УДК 026.06

К. Е. Соколинский

Ассоциация ЭБНИТ, Управление информационно-образовательных ресурсов Санкт-Петербургского государственного университета телекоммуникаций им. проф. М. А. Бонч-Бруевича

Новая технология создания сводных каталогов и корпоративных электронных библиотек в J-ИРБИС 2.0

Проанализирована технология, упрощающая создание и эксплуатацию локальных сводных каталогов и корпоративных электронных библиотек. Обоснован новый подход к процессу дедубликации библиографических записей.

Ключевые слова: J-ИРБИС 2.0, локальные сводные каталоги, корпоративные электронные библиотеки, библиографические записи, дедубликация, алгоритм.

UDC 026.06

Kirill Sokolinsky

Bonch-Bruevich St. Petersburg State University for Telecommunications Russian National Public Library for Science and Technology, Moscow, Russia

New technology of creating union catalogs and corporate digital libraries

The author describes the technology to simplify acquisition and management of local union catalogs, as well as corporate e-libraries. Proposed is a new approach towards bibliographic record deduplication.

Keywords: J-IRBIS 2.0, local union catalogs, corporate e-libraries, bibliographic record, deduplication, algorithm.

Несмотря на стремительное развитие интернет-технологий, электронные ресурсы российских библиотек остаются по преимуществу разрознен-

ными. Поэтому на различных уровнях предпринимаются усилия по интеграции как библиографических, так и полнотекстовых данных. Вузы одного профиля стремятся к созданию сводных каталогов своих фондов и объединению электронных коллекций, чтобы минимизировать свои затраты. Муниципальные библиотеки крупных городов и регионов формируют сводные каталоги для реализации технологии единого читательского билета и формирования краеведческих БД. Нередко у них есть и потребность в агрегации электронных версий – копий раритетных или краеведческих изданий.

На федеральном уровне реализуются сразу несколько интеграционных проектов. Информационная система доступа к каталогам библиотек сферы образования и науки в рамках единого интернет-ресурса (ИС ЭКБСОН) ориентирована на предоставление единой точки доступа к электронным каталогам вузовских и научных библиотек [**11**]. Национальная электронная библиотека (НЭБ) стремится к объединению полнотекстовых документов как защищённых авторским правом, так и являющихся общественным достоянием [**7**]. Сводный каталог библиотек России (СКБР) претендует на формирование эталонных записей и применяется в первую очередь для каталогизации путём заимствования [**3**].

В то же время, несмотря на значительную государственную поддержку крупных интеграционных проектов, роль локальных (межвузовских, региональных) сводных каталогов (СК) и электронных библиотек сохраняется. Федеральные проекты не могут в полной мере учесть цели отдельных библиотечных сообществ. Например, пользователь не может сделать заказ на издание в библиотеке своего города непосредственно через сайт федерального корпоративного портала. Такое решение требует высокого уровня технической интеграции и унификации стандартов, что становится тем более затруднительным, чем больше библиотек участвует в проекте.

С ещё более серьёзными проблемами сопряжено глобальное объединение электронных полнотекстовых коллекций в рамках НЭБ. Технология предоставления, обновления и исключения документов в НЭБ пока реализована лишь для РГБ [**7**]. Её полная реализация требует сложнейшей системы коннекторов к различным системам автоматизации библиотек (САБ). И даже если такое программное обеспечение будет разработано, необходимы значительные усилия для унификации и конвертирования документов, не соответствующих стандартам НЭБ (PDF/A).

Дополнительный скепсис по отношению к федеральным проектам провоцируется и негативным опытом, поскольку зачастую такие проекты не достигали своих целей [Там же] и даже неожиданно прекращали своё существование (как это произошло с Центральной библиотекой образовательных ресурсов).

Наконец, передача полнотекстовых ресурсов в федеральную электронную коллекцию негативно воспринимается, поскольку библиотеки-владельцы лишаются статуса эксклюзивных обладателей документов и уменьшается посещаемость их собственных сайтов.

Поэтому можно утверждать, что определяющее значение в области интеграции библиотечных ресурсов имеют локальные (региональные, межвузовские) проекты. Они могут выступить эффективным посредником при передаче данных от библиотек к государственным агрегаторам и в то же время позволяют удовлетворить специфические потребности отдельных библиотечных сообществ. Необходима легко тиражируемая технология, предъявляющая минимальные требования к обслуживающему персоналу головной организации и к рядовым участникам проекта. Такая технология должна удовлетворять потребность в сводных каталогах и электронных коллекциях, создаваемых с различными целями.

Технологии создания СК[1]

В автоматизированном формировании электронного СК можно выделить четыре ключевых процесса: сбор записей, конвертирование (приведение к единому формату), дедубликацию (исключение дублирующихся записей), консолидацию (слияние существенных данных из дублирующихся записей), актуализацию (добавление, исключение или обновление записей). Целесообразно проанализировать некоторые существующие сегодня СК с точки зрения этой схемы.

Одним из первых электронных СК, работа над которым началась ещё в 1980-е гг., был Российский сводный каталог научно-технической литературы (РСК НТЛ). В рамках этого проекта записи передавались агрегатору (ГПНТБ СССР) в виде файлов в MARC-совместимых форматах и приводились к формату UNIMARC. Нередко из-за несоответствия записей формата такая конвертация была сопряжена со значительными трудностями. В этом СК началось использование технологии дедубликации с помощью библиографической свёртки записей, т.е. путём сравнения текстовых строк, сформированных из полей сравниваемых записей[2] (см. рис. 1).

---

[1] Приведённый ниже обзор сводных каталогов не претендует на полноту. Он предпринят лишь с иллюстративными целями.

