eine gute Quelle für zusätzliche Angaben zur Person. Titelsätze können nach zusätzlichen Merkmalen zum Werk der Person durchforstet werden, durch die die Person von anderen Personen ähnlichen Namens unterscheidbar sind. Einer der Titelsätze bietet zum Beispiel folgende Angaben:

```
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 –
```

Titelsätze enthalten zwei Typen von Vergleichsmerkmalen. Sie enthalten gewöhnlich Werk-spezifische Merkmale, wie zum Beispiel den Titel, und Manifestations-spezifische Merkmale, wie zum Beispiel die ISBN-Nummer. Eine Übereinstimmung von Titelsätzen ist nahezu gleichbedeutend mit einer Übereinstimmung der darin enthaltenen Personen(namen).

Titelsätze weisen außerdem weitere Merkmale auf, die mehrere Werke der Person betreffen können. Diese Merkmale können dann zum Matching herangezogen werden, wenn keine direkten Titelübereinstimmungen vorliegen. Exemplarisch für diesen Merkmaltyp soll im obigen Beispiel der Mitverfasser Dick O’Connor genannt werden. Dick O’Connor könnte mehr als ein Buch gemeinsam mit Diane Glynn verfasst haben. Deshalb ist die gemeinsame Verfasserschaft im Matching zwischen Normdateien ein starkes Vergleichsmerkmal. Selbst wenn dasselbe Werk in beiden nationalen Titeldatenbanken vorhanden ist - aber in einer der Datenbanken als

Übersetzung – kann ein automatischer Titelabgleich erfolglos bleiben. In solchen Fällen ist der Name eines Mitverfassers das eindeutigere Vergleichsmerkmal zwischen den Titeldatenbanken.

Um die zusätzlichen Vergleichsmerkmale aus Titelsätzen im Abgleichsverfahren auswerten zu können, werden alle verfügbaren Titelsätze, in denen Personennamen als Haupt- oder Nebeneintragung oder als Schlagwort enthalten sind, herangezogen, um Interimsdatensätze – „abgeleitete Normdatensätze“ zu bilden. Diese abgeleiteten Normdatensätze werden daraufhin mit den Original-Normdatensätzen zu “erweiterten Normdatensätzen” kombiniert. Da die erweiterten Normdatensätze Zusatzmerkmale aus Titelsätzen enthalten, können sie einen sehr viel stärkeren Matching-Prozess unterstützen, als die Normdatensätze allein es könnten.

Bestätigung von Namensübereinstimmungen

Will man in zwei nationalen Normdateien dieselbe Person ermitteln, ist es ein sinnvoller Weg, einfach die Namen in beiden Dateien zu vergleichen. Da mit Variationen in der Namensform zu rechnen ist, ist die Chance, dass die gefundenen Personen miteinander identisch sind, allerdings entsprechend geringer. Um die Übereinstimmung in einem automatischen Verfahren abzusichern, geht der vorliegende Ansatz davon aus, dass 1. die Namen miteinander kompatibel sein müssen und dass 2. genügend zusätzliche Vergleichsmerkmale übereinstimmen müssen, um die Übereinstimmung zu bestätigen.

