*2001 - Том 4 - Выпуск 4*

# Ретроспективная конверсия каталогов. Организация и технология

*Лаврёнова О.А.*
*Российская государственная библиотека, Москва*
*Адрес: 101000, Москва, ул. Воздвиженка, 3/5*

Рассмотрены вопросы подготовки и ведения проектов ретроконверсии каталогов, формирования запроса на деловые предложения от потенциальных подрядчиков и требований к качеству работ, выбора технологии и исполнителей, а также контроля качества.

В докладе использованы материалы, подготовленные РГБ, отечественными и иностранными фирмами (в частности, консультантом из французской фирмы Jouve Мари - Элиз Фреон, московскими фирмами “ГИПЕР” и “ПроСофт-М”) по проекту “Создание информационной системы для Российской государственной библиотеки”, финансируемого в рамках программы TACIS. Дело в том, что одним из направлений работ по этому гранту, полученному в РГБ на 1,5 года, является пилотный проект ретроконверсии каталогов, реализованный по всем правилам на массиве 10000 библиографических записей из сводного каталога русской книги 1826-1917 гг.

Известны различные сочетания методов ретроконверсии каталогов, существующих в карточной и книжной форме:

- перевод в машиночитаемую форму библиографических записей непосредственно с карточек или из книжного варианта:
  - полная конверсия в формат MARC с ручным набором линейного текста (текстовых данных) и разметкой полей MARC-формата (производится сразу в процессе работы, или коды вписываются в записи на карточках для последующего ввода с клавиатуры);
  - ввод с клавиатуры текстовой информации и создание структуры электронных записей автоматически (распознавание знаков препинания, последовательности символов,использование специальных словарей и лексиконов);

- сканирование карточек с целью создания их машинных (цифровых) образов с последующим созданием машиночитаемых библиографических записей:
  - в случае обработки отпечатанных каталожных карточек хорошего качества - оптическое распознавание символов и приведение записей к стандартному формату базы данных вручную или автоматически;
  - ввод текстовой информации с клавиатуры в поля MARC-формата на основе образов карточек на экране дисплея;
  - использование библиографических записей из других электронных

каталогов, приобретаемых на оптических дисках или получаемых по сетям) и доввод собственных данных библиотеки (индексов классификации, предметных рубрик, шифров хранения и т.д.);
- предварительное изготовление копий библиографических карточек на микрофишах, их сканирование с использованием специальных сканеров и последующая ручная или программная обработка, как описано выше.

Часто используются смешанные технологии, когда метод преобразования записей в машиночитаемые зависит от качества соответствующего фрагмента каталога, и, разумеется, в одной и той же библиотеке различные технологии подбираются отдельно для каждого конкретного каталога.

Выбор той или иной технологии и соответствующего подрядчика для ретроконверсии каталогов зависит преимущественно от качества последних.

Ручной набор записей с клавиатуры требует огромных затрат труда и времени, особенно при ретроконверсии каталогов больших библиотек. Однако этот метод довольно распространен. Привлекаются для этого студенты, школьники. Надо сказать, что многие крупные зарубежные библиотеки так же проводили ретроконверсию несколько лет назад, в частности, отыскивая дешевую рабочую силу (например, студентов, исполнителей из других стран или даже операторов-монахов). Обычные нормы при данной технологии – 40-50 каталожных карточек в день на одного человека, но производительность существенно возрастает при правильной организации труда и приобретении опыта операторами.

В последние годы за рубежом наиболее распространено сканирование карточных каталогов с использованием специальных высокопроизводительных сканеров, приспособленных для автоматической подачи каталожных карточек из пачек, которые закладываются в приемные устройства. В результате сканирования в памяти компьютера формируются массивы факсимильных образов лицевых сторон карточек и, как правило, их оборотных сторон. Современные скорости сканирования очень высоки. Важным является то обстоятельство, что данный способ копирования карточек каталога решает первую проблему – обеспечение сохранности библиографической информации, зафиксированной в каталоге. Более того, далее можно обрабатывать уже не карточки из ящиков, которые не хотелось бы куда-то перемещать из зоны каталогов по различным соображениям, а их образы на машинных носителях.

