VIAF (Виртуальный Международный Авторитетный Файл): Авторитетные файлы имен, связывающие Немецкую Библиотеку и Библиотеку Конгресса

Рик Беннетт
OCLC Онлайновый Компьютерный Библиотечный Центр
Дублин, Огайо, США

Кристина Хенгель-Диттрих
Немецкая библиотека
Франкфурт-на-Майне, Германия

Эдвард. Т. О'Нейлл
OCLC Он-лайновый компьтерный библиотечный центр
Дублин, Огайо, США

Барбара Б. Тиллетт
Библиотека Конгресса
Вашингтон, О.К. США

Заседание:	**123 Каталогизация**
Синхронный перевод:	**Да**

ВСЕМИРНЫЙ БИБЛИОТЕЧНЫЙ И ИНФОРМАЦИОНННЫЙ КОНГРЕСС: 72-я ГЕНЕРАЛЬНАЯ КОНФЕРЕНЦИЯ И СОВЕТ
20-24 Августа 2006 г., Сеул, Корея
http://www.ifla.org/IV/ifla72/index.htm

Резюме

Немецкая библиотека, Библиотека Конгресса и OCLC Онлайновый Компьютерный Библиотечный Центр совместно разрабатывают виртуальный международный авторитетный файл (VIAF) имен лиц, который связывает авторитетные записи этих мировых национальных библиографирующих агентств и будет представлен в свободный доступ через Web. Целью проекта является проверка возможности автоматически связывать авторитетные записи из различных национальных авторитетных файлов и наглядная демонстрация его полезности. Для создания первоначального VIAF, который включает более шести миллионов имен с более чем половиной миллиона связей, были использованы авторитетные и библиографические файлы Библиотеки Конгресса и Немецкой библиотеки. Ключевой задачей проекта была разработка алгоритмов автоматического подбора пар имен на основе сведений как из авторитетных, так и из библиографических записей. Была доказана целесообразность алгоритмической установки связи между именами лиц из разных национальных авторитетных файлов; семьдесят процентов автритетных записей имен, общих для обоих файлов, были связаны автоматически с долей ошибок менее чем один процент. Долгосрочной целью проекта VIAF является объединение авторитетных имен из множества национальных библиотек и иных важных источников в совместную глобальную службу авторитетного контроля.

Введение

Несколько групп Секции по каталогизации Международной федерации библиотечных ассоциаций и учреждений (ИФЛА) одобрили потенциальную возможность виртуального авторитетного файла (VIAF) [1], в котором авторитетные записи из мировых национальных библиографирующих агентств, представляющие одну и ту же сущность, были бы связаны и доступны в Интернет. Подобный VIAF мог бы стать практическим развитием идеи универсального библиографического контроля и включит результаты работы каждого национального библиографирующего агентства. Это позволит различным национальным или региональным вариантам сосуществовать в авторитетной форме, что будет соответствовать всемирным пользовательским потребностям, различающимся по предпочтительности языка, графики и орфографии.

В настоящее время будущее развитие Web предполагает использование онтологических возможностей для превращения Web в инструмент, более понятный для машинной и автоматической обработки. Если VIAF объединится с другими контролируемыми словарями и авторитетными файлами таких источников как реферирующие и индексирующие службы, архивы, музеи, издательства и т.п., он может стать одним из основных блоков в фундаменте «семантического Web» [2]. Библиотеки имеют сейчас возможность внести значительный вклад в это будущее и помочь превратить эту мечту в реальность. Для развития этого направления важно, чтобы пользователям во всем мире был предоставлен свободный доступ к VIAF.

Другие проекты, направленные на связи имен в авторитетных файлах. В проекте LEAF (Связь и Исследование Авторитетных Файлов) предлагалось связать авторитетные записи из множества различных источников, включая библиотеки, архивы, документационные и исследовательские центры. Эти записи представлены в различных форматах, и особенности их типов и содержания имеют значительные различия. В проекте LEAF предполагалось автоматически связать записи при их загрузке в систему. Из-за различий источников авторитеных записей имен было признано, что только имя, включающее ссылки «смотри» и связанные с ним даты, является той общей информацией, которая позволяет установить связи. Поскольку авторитетные записи участников в настоящее время не включают даты, ожидается, что процент ошибок при подборе пар авторитетных записей имен должен быть неприемлемо высок.

Проект InterParty [4] финансируется ЕС и является показательным проектом создания связи между авторитетными файлами разных организаций с приоритетной задачей поддержки управления процессами правовой реабилитации (supporting digital rights management). Предложенная система InterParty должна была обеспечить единую точку доступа ко множеству баз данных, включенных в систему, поэтому ее главной задачей была функция централизованного поиска. Поскольку связи между именами в каждой из баз данных идентифицируются вручную, человек, выявляющий связь, может ее фиксировать в записях. В дальнейшем эти связи могут использоваться автоматически. В зависимости от организации процесса формирования связей, такие связи могут быть достаточно правильными. Введение связи одной стороной не требует ее согласования с другими участниками системы. Проект учитывает возможность алгоритмического подбора пар, но не определяет технические приемы или требавания к данным, необходимые для поддержки возможности формирования связей.