Kompatibilität bedeutet, dass in den Namen keine Abweichungen bestehen, die ausschließen, dass es sich um dieselbe Person handelt. Die Namen können in ihrer Vollständigkeit voneinander abweichen, wie bei John A. Smith und John Allen Smith. Diese Namen sind kompatibel, weil „A.“ für „Allen“ stehen kann. Demgegenüber sind John A. Smith und John B. Smith wegen der sich widersprechenden mittleren Initialen nicht miteinander kompatibel. Sowohl die Ansetzungsformen als auch die Namensvarianten werden beim Testen auf Kompatibilität mit einbezogen.
Ist festgestellt, dass die Namen miteinander kompatibel sind, werden die zusätzlichen Vergleichsmerkmale für die Namen herangezogen, um ihre Übereinstimmung abzusichern.
Die Titeldatenbanken können viele Titel enthalten, die zwar verschieden, aber sehr ähnlich sind, sowie viele zwar verschiedene, aber doch sehr ähnliche Namen. Wenn ein Paar aus Name und Titel in beiden Datenbanken ähnlich vorkommt, kann in jedem Fall davon ausgegangen werden, dass der Name dieselbe Person repräsentiert.
Diese Vorgehensweise wird grundsätzlich auch auf die anderen aus Titeldaten stammenden Vergleichsmerkmale angewendet.
Die Lebensjahre werden getrennt davon als eigene positive Korrelation berücksichtigt. Wenn sie um mehr als ein Jahr voneinander abwichen, wurde keine Übereinstimmung angenommen. Abweichungen um einzelne Jahre wurden allerdings zugelassen. Beim Aufbau des VIAF erwiesen sich kleinere Differenzen in den Lebensjahren als relativ häufig; die zusätzlichen Vergleichsmerkmale reichten aus, um die Übereinstimmung trotz der kleineren Abweichungen in den Lebensjahren zu verifizieren.
Beim Abgleich zweier erweiterter Normdatensätze gilt jedes übereinstimmende Datenelement als Übereinstimmungspunkt (Matching Point). Es werden drei Kategorien unterschieden: starke, mittlere und schwache Übereinstimmungspunkte. Bei kompatiblen Namen gilt ein starker Übereinstimmungspunkt als hinreichend, um abzusichern, dass sie dieselbe Person

repräsentieren. Starke Übereinstimmungspunkte sind Titel, ISBNs, Geburts- und Sterbedaten oder Mitverfasser. Das Geburtsjahr allein wurde nicht als hinreichendes Merkmal zur Unterscheidung von Personen angesehen, es wurde vielmehr als mittlerer Übereinstimmungspunkt behandelt. Als mittlere Überschneidungspunkte wurden Elemente gekennzeichnet, die das Umfeld des Werks der Person beschreiben, wie zum Beispiel die Verlage, in denen die Person publizierte, die im Werk behandelten Sachgebiete oder die Funktion der Person (z.B. Illustrator oder Komponist). Ein großer Verlag wird die Werke vieler Autoren veröffentlichen, von denen zumindest einige ähnliche Namen haben können. Nur die Übereinstimmung vieler mittlerer Übereinstimmungspunkte reicht aus, die Übereinstimmung zu bestätigen.
Schwache Übereinstimmungspunkte werden nur herangezogen, um in unklaren Fällen zur Unterscheidung zu dienen. Beispiele für solche schwachen Übereinstimmungspunkte sind Sprache, Sachgebiet und Erscheinungsland.
Um vorliegende Übereinstimmungspunkte gemeinsam bewerten zu können, wurde jedem eine Zahlenwert zugeordnet. Bei einer Nummer wie der ISBN gibt es entweder eine exakte oder keine Übereinstimmung; daher ist bei einer Übereinstimmung die Punktzahl 1, und 0 bei keiner Übereinstimmung. Für Textelemente wie den Titel wird die Punktzahl abhängig von der Ähnlichkeit des Texts bestimmt, mit einem Wert zwischen 0 und 1. Für die Ähnlichkeitsbestimmung wird dabei ein Trigram-basierter Algorithmus verwendet. Die einzelnen Zahlenwerte werden nochmals nach der Stärke der Übereinstimmungspunkte gewichtet (stark, mittel, schwach) und aufsummiert. Wenn die Gesamtpunktzahl den im Testverfahren festgelegten Schwellenwert übersteigt, wird Übereinstimmung festgestellt. Für den aktuell verwendeten Matching-Algorithmus wurden die Zahlenwerte für die einzelnen Kategorien über umfangreiche Datentests austariert. Es steht zu erwarten, dass weitere Anpassungen vorgenommen werden, wenn weitere Normdateien hinzukommen und Erfahrungen gewonnen werden.

Bildung erweiterter Normdatensätze

Die oben beschriebenen Verfahren setzten erweiterte Normdatensätze sowohl für LCNA- als auch für PND-Sätze voraus. Aus den LC-Titeldaten wurden abgeleitete Normdatensätze aufbereitet (vgl. oben), um damit die LCNA-Datensätze zu erweitern, und die Titeldaten der DNB und des BVB wurden zur Erweiterung der PND-Datensätze genutzt.