Понятно, что образ карточки – это картинка, которую можно хранить и смотреть на экране, но нельзя обрабатывать как текст. Другая серьезная проблема – невозможность поиска библиографических записей в массиве образов без создания каких-либо поисковых признаков, внесенных в компьютерную систему в форме текста. Поэтому нужно связать в машине образ карточки с какой-то записью из букв, цифр и других знаков. Простейший вариант – поставить в соответствие каждой карточке начальную букву или несколько букв алфавита из заголовка описания, как в карточном каталоге, но, разумеется, этого маловато для возможностей электронного каталога, учитывая затраты, вложенные в сканирование.

Можно пойти дальше: сформировать в машиночитаемом виде для каждого образа карточки (учитывая образ оборота) несколько наиболее важных поисковых полей элементов данных. Например, на первое время вполне достаточно ограничиться полями индивидуальных и коллективных авторов, заглавий, года издания. По мере появления средств работа в любое время может быть продолжена. Идеальный вариант – получение полной БЗ в формате представления элементов записей (как правило, в формате типа MARC).

В последние годы метод формирования элементов машиночитаемых библиографических записей по образам карточек для доступа к данным в ЭК (keying on image - koi) получает все большее распространение как наиболее рациональный и экономичный для конверсии старых каталогов. Суть такой работы заключается в том, что оператор видит на экране одновременно образ каталожной карточки (а затем – и ее оборота) и макет ввода элементов библиографической записи, основанный на MARC-формате. Глядя на образ карточки, оператор вносит с клавиатуры данные одно за другим в соответствующие поля формата. Как правило, фирмы, занимающиеся такого рода деятельностью, стараются использовать разделение труда и узкую специализацию своих операторов: каждый из них досконально осваивает и заполняет одно-два поля с их подполями или определенную область библиографической записи. Так, от оператора к оператору формируется БЗ, а затем проходит ряд стадий программного контроля и полной или выборочной проверки человеком – библиографом. Приложение 5 содержит изображение экрана, на котором слева располагается образ каталожной карточки, а справа – макет для ввода данных в поля формата (пример компании Жув, Франция).

Разумеется, методы программного распознавания образов знаков нашли применение и в ретроконверсии каталогов. В общих чертах они работают следующим образом: в машинную программу закладываются образы всевозможных знаков, которые могут предположительно встретиться в данном типе документов, и их различные вариации. Специальная программа распознавания образов считывает образ страницы документа (например, образ карточки) и сравнивает все, что там встретит, с имеющимися в ее памяти образами знаков. В качестве результата программа выдает на экран “расшифрованный” ею текст. В нем обозначаются знаки, в распознавании которых программа не уверена. Понятно, что часть “картинок” она может вовсе не распознать.

В частности, при обработке карточек относительно приличного качества или, например, каталогов в книжной форме, и при использовании сильного программного обеспечения с тщательно продуманными справочниками, словарями, действительно, получается очень серьезная экономия времени и средств. Однако достаточно вспомнить качество наших каталогов, где на карточках есть и машинописный, и рукописный текст, с помарками, нечеткий, то нетрудно представить себе и без специальной подготовки в области компьютерных технологий, насколько сложным и дорогим может оказаться программирование эталонов знаков на все эти случаи из библиотечной практики. Кроме того, считается, что после программного распознавания знаков процент ошибок даже на неплохом материале обычно составляет не менее 2-5 %. Поэтому

суммарные затраты на обработку карточек могут намного превысить затраты на ручной ввод.