[2] Здесь и далее под «полем записи» подразумеваются типичные структурные единицы описания документов, а не конкретные поля форматов семейства MARC.

Автор:	Кузьмина Л. М.
Заглавие:	Конструктор Сухой. Люди и самолеты
ISBN:	5-203-01472-8
Количество страниц:	383
Год издания:	1995
Рабочий лист:	PAZK

Рис. 1. Формирование свёртки для дедубликации

Такой подход в дальнейшем получил большое распространение, став едва ли не эталонным на длительный срок. В то же время был выявлен и его существенный недостаток – стало крайне сложно исключать записи из каталога (например, в случае списания книг). Кроме того, технология проверки на дублетность по свёртке не могла гарантировать высокую точность дедубликации. В связи с этим возникала потребность в ручном контроле качества [**5**].

Радикальной попыткой преодолеть сложности конвертирования и обмена записями стала концепция СКБР: она оставляла гораздо меньше свободы библиотекам, участвующим в наполнении каталога. Рядовые участники (кроме двух национальных библиотек) были обязаны выполнять каталогизацию в едином интерфейсе САБ *OPAC Global* [**8**]. В случае использования САБ, отличной от *OPAC Global*, библиотеки должны были проводить каталогизацию дважды – в собственной САБ и в *OPAC Global*. Поэтому, учитывая высокую трудоёмкость, технология СКБР едва ли может тиражироваться и получить распространение в условиях, когда отсутствует возможность административного влияния на участников.

Примером реализации идентичной технологии, но уже на локальном уровне может выступать Электронный каталог периодических изданий (в рамках Корпоративной сети библиотек Санкт-Петербурга). Однако здесь процесс изначально был облегчён тем, что участники каталога не создавали

запись в СК, а лишь копировали её туда. Это позволило избежать разработки специализированного ПО, но потребовало значительных инфраструктурных инвестиций. Библиотеки города, входящие в число составителей СК, были объединены высокоскоростными каналами связи.

Иной подход использовался при формировании отдельных каталогов в рамках АРБИКОНа (например, *RUSLANet*). Здесь сбор записей и их дедубликация выполнялись полностью автоматически. Единственное требование, которое предъявлялось к участникам, – это обеспечение возможности работы с их каталогами по протоколу Z39–50.

Применение общепризнанного стандарта для предоставления записей стало прогрессивным шагом. Появилась возможность автоматической поддержки актуальности СК путём периодического опроса систем библиотек-участников. Но в то же время потребовались значительные усилия для того, чтобы обеспечить работу Z39–50-серверов в соответствии с профилем АРБИКОНа, поскольку не существует систем автоматизации, поддерживающих предусмотренные стандартом протокола атрибуты и одинаково трактующих формат RUSMARC.

Примечательна технология ведения негосударственного коммерческого электронного каталога «Open for you», объединяющего значительное количество библиотек Востока России. Этот СК не только использует имеющиеся у участников записи, но и корректируется профессионалами-библиографами. Для работы с ним библиотеки-участники должны особым образом настроить САБ ИРБИС и установить специальное ПО, которое автоматически передаёт записи и вместе с тем, если это необходимо, обновляет записи каталогов [**6**].

В этой концепции нашли своё отражение потребности школьных библиотек, которые далеко не всегда готовы самостоятельно каталогизировать, а также реализован ещё один подход к сбору данных. Однако обязательность приобретения и настройки дополнительного ПО участниками, необходимость уплаты членских взносов создают определённые барьеры для вхождения в проект.

Принципиально новый путь сбора данных, практически противоположный предшествующим, был реализован в ИС ЭКБСОН. Здесь практиковалось уже не приведение библиотек-участниц к стандартам СК, а наоборот адаптация системы сбора данных СК к стандартам библиотек-участниц. В рамках проекта были разработаны провайдеры/коннекторы ко многим распространённым библиотечным системам, и за счёт технической простоты получения записей процесс значительно упростился.

Итогом этого проекта стал самый репрезентативный СК в РФ (больше 10 млн записей). Но в то же время требования к скорости сбора и дедубликации данных определили необходимость использования технологии формирования СК, близкой к традиционной.

Проблемы существующих технологий создания СК

Некоторые проекты достигли существенного прогресса в автоматизированной технологии сбора данных. И в настоящее время за счёт резкого сокращения количества распространённых САБ впервые стала реальной задача разработки полного набора коннекторов/провайдеров. Но такому «технологическому» подходу ещё лишь предстоит стать общепризнанным, особенно в локальных СК. Гораздо чаще вопрос унификации стандартов решается путём унификации технологических процессов библиотек-участниц, а не совершенствованием технологий.

Консолидация записей чаще всего ограничивается объединением сигл фондодержателей. Обогащение одной эталонной записи информацией из других не получило распространения. Причины этого как в неоднородности записей, так и в отсутствии эффективных алгоритмов консолидации.

Также существует большое количество проблем, связанных с распространённой технологией дедубликации по библиографической свёртке (рис. 1). Она остаётся существенным препятствием для автоматического создания качественных СК с минимумом дублетных или ошибочно консолидированных записей.

Технология дедубликации по свертке чувствительна к любым погрешностям в библиографических записях. К ошибкам дедубликации могут приводить не только ошибки в записях, но и незаполненные поля. В то же время обеспечение единообразия записей затруднительно не только в корпоративных объединениях, но даже в рамках каталогов одной библиотеки. К сожалению, какими бы странными не казались некоторые погрешности, их приходится рассматривать как данность при создании любого СК.