Für den erweiterten LCNAF konnten 3,8 von 4,2 Millionen Normdatensätzen erweitert werden. Davon wurden lediglich 2,6 Millionen (60 %) mit Datenelementen aus Titeldaten erweitert (aus insgesamt 7,4 Millionen Titelsätzen). Weitere Erweiterungen stammen aus 4,1 Millionen Titeln, die aus den Feldern 670 (Quelle) der Normdatensätze extrahiert werden konnten. Der Titel – wie auch aus der Zusammenfassung der Ergebnisse hervorgeht - ist das wichtigste Erweiterungselement zum Herstellen von Übereinstimmungen.

Für die erweiterte PND konnten 2,4 Millionen von 2,6 Millionen (90 %) der Normdatensätze mit Erweiterungen versehen werden, allerdings auch hier nur 2,0 Millionen (80 %) aus Titelsätzen. Die restlichen 400.000 Datensätze wurden aus Titeln erweitert, die den PND-Sätzen selbst entnommen wurden.

Test der Matching-Verfahren

Die VIAF-Teilnehmer unterstützten den Matching-Prozess durch Genauigkeitsprüfungen und Kommentierungen der Ergebnisse. Zum Beispiel waren Serientitel anfänglich als Übereinstimmungspunkte behandelt, führten in den Tests aber häufig zu falschen Ergebnissen. Jede der Prüfungen führte zu Anpassungen, die entweder die Anzahl der gefundenen Übereinstimmungen erhöhte oder die Zahl der Fehler reduzierte.
In dieser Zeit wurden ein hinreichend genauer Schwellenwert und Bewertungsalgorithmus entwickelt. Hierzu können nur die abschließenden Absicherungstests beschrieben werden.

Um die Genauigkeit und Wirksamkeit des Matching-Prozesses abzusichern, wurden Stichproben übereinstimmender Namen durch erfahrene Normdaten-Bearbeiter aus LC und DNB überprüft. Die erste Stichprobe hatte die beiden Zielsetzungen, die Überschneidungsrate zwischen den Normdateien zu bestimmen sowie herauszufinden, welcher Anteil davon im Matching-Prozess identifiziert werden kann. Die zweite Stichprobe diente dazu, systematische Fehler und Defizite aufzuspüren, um sie zu bereinigen, sowie die zu erwartende Gesamt-Fehlerquote zu schätzen.

Die erste Stichprobe enthielt 391 zufällig ausgewählte PND-Sätze. Im LC Authority File wurden – automatisch und intellektuell - mögliche Übereinstimmungen gesucht.
Im automatischen Verfahren wurden die PND-Sätze zu allen LC-Datensätzen abgeglichen, die denselben Nachnamen haben, was zu 74.000 automatisch zusammengeführten Namenpaaren führte, die weiter zu untersuchen waren.
Der Matching-Algorithmus wurde auf diese 74.000 Namenpaare angewandt, und 79 PND/LC-Normdatensatz-Paare automatisch als übereinstimmend erkannt.

Intellektuelle Überprüfungen der 391 PND-Sätze ergaben weitere 35 Namen mit korrespondierenden LCNA-Sätzen; bei diesen Datensatz-Paaren beruhte aber entweder die Übereinstimmung nicht auf Gleichheit der Familiennamen, oder der Matching-Algorithmus konnte die Übereinstimmung nicht bestätigen. Die 79 automatisch gefundenen Übereinstimmungen wurden durch die intellektuelle Überprüfung als zutreffend bestätigt. Ausgehend von der PND-Stichprobe lässt sich schätzen, dass rund 30 % der PND-Personen Entsprechungen in dem LCNA File haben, und dass hiervon etwa 70 % automatisch durch den Algorithmus als übereinstimmend erkannt werden können. Das lässt sich auf geschätzte 800.000 Datensätze hochrechnen, die in beiden Normdateien übereinstimmen. Davon werden der Schätzung nach 550.000 Datensätze in dem automatischen Matching-Prozess identifiziert werden können.