В качестве примеров использования различных методов работы можно привести две отечественные фирмы, имеющие серьезный опыт в ретроконверсии карточных каталогов. С одной стороны, предлагается максимальная автоматизация процессов распознавания графики знаков, разнесения по полям MARC-формата, контроля грамотности, минимальное участие человека в процессе ретроконверсии. На этой позиции стоит фирма “ГИПЕР”, которая преобразует таким образом в электронную форму каталоги Всероссийской государственной библиотеки иностранной литературы (ВГБИЛ). Необходимое качество автоматизированной обработки образов карточек достигается фирмой в основном тщательной настройкой системы распознавания и использованием разнообразных программ контроля получаемых данных. Это и файлы-эталоны графического изображения символов, и справочники, и классификаторы для правильного разнесения информации по полям формата и проверки результата, а также лингвистические справочники требуемого языка для контроля правописания (“грамотности”), справочники ключевых и стоп-слов, которые позволяют определить области информации на карточке для внесения в конкретные поля MARC-формата, и т.д.

С другой стороны, фирма “ПроСофт-М” при ретроконверсии карточных каталогов библиотек ориентируется на свой высококвалифицированный персонал, формирующий библиографические записи путем ручного заполнения полей на основе изображений карточек на экране, на последующий тотальный и очень жесткий автоматизированный контроль технологии, автоматизацию отдельных (рутинных) процессов, максимально комфортные условия труда и его специализацию. В настоящее время данная фирма ведет в РГБ ретроконверсию сводного каталога русской книги 1826-1917 гг., демонстрируя высокое качество результатов.

Речь идет о целесообразности выбора той или иной технологии и соответствующего подрядчика для ретроконверсии каталогов определенного качества. Методы обеих фирм РГБ тестировала в процессе ведения пилотного проекта ретроконверсии части сводного каталога русской книги 1826-1917 гг. и убедилась в их ответственном отношении к работе. Обе фирмы не используют технологию прямого клавиатурного ввода данных с бумажных карточек. Промежуточным этапом их работы является получение машинных (цифровых) изображений карточек путем сканирования на специализированных документных сканерах. Только потом машиночитаемые записи формируются на основе образов карточек по различным технологиям. Однако качество старых карточек нашего сводного каталога не оставляет никакой надежды на автоматическое распознавание знаков, так что именно технология, предложенная “ПроСофт-М”, соответствует данной задаче. Кстати, аналогичные методы могла предложить РГБ фирма “Жув” (Франция), но остановились на отечественной фирме по вполне понятным соображениям. При этом приятно было осознавать, что выбор опирается на оценку работы фирм по международным требованиям.

Успех ретроконверсии каталога, как и любой другой серьезной работы, во многом

зависит от подготовки к ней и управления технологическими процессами. Так, особенно важно организовать исследование объекта автоматизации, четко разработать постановку задачи, предварительно подготовить к вводу в машину каталог, выявить имеющиеся где-либо электронные копии его фрагментов, выбрать методы работы и исполнителей)

Ретроспективная конверсия каталогов, существующих в карточной или печатной (книжной) форме, в электронную форму проводится для обеспечения их доступности вне зависимости от местонахождения пользователя и времени обращения, обеспечения новых возможностей поиска и сохранности библиографических данных. В настоящее время РГБ, как и ряд других библиотек, занимается такими проектами, как “Создание электронной библиотеки” и “Электронная доставка документов”. Надо отметить, что отсутствие полного электронного каталога в качестве справочно-поискового аппарата для электронных массивов документов и в качестве средства удаленного поиска библиографических записей для заказа документов на различных носителях вносит в реализацию этих проектов дополнительные трудности. Приходится по ходу подготовки электронных копий документов из фонда библиотеки формировать их машиночитаемые библиографические записи (БЗ), если они пока отсутствуют в электронном каталоге (ЭК). Это обстоятельство еще ярче подчеркивает неотложность решения проблемы ретроконверсии.

Ретроконверсия каталогов осуществляется обычно поэтапно как силами самой библиотеки, так и с привлечением других организаций на базе внедрения наиболее прогрессивных информационных технологий в координации с ведущими библиографирующими учреждениями России.