Чтобы подтвердить сказанное, достаточно рассмотреть пример с одним из самых формализованных полей – «Год издания». На первый взгляд, это поле предельно однозначно: оно должно включать лишь четыре цифры и заполняться одинаково. Но ГОСТ 7.1-2003 предписывает указывать «Год издания» в квадратных скобках, если он напечатан не на титульном листе и определён каталогизатором. Кроме того, ГОСТ допускает в случае, если отдельные цифры года не известны, проставлять вместо них знаки вопроса. А такие правила могут как выполняться каталогизатором, так и игнорироваться – из соображений практичности и целесообразности (в некоторых

САБ – для упрощения поиска). Если один сотрудник не решится самостоятельно определять год издания и укажет «Б.г.», второй – может заполнить поле как «[198?]», третий – указать выявленный им год целиком в квадратных скобках [1982], а четвёртый – вообще не заполнить это поле.

Может показаться, что записи, в которых поле «Год издания» заполнено некорректно или не соответствует определённым правилам (например, для поля «Год издания» это четыре цифры, где первая – не больше двойки), достаточно легко игнорировать при дедубликации, подвергая каждый вариант заполнения формально-логическому контролю. Но для технологии библиографической свёртки это не так. Если в одной из записей поле будет заполнено правильно – в соответствии со схемой, а в других нет, то свёртки окажутся различными и, следовательно, записи будут признаны различными.

Следует сказать и о проблемах дедубликации, связанных с опечатками или формально корректным, но различным заполнением полей «Заглавие» и «Автор». Для многих каталогизаторов существенной проблемой является определение заглавия и сведений, относящихся к заглавию. Ошибки при описании нормативных документов допускают даже опытные сотрудники. А в заглавиях журналов название журнала, серии, номера серии и названия серии зачастую конструируются каталогизаторами в самом прихотливом порядке и через непредсказуемые знаки препинания.

Поле «ISBN/ISSN» должно использоваться как основной инструмент для дедубликации, однако значительная часть публикаций, с которыми работают во многих локальных корпоративных проектах (университетские, региональные издания), не имеет этих идентификаторов. ISBN/ISSN зачастую используются неправомерно (для различных изданий одной книги) или откровенно недобросовестно (когда издательство проставляет ISBN другой книги). Поэтому в России ISSN/ISBN не придаётся большого значения, и соответствующие поля зачастую не заполняются даже в записях научных библиотек.

Нет необходимости говорить и об опечатках, которые естественны и неизбежны даже в ключевых полях. Достаточно лишь отметить, что внешне совершенно одинаковые записи могут с технической точки зрения различаться, так как вместо кириллической была использована латинская буква *C*. Эти буквы соответствуют одной клавише клавиатуры и часто подменяют одна другую.

Таким образом, распространённая сегодня технология дедубликации по свёртке является ограниченной и морально устаревшей: она не способна решить задачу автоматического создания СК потенциально возможного качества.

Новая технология формирования СК на основе пороговой дедубликации

Анализ используемых технологий и их ограничений позволяет обозначить совокупность характеристик, которым должна соответствовать современная система формирования СК, ориентированная на потребности региональных или университетских сообществ:

Совместимость с различными технологиями работы библиотек – участниц каталога;

Низкая стоимость и низкие эксплуатационные издержки;

Простота использования и настройки;

Высокое качество создаваемого СК;

Гибкость и расширяемость базовых функциональных возможностей;

Использование внешних авторитетных библиографических ресурсов.

Допустимо утверждать, что этим требованиям отвечает технология автоматического формирования СК и электронных коллекций на базе модуля J-ИРБИС 2.0 САБ ИРБИС. Она была разработана для макета ИС ЭКБСОН, где применялась для объединения каталогов 66 библиотек и формирования СК, объёмом в 2 600 тыс. записей.

J-ИРБИС 2.0 – распространённая система построения библиотечных порталов и электронно-библиотечных систем, ориентированных на стандарты WEB 2.0. Модуль предоставляет широкие поисковые возможности: поиск в стиле Google, автоматическое извлечение обложек из интернета, настройка отображения результатов, специальный интерфейс для мобильных устройств и многое другое.

Интеграция системы формирования СК в портальное решение позволяет библиотечному объединению не только создавать СК, но и разместить его в интернете. С помощью J-ИРБИС может быть обеспечен доступ к корпоративной коллекции полнотекстовых или мультимедийных документов. При этом, в зависимости от корпоративной политики, существует возможность предоставить права на выгрузку документов только определённым категориям пользователей или ограничить доступ просмотром документов в браузере.

Интеграция обеспечивает возможность управления системой СК в визуальном режиме через административную панель модуля. Создание СК может запускаться в интерфейсе сайта любым сотрудником библиотеки путём нажатия одной кнопки.