Die Ergebnisse wurden gleichzeitig auch überprüft, um die Vorgehensweise bei der Zusammenführung der Namenpaare zu verbessern.
Werden dabei nur Nachnamen genutzt, würden für jedes übereinstimmende Paar etwa 1000 Namenpaare dem vollen Matching-Prozess unterzogen werden müssen. Die intellektuellen Matching-Test-Ergebnisse dienten auch dazu, festzustellen, dass für die Zusammenführung der Namenspaare eine Auswahl, basierend auf Nachnamen, Vornamen sowie in bestimmtem Unfang Jahresangaben, genutzt werden sollte, die bereits als grober Annäherungswert an Namenskompatibilität gelten kann. Dieser einfache Index erwies sich als so gut gewählt, dass in 95 % der Fälle pro Match nur vier Namenspaare überprüft werden mussten. Der Index ist

gleichzeitig effizient und effektiv, und kleinere Anpassungen dürften zukünftig noch zu weiteren Verbesserungen führen.

Die Zielsetzung der zweiten Stichprobe lag in der Schätzung der Fehlerquote beim Abgleich. Als Teil des Prozesses überprüfte die Stichprobe die Adaequatheit des vorläufig festgelegten Schwellenwerts für Übereinstimmungen und passte ihn, soweit nötig an.
Bei der Verwendung eines Schwellenwerts ist davon auszugehen, dass die Fehlerquote bei Übereinstimmungen mit einem Punktwert nahe am Schwellenwert größer ist als bei Übereinstimmungen mit einem Punktwert weit über dem Schwellenwert. Die meisten übereinstimmenden Normdatensätze haben Werte, die weit über dem Schwellenwert liegen. Um die bestmögliche Fehlerquote mit dem geringst möglichen Anteil an intellektuellem Arbeitsaufwand zu ermitteln, wurde die Stichprobe basierend auf den Punktwertergebnisse in vier Unterstichproben geteilt. Über intellektuelle Prüfungen wurden die Fehlzuordnungen festgestellt und für jede der Unterstichproben Fehlerquote und Konfidenz bestimmt. Die Teilergebnisse wurden gewichtet und summiert, um die Gesamt-Fehlerquote für das Matching-Verfahren zu festzustellen. Es ergab sich eine Quote von weniger als ein Prozent.

Eine der Unterstichproben reichte in ihren Bewertungszahlen bis nahe an den Schwellenwert heran. Eine Absenkung des Schwellenwerts würde eine zusätzliche Fehlzuordnung pro jeweils drei richtigen Zuordnungen ergeben. Dennoch ist die Absenkung des Schwellenwerts nicht zulässig. In dem Wertebereich unmittelbar über dem Schwellenwert war unter 25 richtigen Zuordnungen nur eine falsche. Da innerhalb dieses Werteinterwalls nur relativ wenige Zuordnungen anfielen, ist die Auswirkung auf die Gesamt-Fehlerquote niedrig und eine große Zahl der validen Zuordnungen bleibt erhalten. Daher wurde der vorläufige Schwellenwert akzeptiert.

Aufbau des Basis-VIAF

Die erweiterten Normdateien wurden von beiden Seiten mit dem Matching-Algorithmus bearbeitet, und die bearbeiteten Datensätze – sowohl übereinstimmende als auch nicht-übereinstimmende – in VIAF-Datensätze konvertiert. Der Vorgang ist in Abbildung 2 dargestellt. Es ergaben sich 6,3 Millionen Datensätze in der VIAF-Datenbank, die sich zusammensetzen aus 500.000 verknüpften Datensätzen, 3,7 Millionen nicht zugeordneten Datensätzen aus dem LC Authority File und 2,1 Millionen nicht zugeordneten Datensätzen aus der PND. Das Ergebnis liegt nahe bei der Schätzung aus der intellektuellen Überprüfung (vgl. oben). Es wird geschätzt, dass etwa 250.000 zusätzliche Normdatensatz-Paare für dieselbe Person existieren, die aufgrund fehlender auswertbarer Daten nicht automatisch zugeordnet werden konnten. Im endgültigen Verfahren wird für solche und andere intellektuell festgestellte Übereinstimmungen eine manuelle Verknüpfung möglich sein. Die Normdatensätze werden eine sequentiell vergebene VIAF-Datensatznummer erhalten.