Проект VIAF

Во время Всемирного библиотечного и информационного конгресса ИФЛА 2003 г. в Берлине Немецкая библиотека (DDB), Библиотека Конгресса (LC) и OCLC Онлайновый компьютерный библиотечный центр договорились создать Виртуальный международный авторитетный файл (VIAF) для имен лиц [5]. Целью проекта VIAF является обеспечение возможности автоматического формирования связей между авторитетными записями из различных национальных авторитетных файлов и наглядно показать пользу VIAF. Проект VIAF будет связывать авторитетные файлы имен Библиотеки Конгресса и Немецкой Библиотеки в единый виртуальный авторитетный файл имен. OCLC разрабатывает программное обеспечение для подбора пар авторитетных записей имен в двух авторитетных файлах. Стратегической целью проекта VIAF является задача связывания авторитетных имен из множества национальных библиотек и других авторитетных источников в распределенную глобальную авторитетную службу имен/наименований лиц, организаций, конференций, населенных пунктов и т.п.

Проект VIAF включает пять этапов:

1. Формирование «Расширенных Авторитетных» записей как из Personennormdatei (PND), так и из Авторитетных записей LC. Этот этап включает выявление соответствующих авторитетных записей для включения их в расширенные авторитетные записи и установление потребности в каких-либо определенных способах решения проблем для входящих файлов.
2. Разработка алгоритмов подбора пар расширенных авторитетных записей PND и LC для формирования первоначальной версии VIAF. На этом этапе повторяются процессы первого этапа, поскольку промежуточные результаты подбора пар выдвинули на первый план потребность в дополнительной информации, которая должна быть найдена и включена в расширенные авторитетные записи для совершенствования процесса подбора парных записей.
3. Построение сервера типа Open Archive Initiative (OAI) [6] для обеспечения доступа к VIAF.
4. Для поддержки базы данных VIAF требуются дополнения и изменения авторитетных и библиографических записей всех участвующих агентств. Это обновление данных и поддержка системы будут строиться на протоколах, используемых в OAI по запросу этих сведений для обновления данных.
5. Для доступа к записям VIAF будет сделан пользовательский интерфейс доступа в открытом Web. Со временем база данных и интерфейс будут поддерживать Unicode и различные шрифты, станут многоязычными. При выполнении прямых запросов в базу данных, обеспеченную, например, версией имен Библиотеки Конгресса, с требованием поиска совпадений с версией PND, в качестве простой связи для поддержки семантических возможностей Web может использоваться HTML.

Изначально проект фокусируется на показе возможности установить в VIAF связь между авторитетными записями имен Personennormdatei (PND) и Авторитетным файлом имен Библиотеки Конгресса (LCNAF). К 31 декабря 2005 г. файл LCNAF включал 4,2 млн. авторитетных записей для имен лиц. К тому времени Библиотека Конгресса создала и распространила суммарно 9,3 млн. библиографических записей.

К концу 2005 г. файл PND содержал 2,6 млн. авторитетных записей для имен лиц. Авторитетный файл PND используется в библиографических записях Немецкой Библиотеки

(DDB) и в библиогрпфических записях Bibliotheksverbund Bayern (BVB). В этих двух библиографических файлах представлено суммарно 15 млн. библиографических записей, связанных с авторитетными записями PND.

Проблема подбора парных имен

Изначально VIAF будет функционировать в качестве немецко-английского и англо-немецкого словаря для имен лиц. Например, если американский пользователь ищет **J. P. De Valk** (форма имени, установленная в LC), имя может автоматически «толковаться» как **Johannes P. De Valk** (форма, установленная в DDB). Как и в данном случае, различные библиографирующие учреждения часто определяют форму имени одного лица по-разному, либо, наоборот, используют одну форму имени для представления разных авторов. Вполне возможно, что форма **J. P. De Valk** могла быть установлена в DDB для совершенно иного автора.

Для одного и того же лица могут использоваться разные формы имени, либо одна и та же форма имени может использоваться разными людьми, что создает трудности для достоверного подбора пар имен из различных авторитетных файлов. В составе обоих авторитетных файлов имеются значительные расхождения; только небольшая доля имен лиц представлена в обоих файлах. Поэтому для обеспечения достоверности подбора пар имен необходимо использовать не только собственно имя, но и другие сведения. В авторитетных записях для имен лиц часто используются даты рождения и/или смерти лица. Сочетаний имени с датами жизни обычно достаточно для различения людей с одинаковыми именами.