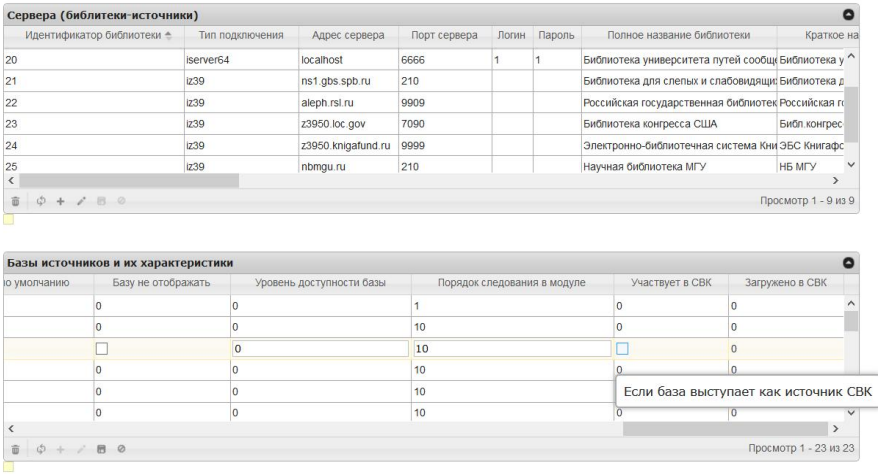

Рис. 2. Выбор баз для использования в СК

Элементарный табличный интерфейс предусматривает не только подписи к параметрам настройки, но и справку по каждому параметру (рис. 2). Несколько стандартных кнопок («редактировать», «удалить», «сохранить») позволяют оперировать данными. Стандартный интерфейс административной панели J-ИРБИС 2.0, используемый для настройки широковещательных запросов, применяется также для ввода данных о библиотеках – участницах СК. Здесь же определяются правила дедубликации записей.

Дополнительным преимуществом, с точки зрения администратора СК, является автоматическое обновление программного обеспечения. Система может обновляться без участия человека непосредственно в момент выхода новых версий.

Таким образом, включение системы формирования СК в модуль создания библиотечного портала одновременно решило несколько задач: эффективного использования каталога и электронных ресурсов, упрощения администрирования и поддержки программного обеспечения в актуальном состоянии.

Технические характеристики системы сбора данных во многом идентичны ИС ЭКБСОН. Здесь также предусмотрен автоматический (роботизированный) поиск с использованием технологии провайдеров, которые могут осуществлять выгрузку записей из библиотек по различным протоколам (Z39-50, WEB-ИРБИС или J-ИРБИС, JSON-RPC формата J-ИРБИС 2.0). Этот набор поддерживаемых протоколов может расширяться. Независимо

от исходного формата и кодировки, записи преобразуются в соответствии с единым форматом – САБ ИРБИС. Для конвертирования также используются конвертеры, сертифицированные Национальной службой развития RUSMARC.

Актуализация СК не требует операционных затрат и выполняется в автоматизированном режиме путём либо получения вновь добавленных в каталоги записей библиотек-участниц, либо создания каталога заново.

Но особый практический интерес представляет сама технология объединения записей, радикально отличающаяся от всех используемых сегодня в России. Алгоритм дедубликации, реализованный в проекте, претендует на обеспечение максимально возможного в настоящее время качества автоматической проверки на дублетность. Поскольку качество дедубликации определяет также качество консолидации (слияние различных записей крайне нежелательно), можно утверждать, что оно определяет и качество СК в целом.

Алгоритм выявления дублетности, реализованный в описываемой системе, имеет мало общего с другими, упомянутыми в современной специальной литературе[**1**, **10**]. Система базировалась на опыте, полученном в ходе решения задач из других областей библиотечной практики. Первым стимулом стала необходимость предельно надёжного слияния созданных вручную и полученных из автоматизированной системы университета записей читателей библиотеки. Ошибочное отождествление двух различных записей приводило к материальным проблемам: читателю приписывались задолженности перед библиотекой, которых он не имел. Не менее проблематичной была и противоположная ситуация, когда запись дублировалась, – в этом случае сотрудник мог не заметить задолженности и тем самым освободить читателя от ответственности за взятую литературу.

С другой стороны, был учтён опыт разработки системы широковещательного поиска для ИРБИС-корпорации, где ставилась задача дать пользователю возможность однозначной идентификации книги и в то же время обеспечить минимизацию вводимых для этого данных. Громадный массив источников ИРБИС-корпорации (более 160 библиотек) позволил оценить разнообразие и характер различных вариантов отклонения записей от стандарта. Кроме того, он дал обширную статистику, позволяющую определить набор данных, необходимый для идентификации записей.

Полученный опыт определил в новой технологии следующую последовательность операций.

1. *Формально-логический контроль записи.* Он выполняется, чтобы определить соответствие записи минимальным требованиям. В случае, если запись по каким-то причинам не содержит базового набора элементов, она

вообще не анализируется и рассматривается как бракованная. Например, не может быть записи без поля «Заглавие».

2. *Формально-логический контроль полей.* До того как проверять запись на дублетность, каждое существенное для дедубликации поле проходит формальный контроль на корректность содержимого. Например, если поле «Год издания» имеет вид «[198?]», оно не будет участвовать в проверке на дублетность, так как не содержит унифицированного значения. Это позволяет не выполнять сравнение полей, значение которых не однозначно.

3. *Создание метабиблиографического цифрового образа документа.* Точно так же, как документ служит основой для создания библиографической записи, сама БЗ используется для создания метабиблиографического образа. Так же, как каталог имеет своей целью ускорение поиска документов, база метабиблиогарфических образов ориентирована на то, чтобы сократить количество операций, связанных с поиском дублетов.

Все поля, признанные годными для сравнения, подвергаются нормализации. Например, ISBN приводится к чисто цифровому значению без дефисов, пробелов и буквенных символов. Поле «Заглавие» очищается от всех знаков препинания и пробелов. Имя и отчество автора, сложные инициалы приводятся к двум символам. И такому преобразованию подвергаются практически все поля записи.

Второй этап – хеширование. Все текстовые поля преобразуются в цифровые значения определённой структуры (хэшей). Особенностью этих хэшей является то, что сравнение двух хэшей (цифр), полученных из полей разных записей, позволяет не только определить идентичность закодированных в них строк, но и выявить незначительные отличия этих строк путём использования простых математических операторов («равно», «больше» или «меньше»).