Abbildung 3 zeigt ein Beispiel für einen VIAF-Satz im MARC 21-Format. Da der Hauptzweck des VIAF in der Verbindung der Normdateien besteht, enthält der VIAF-Satz – jeweils mit einer Herkunftsangabe – für jeden Namen eine Eintragung in einem Feld 700 (Alternative Ansetzungsform). Da keiner der Namen als gemeinsame Ansetzungsform festgelegt ist, ist Feld 100 (Personenname in Ansetzungsform) nicht besetzt. Bei durch den Algorithmus bestätigten

Übereinstimmungen sind zwei alternative Ansetzungsformen im Datensatz angegeben, wenn keine Übereinstimmung vorliegt, ist nur ein einziges 700er Feld besetzt.

Die zusätzlichen Individualisierungsmerkmale sind in erweiterten Normdatensätzen in lokalen Feldern (9xx) ebenfalls angegeben. Um den Abgleich zu vereinfachen, wurden Textelemente durchgehend normalisiert. Dabei wurde eine leicht abgewandelte Version der Normalisierungsregeln der NACO (Name Authority Cooperative Program of the Program for Cooperative Cataloging) verwendet [7]. Die Vorkommenshäufigkeit bestimmter Wörter ist im Unterfeld $9 festgehalten. Da die Angabe hauptsächlich für maschinelle Bearbeitungsvorgänge gedacht ist, wird sie in Präsentationen für End-Nutzer nicht notwendigerweise enthalten sein. Sobald weitere Normdateien hinzugefügt werden, werden diese zunächst mit den bestehenden erweiterten VIAF-Sätzen verglichen. Werden übereinstimmende Datensätze ermittelt, werden die entsprechenden Links in den VIAF-Satz übernommen. Die zusätzlichen Angaben aus den übereinstimmenden Datensätzen werden ebenfalls in den VIAF-Satz eingespielt.

In einer signifikanten Anzahl von Fällen traf eine Ansetzungsform aus der einen Normdatei auf mehrere Ansetzungsformen in der anderen Normdatei. Da die Hauptaufgabe des VIAF in einer 1:1-Umsetzung der Normdatensätze besteht, wurden 1:n-Zuordnungen grundsätzlich nicht bestätigt, und 70.000 algorithmische Zuordnungen wurden aufgrund von Mehrfachzuordnungen eliminiert. Mindestens zwei Ursachen für Mehrfachzuordnungen ließen sich bestimmen.
Erstens gibt es in der PND eine größere Anzahl nicht differenzierter Personennamen, die jeweils mit zwei oder mehreren differenzierten Namen im LCNAF übereinstimmen. Gemäß den deutschen Katalogisierungsregeln RAK-WB wurden Personen gleichen Namens nicht unterschieden. Mit dem Übergang zur Katalogisierung mit einer Normdatei rückte die Deutsche Nationalbibliothek von dieser Praxis ab und begann, Personensätze zu individualisieren, und mittlerweile hat die Individualisierung auch Eingang in die Katalogisierungspraxis der Bibliotheksverbünde im deutschsprachigen Raum gefunden. Trotzdem enthält die PND nach wie vor viele nicht differenzierte Namen. Die Deutsche Nationalbibliothek strebt an, die Personennamen mit mehrfachen Zuordnungen aus dem VIAF-Projekt auf der Basis der Übereinstimmungen zwischen LC- und DNB-Titeldaten zu individualisieren, soweit wie möglich in einem automatischen Abgleich, den Rest intellektuell. Die Korrekturen werden als Teil der laufenden Updates in den VIAF gelangen, und damit werden eindeutige Links zwischen den übereinstimmenden Datensätzen hergestellt werden.
Zweitens gibt es im LCNAF eine Anzahl von Datensätzen, die den AACR-Bestimmungen folgend für jede bibliografische Identität einer Person, z.B. für jedes Pseudonym, einen separaten Datensatz bieten. Umgekehrt als im Falle der nicht-individualisierten PND-Namen, werden hier für ein und dieselbe Person mehrere Normdatensätze gebildet. Die PND hat demgegenüber gemäß den RAK-WB nur einen Normdatensatz, in dem wirklicher Namen und Pseudonyme vereinigt sind. Wie die nicht-individualisierten Namen bieten auch diese “überdifferenzierten” Namen Probleme, für die keine völlig befriedigende Lösung gefunden werden konnte.