Для решения этой проблемы при подборе пар авторитетных записей из авторитетных файлов LC и DDB были выбраны общие имена без дополнительных сведений. Эти пары автортетных записей были просмотрены вручную для того, чтобы удостовериться, что они представляют одно и то же лицо. Эта проверка показала, что около 10% пар имен относились к разным людям. Таким образом, доля ошибок при выявлении совпадений только по установленной форме имени неприемлемо высока. Поскольку в двух национальных авторитетных файлах формы имени не всегда идентичны, подбираются пары похожих, но не идентичных имен, что приводит к более высокму проценту ошибок. Этот простой подход также ведет к неудачам при подборе пар среди многочисленных имен, для которых устанавливаются различные авторитетные формы.

Решение задачи подбора пар имен

Ясно, что дополнительные сведения при подборе пар необходимы для подтверждения либо отказа от потенциально совпадающих пар. Например, рассмотрим следующие авторитетные данные LC для 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)
```

Единственными непосредственно используемыми данными являются имена и даты рождения. В поле 670 (Источник сведений) включены два заглавия, которые могли быть выбраны при машинной обработке. На практике только некоторые заглавия могут быть достоверно получены из этих полей.

Библиографические записи, несомненно, являются источником дополнительной информации о конкретном лице. Такие библиографические записи могут быть источником для

дополнительных сведений о произведении данного лица, которые помогают различать двух разных авторов с одинаковыми именами. В одной библиографической записи указано:

```
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 –
```

Библиографические записи включают два типа дополнительных сведений. Обычно библиографические записи включают информацию о конкретном произведении, например, заглавие, и информацию о его конкретном воплощении, например, номер ISBN. Совпадающее заглавие обеспечивает почти точное совпадение при подборе именых пар.

В библиографической записи имеется также дополнительная информация, которая может касаться множества произведений данного автора. Такая информация может помочь при подборе пар авторов, если точное совпадение заглавия не применимо. Соавтор Dick O'Connor является примером такого типа сведений. Dick O'Connor может быть соавтором более чем одной книги Diane Glynn, что является веским фактом при подборе в авторитетных файлах совпадающей пары имени. Даже если одно и то же произведение представлено в обоих национальных базах данных, но в одной из них представлен перевод этого произведения, подбор парного заглавия будет трудно произвести автоматически. В этом случае имя соавтора является более подходящим и простым способом подтверждения правильности подобора пары в базах данных.

Все имеющиеся библиографические записи, в которые включено данное имя в качестве основного или дополнительного заголовков или как предмет трансформируются для создания промежуточной записи, названной «производной авторитетной». Затем эти производные авторитетные записи соединяются с оригинальной авторитетной записью для создания расширенной авторитетной записи. Поскольку расширенные авторитетные записи включают дополнительную информацию из библиографических записей, связанную с именем, они могут способствовать более точному подбору пар, чем собственно авторитетные записи.

Подтверждение правильности подбора именных пар

Простое сравнение имен в двух национальных авторитетных файлах является приемлемым способом нахождения одного и того же человека. Ожидаемые различия в форме имени снижают вероятность того, что это будет одно и то же лицо. Для автоматического подтверждения того, что подобранные пары имен относятся к данным лицам, в нашем случае (1) имена должны быть совместимы (похожи) и (2) должна присутствовать достаточная дополнительная информация, подтверждающая идентичность пары.

Совместимость требует, чтобы не было расхождений между именами, представляющими одно и то же лицо. Имена могут отличаться по полноте, как, например, John A. Smith и John Allen Smith. Эти имена совместимы, потому что 'A' может быть Allen. Однако, John A. Smith и John B. Smith не совместимы из-за различий второго инициала. И авторитетная форма имени и вариантные формы учитываются при проверке на совместимость.

Если имена уже определены как совместимые, дополнительные сведения о них собираются для подтверждения правильности подбора пары. Библиографические файлы могут включать множество разных, но похожих заглавий и множество разных, но похожих имен. Однако, если пара имя/заглавие в обоих файлах совпадает, весьма вероятно, что имя представляет одного и того же человека. Эта базовая стратегия распространяется и на другие типы сведений, получаемых из библиографических записей.

Даты для точного соотнесения пар рассматриваются отдельно. Если даты отличаются более чем на год, имена признаются несовместимыми и подобранные пары отвергаются. Допуски в датах приняты для разницы в один год. При разработке VIAF было относительно просто найти небольшие расхождения в датах, а дополнительной информации для подбора пар имелось достаточно, чтобы подтвердить совпадение даже при небольших различиях в датах.

При сравнении двух расширенных авторитетных записей каждый парный элемент считается парной точкой совпадения. Парные точки совпадения делятся на три категории: сильные, средние и слабые. Для совместимых имен наличие сильной точки совпадения считается достаточным подтверждением того, что они относятся к однму и тому же лицу. К сильным точкам совпадения пары относятся заглавия, ISBN, даты рождения и смерти или соавторы. Одна только дата рождения признана недостаточной для различения имен и более правильно считать ее средней точкой совпадения. К средним точкам совпадения относятся указания на среду, окружающую произведение, такие как издательства, тематическая область или на роль лица (например, иллюстратор или композитор). Большие издательства публикуют произведения многих авторов и некоторые из них могут иметь одинаковые имена. Совпадение по множеству пар средних точек является достаточным для подтверждения соответствия. Слабые точки соответствия считаются достаточными только для различения пар, подборанных иными способами, но вызывающими сомнения. Примеры таких слабых точек совпадения включают язык, тематическую область и страну публикации.

Для объединения точек подбора пар подсчитывается сумма баллов, определенных для каждой точки совпадения. Цифровой номер, например, ISBN, или точно совпадает, или не совпадает, что в сумму баллов добавляет единицу при совпадении или ноль при несовпадении. Текст, например, заглавие, в зависимости от того, насколько текст сходен, может добавить в сумму от ноля до единицы баллов.

Методика определения трехзначной суммы баллов используется для определения сходимости текста. Отдельные результаты подсчетов модифицируются в зависимости от значимости (сильный, средний, слабый) и суммируются. Если общая сумма баллов превышает порог, определенный при тестировании, правильность подбора пары считается подтвержденной. В существующем алгоритме подбора пар, тестирование множества записей позволило откорректировать подсчет баллов в рамках указанных категорий, и мы предполагаем, что такая корректировка будет продолжена как при подключении к системе новых авторитетных файлов, так и опытным путем.