Если стандартное определение подобия двух строк, отличающихся одной опечаткой, обычно требует больших вычислительных ресурсов (необходимо посимвольно сравнивать эти строки по сложному алгоритму), то при такой технологии достаточно лишь одной математической операции. Каталогизатор может добавить лишнюю букву, пропустить букву или заменить одну букву другой – все эти типичные ошибки легко обрабатываются. Кроме того, алгоритм позволяет различить полную идентичность и подобие строк, за счёт чего появляется возможность более точно диагностировать дублетность на следующих этапах.

Наряду со свёртками полей, метабиблиографический образ содержит служебные данные, которые не имеют прямой связи с библиографией – адрес записи в БД и её «индекс качества». За счёт них реализуется возможность извлечения записи и выбора наилучшей.

4. *Запись метабиблиографического образа*. Как следует из описания алгоритма, создание метабиблиографического образа – достаточно ресурсоёмкая процедура. Поэтому её результат сохраняется и используется многократно. Когда в каталог добавляется новая запись, её вновь созданный метабиблиографический образ сравнивается с тем, что уже подготовлен и сохранён ранее.

В зависимости от размеров каталога сохранение образа может выполняться либо в реляционную базу, либо в оперативную память. Применение внешнего хранилища снимает с процесса дедубликации целый ряд ограничений, заложенных в САБ, структуре поисковых индексов и алгоритмах их формирования. Во-первых, это избавляет от необходимости формирования ненужных с точки зрения дедубликации индексов и, следовательно, ускоряет процесс. Во-вторых, даёт возможность использования для формирования метабиблиографического образа алгоритмы, которые значительно сложнее стандартных алгоритмов создания индексов. В-третьих, позволяет избежать продолжительных и неоправданных операций записи новых данных обычно в достаточно медлительную библиографическую СУБД. Внешний поисковый метабиблиографический индекс даёт значительную свободу и ускорение при использовании описываемых сложных алгоритмов.

5. *Поиск дублетов по свёртке метабиблиографической записи.* Поскольку технология поиска по свёртке остаётся самой быстрой, этот метод в описываемом алгоритме также применяется в первую очередь. Он даёт исключительно высокий эффект, так как позволяет определить соответствие более чем в 95% случаев. В то же время недостатки этого метода легко нивелируются на следующем этапе.

6. *Определение рейтинга соответствия*. В том случае, если запись не была найдена по метабиблиографической свёртке, начинает выполняться основной алгоритм. Его первый шаг – отбор минимально похожих записей по небольшому набору полей, ошибки в которых маловероятны. Среди таких полей могут быть, например: «Вид записи», «Номер тома» и ISBN. Чем больше этих полей, тем выше будет скорость и тем ниже качество дедубликации.

После выполнения первичного поиска каждый метабиблиографический образ из подмножества, полученного в результате первичного отбора, сопоставляется с образом базовой записи. При этом выделяются две степени соответствия: полная идентичность и подобие. Для буквенных полей подобными могут считаться поля с одной ошибкой. Для цифровых полей допустимая погрешность определяется индивидуально, также в цифровой форме. Например, погрешность в количестве страниц может достигать ± 5 страниц. Для поля «Год издания» погрешность в принципе не предусмотрена.

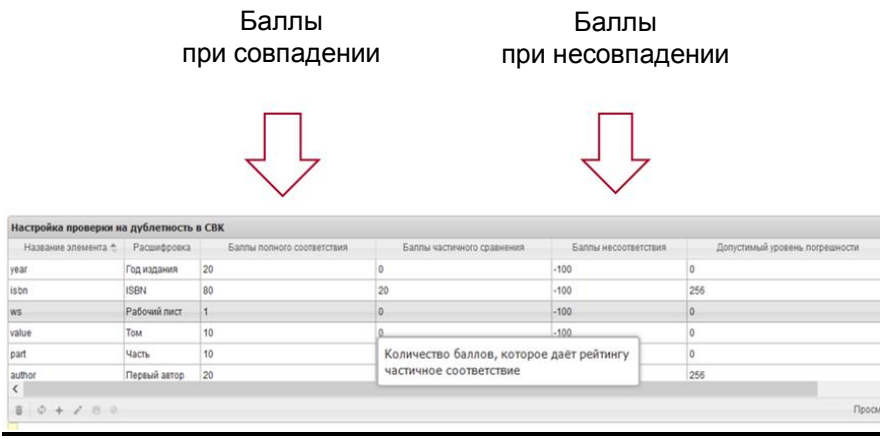

Рис. 3. Настройка рейтинга совпадения элементов

В алгоритме учитывается значение каждого поля для диагностики дублетности (рис. 3). Например, совпадение поля «Год издания» совсем не означает, что записи идентичны – в один год может быть издано несколько тысяч книг. В то же время совпадение «Года издания» и «Автора» уже значительно повышает вероятность того, что записи описывают одну книгу. Если же совпадает «Год издания», «Автор», «Заглавие» и ISBN, можно с большой вероятностью говорить о дублетности.

В то же время, если «Год издания» не совпадает, то с высокой вероятностью можно предположить, что записи составлены на разные издания одной книги.

Такие характеристики, как «соответствие», «неполное соответствие», «несоответствие», сопоставлены с определённым (отрицательным или положительным) количеством баллов, суммой которых (индекс соответствия) оценивается степень идентичности записей. Индекс соответствия в цифровой форме выражает степень идентичности записей. Записи признаются дублетными в том случае, если индекс превышает определённую пользователем пороговую величину (например 63).