Die miteinander verlinkten Namen können direkt zur automatischen Übersetzung von LC-Namen zu PND-Namen verwendet werden und umgekehrt. Damit werden die Anforderungen des Semantic Web oder übergreifender Suchsysteme unterstützt, die auf dieses Feature angewiesen sind. Die Beibehaltung der Anzeige von Namensvarianten kann den menschlichen Benutzern zusätzliche Informationen vermitteln.

Die Normdatennummern der beteiligten Normdateien oder die VIAF-Nummern selbst können auch als Basis von URIs genutzt werden. Dies würde Möglichkeiten für einen Normdaten-Resolvingdienst eröffnen. Ausgehend von irgendeiner als Zitat verwendeten URI in einem Dokument, einem Datensatz oder einer WebSite würde der Benutzer zu allen Materialien, Datensätzen, Ressourcen etc. geführt, mit denen die Normdaten, die in der URI gemeinsam repräsentiert sind, in Beziehung stehen – natürlich auch zu den beteiligten Normdatensätzen selbst.

Laufendes Verfahren

Die nationalen Normdateien und Titeldatenbanken werden ständig geändert. Eine Datenbank mit Links zwischen zwei oder mehr der sich ändernden Normdateien muss dementsprechend fortlaufend überprüft und upgedatet werden. Die Logik und die Software des für den Basis-VIAF entwickelten Verfahrens muss deshalb so angepasst werden, dass ein kontinuierliches Update der Datensätze gewährleistet wird. Sobald neue Titel- oder Normdatensätze eingehen, müssen die erweiterten Normdatensätze angepasst und auch der Abgleich zwischen den Normdateien erneut ausgewertet werden. Neue Übereinstimmungen werden fortlaufend hinzukommen, andere Übereinstimmungen können aufgrund von Änderungen in der Quellenlage aufgelöst werden. Wenn Links aufgelöst werden, wird der nicht mehr zutreffende Link als Referenzierung in jedem verbundenen Datensatz erhalten bleiben.

Längerfristig soll das VIAF-Updateverfahren über OAI abgewickelt werden, zwischenzeitlich werden innerhalb des Projekts traditionellere Datenaustauschverfahren wie FTP zum Einsatz kommen.

Mit diesem großen, an einem Standort konzentrierten Datenaufkommen können unterschiedlichste Methoden und Verfahren zum Datenzugriff und –organisation getestet werden. Die Links können – als Teil des Semantic Web - dazu genutzt werden, den Personennamen für den End-Nutzer in das von ihm gewünschte Format zu übersetzen. Tools können entwickelt werden, mit denen in verschiedenen Datenbanken eine automatische Suche mit den jeweils zutreffenden Namensformen ermöglicht wird, und ähnliche Tools, die die Katalogisierung und die Normdatenkontrolle unterstützen, indem automatisch die für den jeweiligen Datensatz zutreffende Namensform bestimmt wird. Selbstverständlich wird die VIAF-Datenbank auch direkt recherchierbar sein.

Ergebnisse

Die PND hat bereits beachtliche Vorteile aus dem Projekt gezogen. Aufgrund der Dubletten-Selbsttests, die in beiden Normdateien durchgeführt wurden, wurden bereits umfangreiche Upgrades durchgeführt; die DNB erhofft sich außerdem wesentliche Unterstützung bei der Individualisierung nicht-individualisierter Namen durch die Titelangaben in den erweiterten Normdatensätzen. Die Abgleichsverfahren und –algorithmen sind auf viele andere Anwendungsfälle übertragbar. Möglichkeiten für Dienste werden erforscht, in denen die Links zwischen Personennamen genutzt werden, um den Zugang zu bibliografischen Daten zu verbessern und weitere Katalogisierungsvorhaben der Teilnehmer zu unterstützen.