Прежде чем планировать ретроконверсию каталогов необходимо правильно оценить имеющиеся кадровые и финансовые ресурсы, точно определить цели и контролировать процесс на основе жестких инструкций и графиков. Кроме того, необходимо убедиться, что не представляется возможным и экономичным использовать соответствующие машиночитаемые записи из имеющихся электронных каталогов или иных электронных ресурсов.

Для нормальной организации ретроконверсии необходимо получать гарантированное регулярное ежегодное финансирование. Этапность работ (выбор каталога) зависит не только от необходимости его предоставления пользователям, но и от реальности финансирования работы до ее завершения, чтобы не допускать образования множества незавершенных работ.

В тех случаях, когда нельзя быть уверенным в регулярности финансирования ретроконверсии большого каталога, рекомендуется выбирать для обработке какие-либо части, которые в определенном смысле можно считать законченными. Например, можно отобрать издания определенной группы авторов (русские писатели, европейские писатели), конкретных видов коллективных авторов (вузы, государственные органы), что вполне реально сделать в порядке алфавита, издания по определенной тематике (если это систематический каталог). При таком решении вопроса об этапности ретроконверсии работа, которую удастся завершить, будет представлять самостоятельный интерес для пользователей до

окончания обработки всего каталога, причем такого рода части каталога редких изданий или публикаций по актуальной тематике могут рассматриваться как коммерческий продукт, который библиотека реализует для продолжения дальнейших работ.

В процессе проведения ретроконверсии должны быть организованы следующие основные работы:

- выбор каталога, описание структуры и потока данных, элементов библиографической записи и формата их представления;
- изучение возможностей использования машиночитаемых записей из других библиотек или иных источников;
- выбор оптимальной организации работ, наиболее выгодной технологии, программного и технического обеспечения;
- определение исполнителей и порядка финансирования работ, графика их проведения;
  приобретение техники;
- подготовка каталога к ретроконверсии;
- разработка инструкций по формированию машиночитаемых библиографических записей на основе традиционных, подготовка словарей, кодификаторов для некоторых элементов записей; непосредственный перевод содержания карточек каталогов на машиночитаемые носители, контроль качества, редактирование;
- обеспечение доступа к базе данных в локальной сети и в Internet.

Можно выделить следующие общие принципы проведения ретроконверсии каталогов:

- информация вводятся в том виде, как она задана на оригиналах карточек;
- в машинную запись не вносится никакой дополнительной информации;
- можно получить некоторые элементы машиночитаемой записи из имеющихся в традиционном варианте данных (это не ретроспективная каталогизация и не “перекаталогизация”, а преобразование записей в другую форму - электронную);
- рекомендуется структурировать информацию в соответствии с международными стандартами.

Большое значение имеет определение целей на стадии предпроектного

обследования, четкая постановка задачи.

В частности, в самом начале принимаются решения, целесообразно и возможно ли расчистить предназначенный для ретроконверсии каталог или конвертировать все записи, стоит ли провести инвентаризацию фондов, чтобы не конвертировать карточки на те книги, которых уже нет в фондах библиотеки.

Важнейшим результатом предпроектного обследования должно стать подробное описание системы каталогов библиотеки: соответствие каталогов фондам, типы каталогов (алфавитные, предметные, систематические, топографические), вид карточек в каталогах (рукописные, отпечатанные на принтере, отпечатанные на пишущей машине, карточки с наклеенными надписями).

В мировой практике приоритетными для проведения ретроконверсии считаются:

- каталоги - источники наиболее полной и точной информации,
- каталоги, отражающие наиболее важные или наиболее часто используемые читателями части фондов,
- каталоги, имеющие наибольшее количество пересечений с другими (если одни и те же книги, отражаются в разных каталогах библиотеки).

Следующей работой, определяющей успех ретроконверсии, следует признать подготовку выбранного каталога к вводу в компьютер.