Таким образом, у администратора СК появляется возможность, не вдаваясь в технические детали и манипулируя лишь одним цифровым показателем, получать каталог с нужными характеристиками. Если требуется снизить погрешность неправомерного слияния записей, администратор может повысить показатель, если нужно минимизировать дублетность в каталоге – понизить.

Максимальная чувствительность алгоритма может потребоваться для формирования корпоративных электронных библиотек. В этом случае мо-

жет использоваться применяемый некоторыми агрегаторами электронных ресурсов (*Summon*, например) принцип «контент важнее публикации». Он постулирует: поскольку для пользователя важен текст, а не печатная публикация, то «Евгений Онегин» 2010 года имеет такое же значение, как и «Евгений Онегин» 2011 года. Поэтому нет смысла формировать отдельные библиографические записи на оба издания – достаточно одной схематичной записи и одного текста. Такой подход сегодня мог бы быть очень продуктивным в университетах. Распространённые учебники насчитывают по 20 переизданий, каждое из которых обладает с точки зрения студента практически одинаковой ценностью, поэтому оцифровывать все издания нет смысла. В таких случаях единственной функцией системы формирования СК будет корректная дедубликации исходя из «Автора» и «Заглавия».

Помимо предоставления возможности управления количеством баллов за совпадение отдельных полей, использование порога соответствия повышает надёжность, гибкость и управляемость. Но у технологии есть и ряд недостатков. Она отличается высокой сложностью и значительно медленнее формирует каталог, чем стандартная технология дедубликации по свёртке, поэтому её применение будет обеспечивать приемлемую скорость создания каталогов в локальных объединениях, включающих не более 100 библиотек.

Консолидация записей в большинстве СК сводится к дополнению записи из доверенного источника сиглами библиотек – держателей документа. В некоторых решениях это выполняется с помощью статичного программного кода, который используется только для объединения конкретных полей записей при определённых условиях. А в качестве базовой записи (которая дополняется элементами других) выбирается запись самого авторитетного источника.

В J-ИРБИС 2.0 базовая запись выбирается по произвольному формальному признаку: им может быть авторитетность источника, размер записи, количество полей или совокупность нескольких характеристик. В каждом случае записи присваивается абстрактный цифровой «индекс качества», который в дальнейшем определяет её право выступать базовой записью.

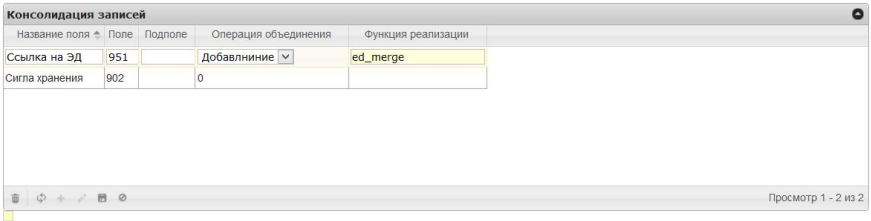

Рис. 4. Настройка консолидации записей

Особенностью алгоритма объединения полей является возможность его визуальной настройки для любых выбранных пользователем полей и подполей (рис. 4). Таким образом, могут быть определены два типа объединения.

Первый тип – «замена», когда одно значение неповторяющегося поля должно быть заменено на другое, приемлем, например, для поля «Аннотация» – когда аннотация более авторитетной организации должна заменить аннотацию менее авторитетной.

Второй тип – «добавление», когда повторяющееся поле дополняется ещё одним – оригинальным, может использоваться, например, для поля «Ключевые слова», если важно сформировать наиболее полный поисковый образ документа.

Третий тип – «сложное объединение», когда используются типовые или наоборот специализированные программные функции, включающие нужные алгоритмы объединения полей и их параллельной проверки соответствия стандартам. Например, при формировании обратных ссылок (из СК на сайт библиотеки – держателя документа) требуется сконструировать адрес записи на сайте библиотеки-держателя. Для этого нужно зафиксировать базу, из которой было выполнено заимствование записи, адрес сайта библиотеки – держателя документа, а затем добавить к этому дополнительные параметры. В результате не просто объединяются имеющиеся данные, но и формируются новые.

Перспективным направлением развития технологии является консолидация с использованием записей ЭКБСОН – открытого и общедоступного ресурса. Этот подход имеет смысл использовать при условии невысокого качества каталогов библиотек-участниц. Но он может быть применён лишь для той части изданий, которая имеет достаточно надёжные идентификаторы (например, ISBN).

Таким образом, настройка консолидации реализуется через гибкий и простой инструментарий. Пользователь может настроить процесс на выполнение элементарных операций, встроенных или собственных функций консолидации.

Агрегация полных текстов – одно из самых важных направлений развития системы СК на базе J-ИРБИС 2.0. С помощью имеющейся технологии существует возможность извлекать электронные документы через сайты практически всех библиотек – пользователей САБ ИРБИС, а также получать открытые версии документов по абсолютным HTTP- или FTP-ссылкам (например, http://library.ru/books/doc.pdf). Но, несмотря на техническую простоту задачи, она может быть решена далеко не во всех случаях. Выгрузить документы из специально защищённых (например, с помощью *Vivaldi*) от выгрузки электронных коллекций пока невозможно. Для таких

систем требуется разрабатывать специальные провайдеры/коннекторы или согласовывать вопросы технологии с разработчиками.

Перспективы использования технологии

Таким образом, в качестве дополнения к модулю САБ ИРБИС, предназначенному для создания библиотечных порталов, разработан *framework* (конструктор), в котором реализованы функции создания СК и корпоративных электронных библиотек. По согласованию с библиотечными консорциумами из него формируется нужная конфигурация системы создания СК. Система не требует ни изменения технологий работы библиотек – участниц СК, ни специального обучения персонала, ни дополнительных трудозатрат на поддержку. При этом она позволяет формировать СК потенциально возможного в настоящее время качества.