Формирование расширенных авторитетных записей

Описанная выше методика использовалась при создании расширенных авторитетных записей и для авторитетных записей имен PND, и для авторитетных записей имен LC. Библиографические файлы LC были обрабатаны с целью получения производных авторитетных записей для расширенного авторитетного файла LC, а библиографические файлы DDB и BVB обрабатывались для расширения авторитетных записей PND. В таблице 1 представлена общая схема информационных потоков, образующих расширенные авторитетные записи.

Для расширенного авторитетного файла LC должны были быть расширены 3,8 милиона (90%) авторитетных записей из 4,2 миллиона. Только 2,6 миллиона (60%) были дополнены сведениями из библиографических записей, всего 7,4 миллиона заглавий. Другие дополнения были проведены за счет использования 4,1 миллиона заглавий, извлеченных из 670 полей авторитетных записей (Источник сведений). Заглавия являются самым важным дополнительным элементом для подбора пар, что будет показано в заключительном разделе.

Для расширенного авторитетного файла PND 2,6 миллиона (90%) авторитетных записей получили некоторые дополнения, но только 2,0 миллиона (80%) были расширены за счет библиографических записей. Остальные 400 тысяч записей были дополнены заглавиями, извлеченными из самих авторитетных записей PND.

Методика тестирования подобранных пар

Участники VIAF помогали при разработке процесса подбора пар, проводя сплошной просмотр и комментируя результаты. Например, изначально использовались заглавия серий, но затем было установлено, что это часто приводит к неправильному подбору пар имен. При каждом просмотре в некоторых модификациях выявлялось либо увеличение количества подобранных пар, либо снижение количества ошибочно подобранных пар. Со временем были разработаны приемлемый порог количества и алгоритм подсчета баллов. Здесь мы приводим только окончательные проверки на подтверждение.

Для подтверждения точности и эффективности подбора пар экспериментальные примеры подобранных пар имен были просмотрены опытными в авторитетном контроле каталогизаторами из DDB и LC. Первый экспериментальный подбор пар проводился с двумя задачами: определить частичное совпадение имен между двумя авторитетными файлами и выяснить, какая часть этих пар имен может быть идентифицирована в процессе подбора. Второй пробный подбор был обращен на выявление систематических ошибок или недостатков, которые могут быть исправлены, и на приблизительную оценку общей доли ошибок.

Первая проба включала случайную выборку из 391 авторитной записи PND. Для подбора пар к этим записям были проведены ручной и автоматический поиски в авторитетном файле LC. Для увеличения доли автоматической обработки в этом эксперименте к авторитетным записям PND были подобраны авторитетные записи LC с общими фамилиями, в результате чего для проверки было подобрано 74 тысячи пар записей. Алгоритм подбора пар был применен ко всем 74 тысячам пар имен и автоматически было подобрано 79 пар авторитетных записей PND/LC.

Ручной просмотр всех 391 авторитетных записей PND выявил дополнительно 35 имен, на которые имелись соответствующие авторитетные записи LC, но которые не были выявлены ни при подоборе пар по фамилии, ни через алгоритм подбора пар при проверке правильности подбора пар. При ручном просмотре была подтверждена точность выбора 79 автоматически подобранных пар. Основываясь на данной экспериментальной выборке PND, определено, что около 30% имен PND представлены также и в авторитетных записях LC, и что алгоритм может подобрать в пары около 70% этих общих имен. Экстраполяция этих результатов на предполагаемые 800,000 имен, общие для двух авторитетных файлов, позволяет ожидать, что при автоматическом подборе пар будет идентифицировано 550,000.

Результаты были также рассмотрены с точки зрения совершенствования процесса подбора пар. Используя только фамилии почти для 1000 пар имен необходимо провести весь процесс полного подбора по каждой паре. Ручная проверка результатов подбора пар проводилась для того, чтобы убедиться, что стратегия, основанная на фамилии, имени плюс ограниченные сведения о датах, может быть использована для предварительного определения совместимости имен. Этот простой способ полезен и эффективен, а его небольшие корректировки обещают привести к дальнейшим улучшениям.

Второй пробный подбор был предпринят для выявления доли ошибок при подборе пар. Частью этого процесса явилась проверка экспериментальной выборки на адекватность предварительного расчета порога ошибок и, при необходимости, его корректировка.

Испльзование порога ошибок предполагало, что близкое к пороговым значениям количество ошибок при подборе пар будет больше, чем подсчитанное при подборе пар, и много больше, чем пороговое значение. Большая часть подобранных пар авторитетных записей имен показала результаты выше, чем пороговые значения. Для снижения доли ошибок до минимума, установленного при ручном просмотре, по итогам подсчетов выборка была разделена на четыре подгруппы. Ошибки выявлялись при ручном просмотре всех подобранных пар, и для каждой подгруппы была установлена доля ошибок и степень достоверности. Эти промежуточные результаты были оценены и обобщены для определения общей доли ошибок в методике подбора пар. Количество ошибочно подобранных пар составило менее одного процента.