Das Projekt hat nachgewiesen, dass die automatische Verlinkung von Personennamensätzen unterschiedlicher Normdateien durchführbar ist. Siebzig Prozent der in beiden Normdateien übereinstimmend vorhandenen Personen wurden mit einer Fehlerquote unter einem Prozent miteinander verlinkt. Das Verfahren, die Normdatensätze mit zusätzlichen Elementen aus Titelsätzen anzureichern, verbesserte die Übereinstimmungsquote erheblich und reduzierte gleichzeitig die Anzahl der falschen Zuordnungen. Kleinere Änderungen in den Normdatensätzen wurde die Zuordnung zudem noch erheblich verbessern. Vielfach missglückte der Abgleich, weil Feld 670 (Quelle) nicht korrekt geparst werden konnte. Eine verbesserte Strukturierung des Texts, bei der verkürzte Namen und Titel vermieden werden, oder explizite Links zu dem als Quelle angegebenen Titelsatz wären sehr hilfreich. Auch die explizite Angabe der üblichen Rollen oder Gebiete (Komponist, Illustrator, Mathematiker, etc.) würde den Abgleich automatisch und intellektuell - weiter verbessern, ebenso wie die Einbeziehung vollständigerer Namensformen, zumindest als Verweisungen.
Die Studie bietet ein überzeugendes Plädoyer für Normdatenkontrolle, für für Normdatennutzung, für Vernetzung und gegenseitige Verlinkung – und dafür ein semantisches Netz für Bibliotheken aufzubauen.
Für die Bibliotheken und Bibliotheksverbünde im deutschsprachigen Raum, die über Titelsätze mit LCNA-Sucheinstiegen verfügen, kann der VIAF als Plattform dienen, von der einen in die andere Normdatei umzusteigen, entweder um die LCNA-Sucheinstiege durch PND-Sucheinstiege zu überschreiben, oder um über den VIAF die durchgehende Suche mit PND-Formen zu ermöglichen. In multinationalen und multilingualen Portalen wie z.B. dem Portal von TEL (The European Library) könnten über den VIAF automatisch kombinierte Suchfragen in beiden Normdateien gestellt und der Benutzer zu den Titeldaten aus beiden Quellen geführt werden.

Mit den vorhandenen Abgleichsverfahren soll ein Update-fähiges System entwickelt werden, in dem die laufenden Updates aus den Normdaten- und Titeldatenbanken der Teilnehmer eingesammelt werden. Hierfür sollen OAI-Verfahren genutzt werden. Das System ist skalierbar und dafür ausgelegt, neue Teilnehmer, die Normdaten gemeinsam nutzen wollen, aufzunehmen. Etwaige Grenzen für die Skalierbarkeit des VIAF können erst herausgefunden werden, wenn weitere Institutionen zu dem Projekt hinzugestoßen sind.

Das VIAF-Projekt hat sich bisher hauptsächlich mit dem Problem beschäftigt, Normdatensätze gegeneinander abzugleichen. Um den VIAF zu führen, zu erweitern und in Dienste einzuführen, sind seine langfristige Pflege und eine erfolgreiche Führungsstrategie notwendig. Entscheidungen stehen an. Außerdem soll die Leistungsfähigkeit des Systems gesteigert warden, um den Unicode-Zeichensatz bedienen zu können. Unicode wird die Einbeziehung nicht-lateinischer Schriften ermöglichen, aber auch den Matching-Algorithmus darauf auszudehnen, wird eine Herausforderung sein, insbesondere für ideographisch geprägte Schriften wie Koreanisch, Chinesisch und Japanisch.

Referenzen

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]