Экономические характеристики такой системы дают возможность конкурировать с другими решениями за счёт примерно десятикратной разницы в стоимости внедрения и администрирования.

СПИСОК ИСТОЧНИКОВ

1. **Достовалов С. С.** Библиопортал: что ожидает пользователь? [Электронный ресурс] / С. С. Достовалов // XI Международ. науч.-практ. конф. и выставка «Корпоративные информационно-библиотечные системы: технологии и инновации» (11; 2013; Санкт-Петербург) : сб. материалов конф. – Электрон. дан. – (Электронная библиотека СПбПУ). – Режим доступа: http://dl.unilib.neva.ru/dl/2/3298.pdf. – Загл. с экрана.

Dostovalov S. S. *Biblioportal: chto ozhidaet polzovatel? [Elektronnyy resurs] / S. S. Dostovalov // XI Mezhdunarod. nauch.-prakt. konf. i vystavka «Korporativnye informatsionno-bibliotechnye sistemy: tehnologii i innovatsii» (11; 2013; Sankt-Peterburg) : sb. materialov konf. – Elektron. dan. – (Elektronnaya biblioteka SPbPU). – Rezhim dostupa: http://dl.unilib.neva.ru/dl/2/3298.pdf. – Zagl. s ekrana.*

2. **Кулиш О. Н.** Работа со сводным каталогом библиотек России (ЛИБНЕТ): успехи и трудности [Электронный ресурс] / О. Н. Кулиш // XI Международ. конф. «Крым–2004: Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» (11; 2004; Судак) : сб. материалов конф. – Электрон. дан. – Режим доступа: http://gpntb.ru/win/inter-events/crimea2004/252.pdf. – Загл. с экрана.

Kulish O. N. *Rabota so svodnym katalogom bibliotek Rossii (LIBNET): uspehi i trudnosti [Elektronnyy resurs] / O. N. Kulish // XI Mezhdunarod. konf. «Crimea–2004: Biblioteki i informatsionnye resursy v sovremennom mire nauki, kultury, obrazovaniya i biznesa» (11; 2004; Sudak) : sb. materialov konf. – Elektron. dan. – Rezhim dostupa: http://gpntb.ru/win/inter-events/crimea2004/252.pdf. – Zagl. s ekrana.*

3. **Логинов Б. Р.** СКБР: от профессиональной каталогизации к публичному поиску [Электронный ресурс] / Б. Р. Логинов // Унив. кн. – Электрон. дан. – 2011. – № 9. – С. 44–48. – Режим доступа: http://www.unkniga.ru/biblioteki/fonds/254-interview-loginov-skbr.html. – Загл. с экрана.

Loginov B. R. *SKBR: ot professionalnoy katalogizatsii k publichnomu poisku [Elektronnyy resurs] / B. R. Loginov // Univ. kn. – Elektron. dan. – 2011. – № 9. – S. 44–48. – Rezhim dostupa: http://www.unkniga.ru/biblioteki/fonds/254-interview-loginov-skbr.html. – Zagl. s ekrana.*

4. **Надпорожская Е. В.** Базы данных Корпоративной сети общедоступных библиотек Санкт-Петербурга в среде ИРБИС: программные и технологические решения [Электронный ресурс] / Е. В. Надпорожская, Л. А. Яковишина, С. К. Егоров // XVIII Международ. конф. «Крым–2011: Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» (18; 2011; Судак) : сб. материалов конф. – Электрон. дан. – Режим доступа: http://www.gpntb.ru/win/Inter-Events/crimea2009/disk/57.pdf. – Загл. с экрана.

Nadporozhskaya E. V. *Bazy dannyh Korporativnoy seti obshchedostupnyh bibliotek Sankt-Peterburga v srede IRBIS: programmnye i tehnologicheskie resheniya [Elektronnyy resurs] / E. V. Nadporozhskaya, L. A. Yakovishina, S. K. Egorov // XVIII Mezhdunarod. konf. «Crimea–2011: Biblioteki i informatsionnye resursy v sovremennom mire nauki, kultury, obrazovaniya i biznesa» (18; 2011; Sudak) : sb. materialov konf. – Elektron. dan. – Rezhim dostupa: http://www.gpntb.ru/win/Inter-Events/crimea2009/disk/57.pdf. – Zagl. s ekrana.*

5. **Рагимова М. А.** Российский сводный каталог: этапы большого пути [Электронный ресурс] / М. А. Рагимова// XII Международ. конф. «Крым–2005: Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» (04; 2005; Судак) : сб. материалов конф. – Электрон. дан. – Режим доступа: http://www.gpntb.ru/win/inter-events/crimea2005/disk/203.pdf. – Загл. с экрана.

Ragimova M. A. *Rossiyskiy svodnyy katalog: etapy bolshogo puti [Elektronnyy resurs] / M. A. Ragimova// XII Mezhdunarod. konf. «Crimea–2005: Biblioteki i informatsionnye resursy v sovremennom mire nauki, kultury, obrazovaniya i biznesa» (04; 2005; Sudak) : sb. materialov konf. – Elektron. dan. – Rezhim dostupa: http://www.gpntb.ru/win/inter-events/crimea2005/disk/203.pdf. – Zagl. s ekrana.*

6. **Руководство** по каталогизации в Сводном каталоге электронного издания «Open for you» для категории «Библиотека – пользователь СК» [Электронный ресурс] / ООО «ЭйВиДи-систем». – Электрон. дан. – Режим доступа: http://www.open4u.ru/doc/doc/Руководство для Библиотеки-пользователя СК.pdf. – Загл. с экрана.