Одна из подгрупп проверялась на пределы порогового значения. Если порог понижался, одна неправильно подобранная пара добавлялась к каждым трем правильным подборам. Понятно, что снижение порога не оправдано. При расчетах значений только выше пороговых, находилась только одна неправильная пара из 25 подбранных. Поскольку для этих расчетов использовалось относительно небольшое количество подборов, влияние ошибок на общий уровень является незначительным при сохранении значительной доли правильных подборов. Поэтому предварительный пороговый уровень был признан приемлемым.

Формирование начального VIAF

Расширенные авторитетные файлы из обоих источников были пропущены через алгоритм подбора пар, а полученные в результате этого записи, совпавшие и несовпавшие, были превращены в записи VIAF. Этот процесс представлен в таблице 2. В полученном VIAF представлено 6.3 млн. записей, включая 500,000 связанных записей, 3.7 млн. несовпавших записей из авторитетного файла LC и 2.1 млн. несовпавших записей из авторитетного файла PND. Это очень близко к значениям, установленным при ручном тестировании. Установлено, что существует дополнительно 250,000 пар авторитетных записей, представляющих одних и тех же людей, которые не могут быть подобраны автоматически из-за отсутствия необходимой информации. Окончательная версия системы в таких случаях будет давать возможность осуществлять подбор пар вручную и иную интеллектуальную идентификацию подобранных пар. Авторитетные записи будут включать последовательный номер записи VIAF.

В таблице 3 представлен пример записи VIAF в формате MARC 21. Поскольку основной целью VIAF является обеспечение связи между файлами, запись VIAF включает доступ по каждому имени в поле 700 (Heading Linking Entry = Связанный заголовок) вместе с указанием его источника. Поскольку единое авторитетное имя отсутствует, поле 100 (Заголовок имени лица) не используется. Если подобранная пара определена через алгоритм, в записи представлены два связанных заголовка. Если пара к имени не подобрана, представлено только одно поле 700.

Дополнительные сведения также включаются в расширенные авторитетные записи в локальные (9xx) поля. Локальные поля, используемые в расширенных автортетных записях кратко описаны в Таблице 4. Для упрощения процесса подбора пар весь текст стандартизован с использованием модифицированной версии правил стандартизации NACO (Name Authority Cooperative Program of the Program for Cooperative Cataloging = Программа кооперации авторитетного контроля имен Программы кооперативной каталогизации). [7] Количество случаев встречаемости специфического термина хранится в подполе $9. Поскольку данная информация предназначена преимущественно для машинной обработки, ее не обязательно представлять в выводных формах записей для конечного пользователя. При включениях последующих национальных авторитетных файлов они прежде всего будут сравниваться с

уже существующими расширенными записями VIAF, дополнительно подобранные пары включаются в записи VIAF таким же образом. Когда подбор пар произведен, дополнительная информация из подобранных пар записей также объединяется.

В значительном количестве случаев при подборе пар авторитетная форма имени из одного файла совпадает с множеством авторитетных имен из другого файла. Поскольку задачей VIAF является установление связей один-к-одному, подбор пар не получал подтверждения, если появлялось множество таких подобранных пар, и 70,000 алгоритмически подобранных пар были исключены из-за множественности подобранных пар. Были выявлены по крайней мере две причины такой множественности при подборе пар.

Во-первых, в PND существует некоторое количество недифференцированных имен, каждое из которых попадает в пару с двумя или более дифференцированных имен в LCNAF. В соответствии с Немецкими правилами каталогизации RAK-WB в немецкой каталогизационной практике допускалось не дифференцировать имена лиц. Когда DDB начала каталогизировать с авторитетным файлом, от этой практики пришлось отказаться и DDB больше не создает авторитетных записей с недиффернцированными именами лиц. Однако, PND все еще включает много недифферецированных имен. DDB будет дифференцировать имена с множеством подобранных пар, насколько это возможно, автоматически на основе совпадающих между LC и DDB пар заглавий, включенных в расширенные авторитетные записи, остальное – интеллектуальная ручная работа. Поправки обогатят VIAF в качестве текущих обновлений и увеличат количество точно подобранных пар записей.

Во вторых, некоторое количество авторитетных записей LC отражают практику AACR2, позволявшую создавать отдельные авторитетные записи для каждой библиографической индивидуальности, используемой одним лицом, такой как псевдонимы. Это – случай, обратный недифференцированным записям PND. В этом случае множество авторитетных записей создается для одной личности. PND, следуя правилам RAK-WB, включает только одну авторитетную запись имени для всех личностей. Как и недифференцированные имена, эти «слишком дифференцированные» автортетные записи рождают проблемы, для которых не были найдены полностью удовлетворяющие решения.

Связанные имена могут использоваться как простой перевод авторитетной формы имени из LC в PND или как visa versa. Это может соответствовать потребностям семантического Web или объединенных поисковых систем, что является одним из их основных требований. Поддержка трассировок «ссылка от» может обеспечить пользователю просмотр дополнительной информации.