При этом обычно проводятся отбор карточек для конверсии, а также подготовка информации, содержащейся на каталожных карточках, к вводу в компьютер. Формирование библиографических записей может проводиться по основным карточкам, если они содержат всю информацию (дополнительные записи или заглавия, наличие шифров всех копий документа); в противном случае обрабатываются все карточки, но при этом необходима очистка базы данных от повторов, что приводит к удорожанию работ.

На карточках необходимо вычеркнуть всю информацию, которую не нужно будет вводить в базу данных, или составить список такой информации в инструкциях по проведению работ (например, ссылки на другие каталоги, инициалы каталогизаторов и т.д.). Следует также проверить, не дублируется ли та же самая информация на карточке в нескольких местах (например, разные формы имени автора и т.д.).

Если ретроконверсии подвергаются печатные каталоги, необходимо:

- определить, собираются ли библиографические записи по определенной схеме (вычеркнуть повторные названия; если иерархическая структура не задана — пронумеровать их);
- если в библиографических записях встречаются перекрестные ссылки,

определить, нужно ли вводить их в базу данных;

- если будут вводиться не все библиографические записи, вычеркнуть ненужные, проанализировать указатели: общий указатель для ряда данных: имена индивидуальных авторов, имена коллективных авторов, предметные рубрики, наименования произведений анонимных авторов (разработать систему кодирования для идентификации каждого вида данных);
- определить, можно ли использовать какие-либо номера (коды, шифры) для связи с библиографическими записями в качестве отличительных признаков (например, номер записи или шифр хранения) и по необходимости дополнить их.

Для всех каталогов полезно собрать образцы разных шифров хранения (полочных индексов), используемых в библиотеке.

Затем следует описать структуру записей и определить формат представления элементов записей. Естественно, при этом необходимо учитывать формат, используемый в других электронных каталогах библиотеки.

После проведения подготовительных работ можно выбирать исполнителя. Разумеется, многие библиотеки проводят ретроконверсию своими силами, но международный опыт показывает, что более целесообразно поручить ее профессионалам в данной области деятельности. Однако, будь то сторонние исполнители или собственные, необходима четкая постановка задачи, оформленная в виде утвержденного руководством библиотеки документа (проекта), содержащего требования к результату. В международной практике такой документ часто называют “спецификацией”.

Если библиотека планирует поручить работу специализированной фирме, обычно полагается провести конкурс (тендер) на лучшие предложения от возможных исполнителей, но в любом случае нужно убедиться, что будущий подрядчик понимает задачу и способен ее выполнить в приемлемые сроки в соответствии с требованиями, содержащимися в спецификации, и с оптимальным соотношением “цена – качество”.

В документе, который представляется потенциальным подрядчикам в качестве постановки задачи, требуется описать:

- свою библиотеку как объект автоматизации (статистические данные о фондах, читателях, информационных потоках и т.д.);
- каталоги, подлежащие ретроконверсии;
- библиографическую информацию, представленную на карточках, способ ее представления;
  поля, заполняемые с помощью информации из других полей или справочников (например, коды стран, коды языков);

- исключения;
- требуемый уровень качества (допустимое количество ошибок на определенный массив записей, количество ошибок, при котором массив нужно будет вводить заново);
- любые виды дополнительной обработки конвертированных записей, если необходимо;
- обязательные процессы: тестирование загрузки, тестирование ввода и т.д.; требуемые форматы и стандарты.

К документу необходимо приложить образцы карточек из каталога.

Рекомендуется также предложить фирме провести тестовую обработку массива данных из каталога библиотеки и на ее основе проверить способность исполнителя выполнить предъявленные требования.