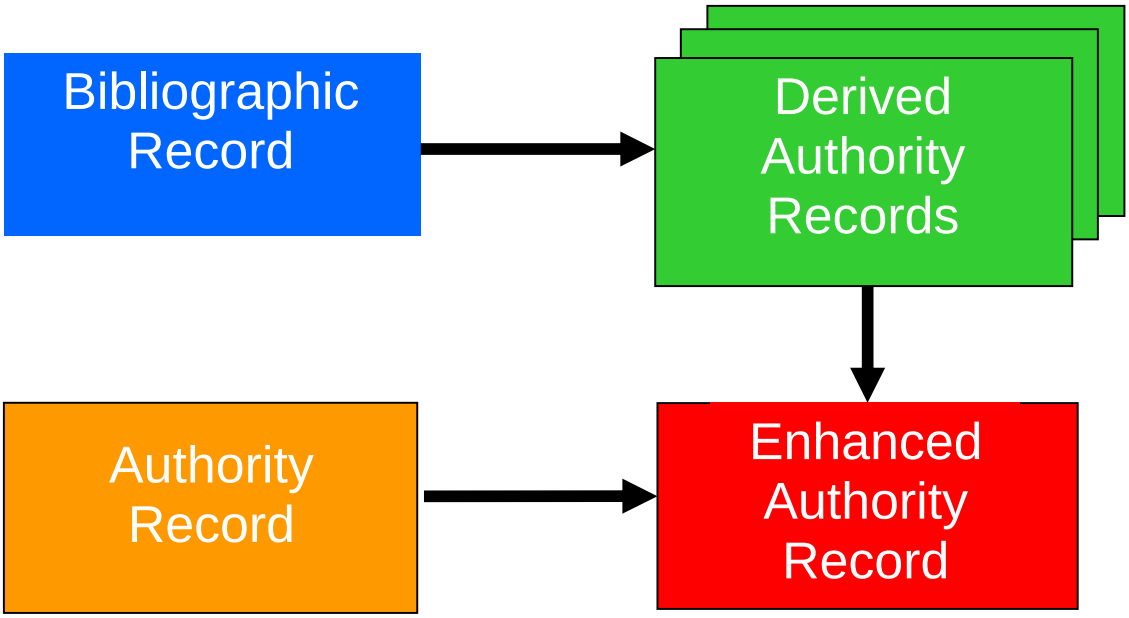

Figure 1. Creating the Enhanced Authority Record

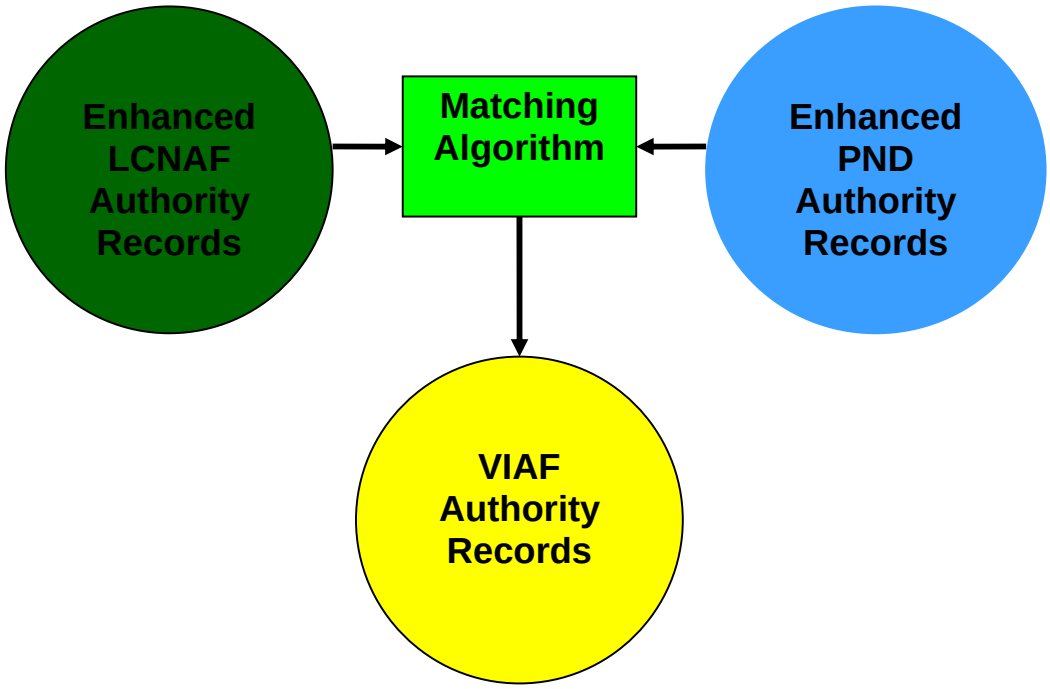

Figure 2.

Creating the VIAF Authority Records.

```
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
Enhanced Record Formats

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
		Authority Record	Extracted from Name/Title authority records, field 100
	917	Uniform Title	$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