Номера авторитетных записей из файлов-участников или номера собственно VIAF также могут быть основой URI. Это могло бы обеспечить потенциальную возможность для функционального решения авторитетного контроля URI. Начиная с какой-либо ссылки URI, представленной в документе, записи или Web-сайте, пользователь мог бы придти ко всем материалам, записям, ресурсам и т.п., которые связаны с авторитетными источниками, представленными в URI, а также к самим авторитетным записям.

Развитие системы

Национальные авторитетные файлы имен и библиографические базы данных постоянно изменяются. Для формирования связанной базы данных из двух или более изменяющихся файлов связи должны постоянно пересматриваться и дополняться. Логика и программное обеспечение начальной системы VIAF модифицируются для обеспечения возможности продолжать дополнение записей. Как только поступают новые библиографические или

автортетные записи, расширенные авторитетные записи модифицируются и база данных перекрестно связанных записей постоянно пересматривается. Новые подборы пар будут производиться постоянно, а подобранные пары, которые не должны далее поддерживаться из-за изменений в основных записях, должны быть разбиты. Если подобранные пары разбиваются, в каждой из пары таких записей должна сохраняться для справок история предыдущих пар.

В будущем система VIAF будет поддерживать преимущества OAI по хранению исходной базы данных, если такие возможности будут реализованы. Пока что для тестирований проекта будут использоваться более традиционные инструменты доступа к файлу, такие как FTP.

Могут рассматриваться многие другие методы обеспечения доступа и использования данных, пригодные при большом объеме данных, размещаемых в одном месте. Как часть семантического Web, связи могут использоваться для перевода имени лица в желаемый пользователем формат. Могут быть разработаны инструменты для автоматического поиска в альтернативных базах данных при поддержке формы имени для такой базы данных. Инструменты каталогизации и авторитетного контроля могли бы строиться подобным же образом, идентифицируя соответствующую форму для имени, включенного в запись. Конечно, база данных VIAF также будет непосредственно доступна для прямого поиска.

Выводы

Файл PND уже получил значительную пользу от проекта. Проверки подобравшихся пар в обоих файлах инициировали значительное повышение качества в PND, а DDB ожидает значительную поддержку работы по дифференциации имен лиц при выявлении одинаковых заглавий в парах расширенных записей. Разработанные для пректа технологии и алгоритмы могут применяться для решения многих других задач. Исследуются возможности использования данных по подобранным парам имен лиц для совершенствования доступа к библиографической информации и поддержки каталогизационной деятельности участников.

Проект показал практическую возможность автоматического установления связи между именами лиц в двух национальных авторитетных файлах. Семьдесят процентов авторитетных записей для лиц, представленных в обоих файла, были связаны с долей ошибок менее чем один процент. Стратегия дополнения оригинальных автортетных записей информацией из библиографических записей значительно улучшила показатели подбора пар, снизив количество ошибочных подборов. Незначительные изменения в авторитетных записях могут значительно улучшить подбор пар. Многие потери при подборе пар обусловлены недостатками грамматического разбора поля 670 (Источник данных). Дополнительный фактор, позволяющий избегать использования кратких имен и заглавий, или представляющий эксплицитные связи для основной библиографической записи, будет очень полезен. Эксплицитная идентификация области деятельности или специальности (композитор, иллюстратор, математик и т.д.) позволит расширить подбор пар и автоматически, и вручную, как для включения более полных форм имен, так и в качестве перекрестных ссылок.

Данное исследование представляет широкие возможности для совершенствования авторитетного контроля, для использования авторитетных записей, для сетевой и перекрестной связи и для построения семантического Web для библиотек. Для тех библиотек и библиотечных сетей в Германии, которые получают или хранят библиографические записи с точками доступа LCNAF, VIAF станет платформой для перехода из одного авторитетного файла в другой, так же как и для транскрибирования точек доступа LCNAF в библиографических записях в точки доступа PND, или для поиска и нахождения заголовков

PND через VIAF. Внедренный в многонациональный и многоязычный портал, например, в European Library portal [Европейский библиотечный портал], VIAF сможет объединять запросы и к LCNAF и к PND, ведя пользователя к связанным библиографическим записям из обоих источников.

При наличии технологии подбора пар, планируется создать способную к модернизации систему, в которой будут накапливаться авторитетные и библиографические данные об именах лиц от участников с использованием преимуществ OAI. Созданная система способна к расширению, поэтому будут приветствоваться новые участники, желающие объединить свои авторитетные и библиографические данные. Ограничения в возможностях расширения VIAF не выявлены, пока к проекту не присоединилось большее количество учреждений. Существуют планы по расширению возможностей системы при подключении набора символов Unicode. Unicode позволит включить не-римские шрифты, но при этом придется пересмотреть алгоритм подбора пар, особенно при использовании идеографических шрифтов, таких как корейский, китайский или японский.

Ссылки

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/EText/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]

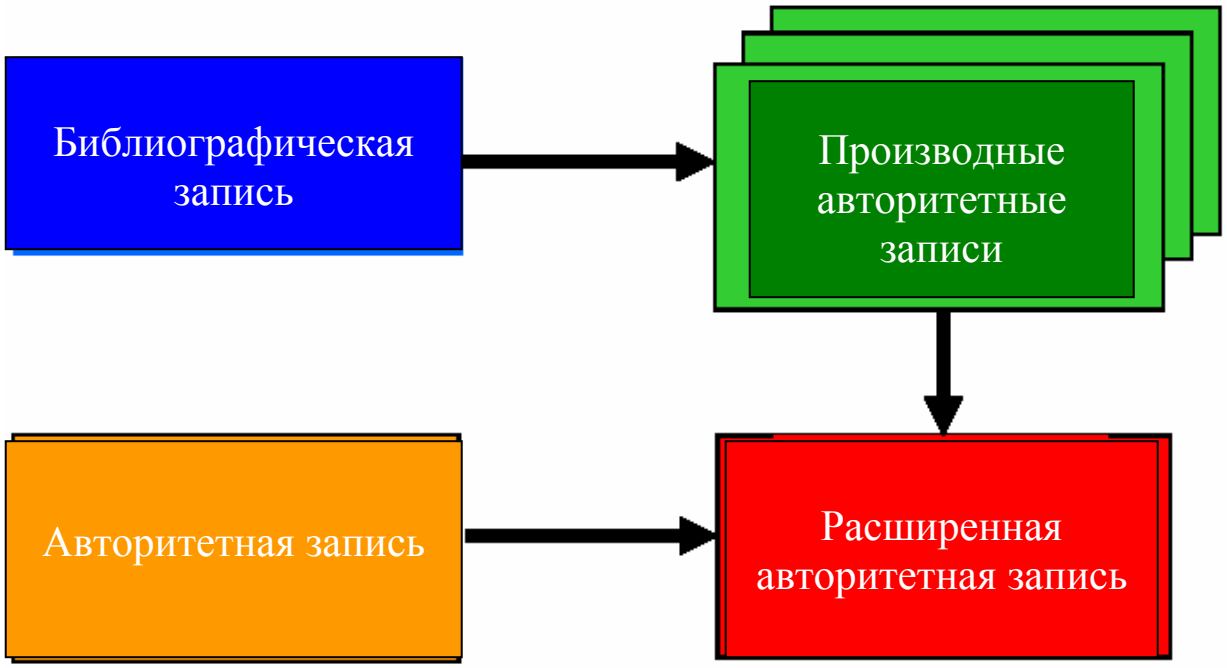

Таблица 1. Создание расширенной авторитетной записи

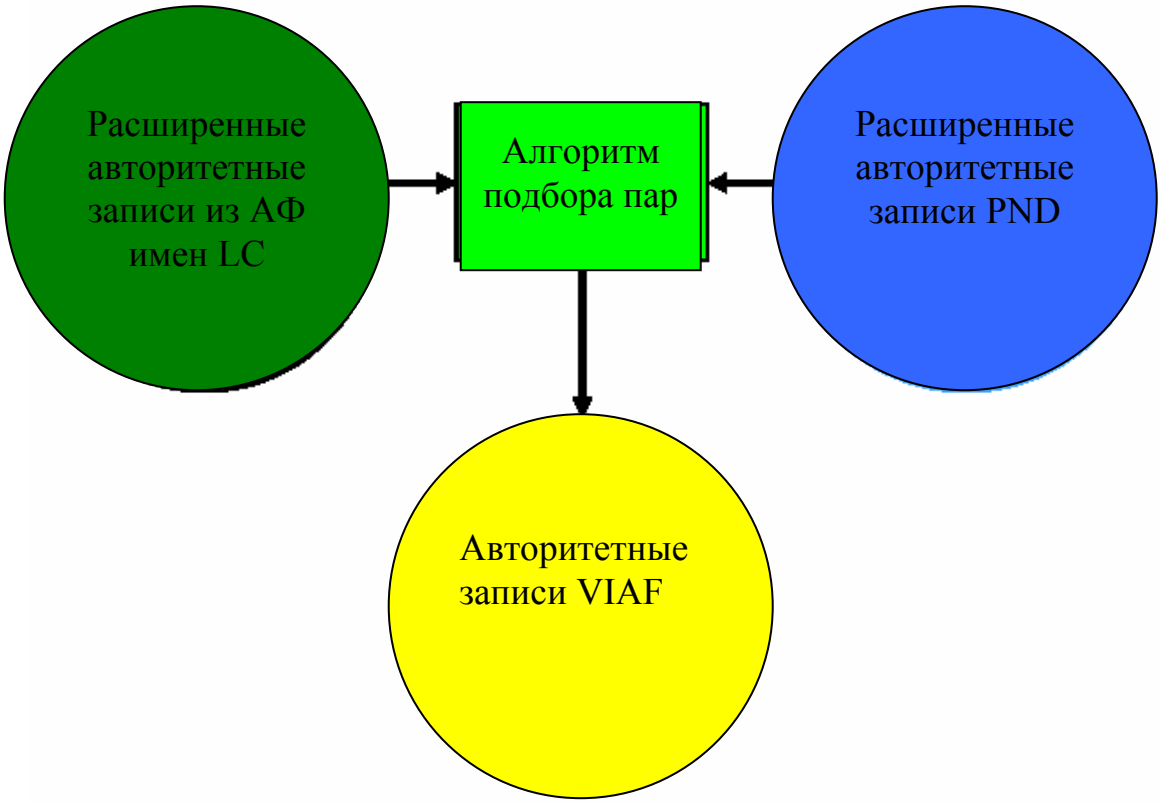

Таблица 2. Создание авторитетных записей VIAF