Контроль качества больших массивов данных обычно производится на основании случайной выборки, если сплошная проверка качества получаемого массива БЗ не представляется возможной. В ISO 2859 описывается план проведения выборки для проверки массивов данных (массивы выделяются на основе объема отконвертированных данных и времени работы), а также принципы определения допустимого уровня качества (процент записей с ошибками от общего количества записей). Разумеется, чем выше требуемый уровень качества, тем выше стоимость работ.

Если при проверке выборки выясняется, что уровень качества не соответствует требованиям, должен быть составлен официальный документ с перечислением типов обнаруженных ошибок и примерами для каждого типа ошибок, а массив карточек передан снова на обработку. После повторного ввода массива придется еще раз провести выборочную проверку.

Для проведения ретроконверсии необходимо подготовить четкие и детальные инструкции: для каждого поля следует описать конвертируемые данные, их представление на карточках и правила их перевода в электронную форму.

В частности, требуется описать:

- правила ввода символов;
- признаки определения лишней информации, которая встречается на карточках;
- правила распределения информации по полям и подполям главным образом на основании формальных признаков;

- правила стандартного преобразования шифров, кодов в машиночитаемую форму (например, шифров хранения);

- правила работы в специфических случаях (периодические издания, многотомники и т.д.).

В инструкциях приводятся примеры для иллюстрации заполнения разных полей и несколько образцов полных записей с приложением оригиналов соответствующих карточек.

Параллельно с ретроконверсией полезно заняться формированием нормативных/авторитетных записей для имен авторов, наименований коллективов, географических названий и т.д. и провести контроль БЗ по таким записям или по имевшимся ранее в библиотеке файлам нормативных/авторитетных записей. Нередко параллельно с ретроконверсией производится нанесение штрих-кодов на единицы хранения.

Таким образом, библиотека передает выбранной фирме-исполнителю библиографические записи на каталожных карточках (это называется “информацией на входе системы”). На выходе технологического процесса, обеспечиваемого фирмой, библиотека получает не только картинки (графические образы) карточек, но и файлы (массивы) библиографических записей в машиночитаемой форме, в которых каждый знак как бы набран с клавиатуры. И это еще не все: элементы библиографической записи распределяются по полям определенного формата представления элементов таких записей (для РГБ использовался формат USMARC).

Данное выше краткое описание технологий демонстрирует, в частности, насколько непростым делом является профессиональный процесс ретроконверсии каталогов, требующий использования дорогих программных продуктов и высококлассных технических средств, которые нужно постоянно поддерживать в рабочем состоянии. Поэтому большинство крупных библиотек в мире предпочитают поручать эту работу специализированным фирмам, а другие библиотеки стараются опереться на готовые библиографические базы данных.

---

**Лаврёнова Ольга Александровна** - канд. филологических наук по специальности “Структурная, прикладная и математическая лингвистика”; ученое звание – ст.н.сотр. Работает зав. научно-исследовательским отделом развития компьютерных технологий и лингвистического обеспечения в Российской государственной библиотеке. Имеет более 80 публикаций по компьютерной лингвистике, созданию информационных систем, в том числе – информационно-библиотечных, проблемам представления знаний в связи с автоматизацией информационных процессов. Руководила целым рядом проектов в указанных областях деятельности. В рамках описываемого в статье проекта РГБ/TACIS выполняла, в частности, обязанности координатора подпроекта ретроспективной конверсии карточного каталога русской книги 19 века.

Российская государственная библиотека - ведущая библиотека Российской Федерации, вторая в мире и первая в Европе по объему фонда. Ее фонд составляет более 42,2 млн. единиц хранения. Библиотека собирает и постоянно хранит отечественные и иностранные издания с начала книгопечатания и до настоящего времени на всех живых, мертвых и искусственных языках, а также располагает богатейшим собранием рукописных материалов начиная с VI века. РГБ принимает до 7 тыс. посетителей, предоставляет их пользование более 30 тыс. изданий ежедневно. РГБ поддерживает электронные каталоги и другие информационные ресурсы на своем сайте в Internet (адрес – http://www.rsl.ru).

Лавренова О.А. ©