Rukovodstvo *po katalogizatsii v Svodnom kataloge elektronnogo izdaniya «Open for you» dlya kategorii «Biblioteka – polzovatel SK» [Elektronnyy resurs] / OOO «EyViDi-sistem». – Elektron. dan. – Rezhim dostupa: http://www.open4u.ru/doc/doc/Rukovodstvo dlya Biblioteki-polzovatelya SK.pdf. – Zagl. s ekrana.*

7. **НЭБ**: магистральное направление развития библиотечной отрасли [Электронный ресурс] / Г. П. Ивлиев [и др.] // Унив. кн. – Электрон. дан. – 2015. – № 1/2. – С. 46–54. – Режим доступа: http://www.unkniga.ru/biblioteki/bibdelo/4031-neb-magistralnoe-napravlenie-razvitiya-bibliotek.html. – Загл. с экрана.

NEB: *magistralnoe napravlenie razvitiya bibliotechnoy otrasli [Elektronnyy resurs] / G. P. Ivliev [i dr.] // Univ. kn. – Elektron. dan. – 2015. – № 1/2. – S. 46–54. – Rezhim dostupa: http://www.unkniga.ru/biblioteki/bibdelo/4031-neb-magistralnoe-napravlenie-razvitiya-bibliotek.html. – Zagl. s ekrana.*

8. **Размещение** электронного каталога в Сводном каталоге электронных ресурсов [Электронный ресурс]: соглашение об информационной услуге / Автономная некоммерческая организация «Национальный информационно-библиотечный центр ЛИБНЕТ». Приложение 1 : регламент. – Электрон. дан. – Режим доступа: http://www.nilc.ru/nilc/documents/agreement_sker.zip. – Загл. с экрана.

***Razmeshchenie** elektronnogo kataloga v Svodnom kataloge elektronnyh resursov [Elektronnyy resurs]: soglashenie ob informatsionnoy usluge / Avtonomnaya nekommercheskaya organizatsiya «Natsionalnyy informatsionno-bibliotechnyy tsentr LIBNET». Prilozhenie 1 : reglament. – Elektron. dan. – Rezhim dostupa: http://www.nilc.ru/nilc/documents/agreement_sker.zip. – Zagl. s ekrana.*

9. **Степанов В. К.** НЭБ как платформа интеграции библиотек в систему цифровых коммуникаций, или Какая национальная электронная библиотека нужна России [Электронный ресурс] / В. К. Степанов // XXI Международ. конф. «Крым–2014: Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» (21; 2011; Судак) : сб. материалов конф. – Электрон. дан. – Режим доступа: http://www.gpntb.ru/win/inter-events/crimea2014/disk/088.pdf. – Загл. с экрана.

***Stepanov V. K.** NEB kak platforma integratsii bibliotek v sistemu tsifrovyh kommunikatsiy, ili Kakaya natsionalnaya elektronnaya biblioteka nuzhna Rossii [Elektronnyy resurs] / V. K. Stepanov // XXI Mezhdunarod. konf. «Crimea–2014: Biblioteki i informatsionnye resursy v sovremennom mire nauki, kultury, obrazovaniya i biznesa» (21; 2011; Sudak) : sb. materialov konf. – Elektron. dan. – Rezhim dostupa: http://www.gpntb.ru/win/inter-events/crimea2014/disk/088.pdf. – Zagl. s ekrana.*

10. **Фронкин А. В.** Управление процессами информационного обмена в распределённых библиотечных системах регионального уровня [Текст] : автореф. дис. … канд. техн. наук: 05.13.10 / А. В. Фронкин ; Науч.-исслед. центр проблем качества подготовки специалистов. – Великий Новгород, 2001. – 19 с. – Библиогр.: с. 18–19 (11 назв.).

***Fronkin A. V.** Upravlenie protsessami informatsionnogo obmena v raspredelennyh bibliotechnyh sistemah regionalnogo urovnya [Tekst] : avtoref. dis. … kand. tehn. nauk: 05.13.10 / A. V. Fronkin ; Nauch.-issled. tsentr problem kachestva podgotovki spetsialistov. – Velikiy Novgorod, 2001. – 19 s. – Bibliogr.: s. 18–19 (11 nazv.).*

11. **Шрайберг Я. Л.** ИС ЭКБСОН: проект сдан и готов к эксплуатации [Электронный ресурс] / Я. Л. Шрайберг // Унив. кн. – Электрон. дан. – 2014. – № 1/2. – С. 52–55. – Режим доступа: http://www.unkniga.ru/biblioteki/bibdelo/2557-is-ekbson-proekt-sdan-i-gotov-k-ekspluatatsii.html. – Загл. с экрана.

***Shrayberg Ya. L.** IS EKBSON: proekt sdan i gotov k ekspluatatsii [Elektronnyy resurs] / Ya. L. Shrayberg // Univ. kn. – Elektron. dan. – 2014. – № 1/2. – S. 52–55. – Rezhim dostupa: http://www.unkniga.ru/biblioteki/bibdelo/2557-is-ekbson-proekt-sdan-i-gotov-k-ekspluatatsii.html. – Zagl. s ekrana.*

***Kirill Sokolinsky,** Head of Information and Education Resource Center, Bonch-Bruevich St. Petersburg State University for Telecommunications, ELNIT Association programmer;*

sokolinsky_k_e@mail.ru

2 Bolshevikov prospekt, St. Petersburg, 193232, Russia