```
000    nz  n
001    viaf30543
005    20050826163535.0
008    050826n||anannabbn |a aaa
040    VIAF  $c VIAF
400 10 $w nnaO'Connor, Diane, 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
```

Таблица 3. Запись VIAF

Таблица 4
Форматы расширенной записи

90x	**Контрольные номера**		
	901	ISBN	$a Цифровая часть ISBN (без различения чисел и тире)
	902	ISSN	$a Цифровая часть ISSN (без различения чисел и тире)
	903	LCCN	$a Цифровая часть LCCN (без различения чисел и тире)
91x	**Поля заглавий**		
	910	Заглавие из 245	Подполя a & b
	911	Сокращенное заглавие из 210	Подполя a & b
	913	Унифицированное заглавие из 130 или 240	Подполя a & b
	914	Перевод заглавия из 242	Подполя a & b
	915	Обобщающее унифицированное заглавие из 243	Все подполя
	916	Вариантное заглавие из 246	Подполя a & b
		Авторитетная запись	Извлеченная из авторитетных записей Имя/Заглавие, поле 100
	917	Унифицированное заглавие	$t
	919	Заглавие, извлеченное из иного текста	Из различных примечаний или других подобных полей
92x	**Поля издательств**		
	920	Издательский номер	$a Издательский номер из ISBN
	921	Имя издателя	$a Имя издателя из полей 260 b или 533 c.
	922	Место публикации	$a Страна публикации код из 008
93x	**Употребляемость**		
	930	Употребление имени	$a Форма имени, указанная в сведениях об ответственности, 245 подполе c
94x	**Характеристики**		
	940	Язык	$a Код языка из 008 или 041 подполе a
	941	Роль автора	$a Код отношения из 700, подполя e и/или 4
	942	NATC Subject	$a NATC survey line number.
	943	Десятилетие публикации	$a Десятилетие публикации
	944	Формат	$a Тип и биб уровень (008/06-07)
	945	Conspectus Subject	Рспространенное использование, см. дискуссию PND
95x	**Соавторы**		
	950	Индивид. авторы	Подполя $a, $b, $c, $d, и $q либо из 100 либо из 700 полей
	951	Кол. авторы	Подполе $a либо из 110 либо из 710 полей
96x	**Имя как предмет**		
	960	Имя как предмет	Подполя $a, $b, $c, $d, и $q из поля 600
	969	Subject usage	Текст «Темы», указывающей авторитетный заголовок, использованный как тема и извлеченный из поля 600
99x	**Специальные поля**		
		Связанные	$a Общее количество записей
	999	библиогр. записи	$b Контрольный номер записи $2 Источник записи