

Организация и технология ретроспективной конверсии каталогов

О.Л. Лавренова

Российская государственная библиотека

Мари-Элиз Фреон

компания "Жув"

Доклад подготовлен совместно сотрудником РГБ и специалистом французской фирмы Jouve ("Жув") в рамках проекта "Создание информационной системы для Российской государственной библиотеки", финансируемого TACIS. Одним из направлений работ по этому гранту является пилотный проект ретроконверсии каталогов. В работе рассматриваются общие принципы проведения ретроконверсии каталогов, организация предпроектного обследования объекта ретроконверсии, процессы подготовки каталогов к ретроконверсии, вопросы управления ею, включая разработку требований, выбор фирмы-исполнителя и контроль качества результатов от сканирования карточек до получения библиографических записей в MARC – формате; проводится анализ известных методов ретроконверсии каталогов.

1. Проблема ретроконверсии традиционных каталогов в электронную форму останется **чрезвычайно актуальной** до тех пор, пока каталоги крупнейших библиотек не будут преобразованы в электронные. Затем процессы ретроконверсии каталогов других библиотек станут простым процессом отбора нужных записей и дополнения их локальной информацией.

Фонды крупнейших российских библиотек представляют собой важнейшую составную часть библиотечного фонда страны. Например, Генеральный алфавитный каталог Российской государственной библиотеки отражает отечественные документы, документы стран СНГ, а также богатейшие собрания зарубежных изданий на всех языках мира. **В каталогах содержатся** уникальные сведения о редких и ценных печатных и рукописных книгах начиная с 15 века. Объем только алфавитных каталогов составляет 15 млн. карточек, что несравнимо ни с какой другой библиотекой России. Кроме того, РГБ поддерживает ряд сводных каталогов, отражающих отдельные виды изданий из фондов нескольких крупных библиотек (например, сводный каталог русской книги 1826 - 1917 гг.). Только в РГБ ведутся общероссийские сводные каталоги зарубежных нотных и картографических изданий. **Совокупность каталогов главной библиотеки страны на 90-95% отражает по названиям национальный репертуар печати и состав единого библиотечного фонда России.**

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 **актуальность проблемы ретроконверсии**. Разумеется, те библиотеки, которые приступают к внедрению подсистем обслуживания читателей, также столкнутся с вполне определенными проблемами при регистрации в системе статуса единиц хранения в процессе их циркуляции в библиотеке для тех изданий, библиографические записи которых еще отсутствуют в электронном каталоге.

2. Основные обстоятельства, которые определяют важность задачи ретроконверсии:

- Единственный в своем роде справочный аппарат РГБ, как и каталоги большинства ведущих библиотек страны, существует в основном в карточном виде, и, следовательно, физически невозможно **обеспечить его сохранность**. Утрата карточки в алфавитном каталоге может означать потерю пути к этому документу на полке навсегда.
- В электронной (машиночитаемой) форме каталоги могли бы **стать доступными теоретически любому человеку в различных регионах России и зарубежья**, причем с его служебного или домашнего компьютера.
- Кроме того, поиск в карточном каталоге или картотеке может осуществляться только в одном направлении в зависимости от порядка расстановки карточек, в то время как в машинной базе данных пользователю предоставляется возможность **найти информацию одновременно по различным характеристикам**.

Другие обстоятельства не столь существенны.

Итак, ретроспективная конверсия каталогов в электронную форму проводится для обеспечения сохранности и широкой доступности хранящейся в них библиографической информации, а создание ЭК ведущих библиотек должно послужить основой для ретроспективной конверсии каталогов других библиотек России.

3. Доклад подготовлен совместно сотрудником РГБ и специалистом французской фирмы Jouve ("Жув") в рамках проекта "Создание информационной системы для Российской государственной библиотеки", финансируемого TACIS. Одним из направлений работ по этому гранту является **пилотный проект ретроконверсии каталогов**. Он успешно реализован на массиве 10000 библиографических записей из сводного каталога русской книги 1826-1917 гг. Может показаться, что по отношению к общему объему данного

каталога (600 тыс. записей) это не очень большая часть работы. Однако, **вопросы организации ретроконверсии, подготовки к ней, выбора фирмы - исполнителя прорабатывались рабочей группой, состоящей из специалистов РГБ и зарубежных консультантов, для задачи в целом.** В этом отношении пилотный проект сыграл очень важную роль. Например, полезно не только получить информацию об используемых в мире методах, обсудить и спланировать работы с опытными в конкретном деле специалистами, но и убедиться, что прежние разработки библиотеки и ранее выбранные партнеры достойно выглядят в глазах признанных международных экспертов.

Сразу отметим, что ретроконверсией каталогов в данном проекте, как это нередко встречается в зарубежной практике, считается не только преобразование традиционных каталогов в электронную форму, но и конверсия библиографических записей из существовавших ранее систем в одну новую.

4. Процесс ретроконверсии каталогов включает следующие основные задачи:

- выбор каталога, описание структуры и потока данных, элементов библиографической записи и формата их представления;
- выбор оптимальной организации работ, наиболее выгодной технологии, программного и технического обеспечения;
- определение исполнителей и порядка финансирования работ, графика их проведения;
- приобретение техники;
- подготовку каталога к ретроконверсии;
- разработку 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.

5. Общие принципы проведения ретроконверсии каталогов:

- данные вводятся так же, как они представлены 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);
- рекомендуется структурировать информацию в соответствии с международными стандартами.

Организовать процесс ретроконверсии можно несколькими способами. Выбор оптимального метода для конкретной библиотеки зависит от подготовленности и количества сотрудников, размеров финансирования, величины фондов и потребностей библиотеки.

6. Предпроектное обследование объекта ретроконверсии проводится для уточнения цели, оценки возможных трудностей, определения уровня качества, которого библиотека хочет добиться.

Принимаются решения, целесообразно и возможно ли:

- расчистить каталог или конвертировать все записи;
- провести инвентаризацию фондов: нет необходимости конвертировать карточки на те книги, которых уже нет в фондах библиотеки;
- подготовиться к штрих-кодированию до проведения ретроконверсии.

При анализе библиотечных каталогов определяется 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).

Возможен также вариант ретроконверсии печатных каталогов (машинописных или выпущенных типографским способом), а также использование при ретроконверсии карточных каталогов печатных копий их фрагментов.

Целесообразно **начать работу с какой-либо значимой части выбранного каталога**. Если ясно, что ретроконверсия всего каталога не получится в обозримый период времени, можно выбрать более или менее законченную часть, чтобы конечный продукт имел какой-то смысл и, может быть, даже коммерческий вид. Например, конвертирование записей из алфавитного каталога от А до Д (допустим, на большее не достали денег) не даст возможности какого-либо использования получаемого машинного массива, а перевод в машиночитаемую форму записей для индивидуальных авторов или коллективов, круг которых определим заранее, может дать законченную и полезную базу данных. Возможно, удастся ее продать на оптическом диске и заработать средства на продолжение ретроконверсии. Поставив удачный (интересный) фрагмент каталога в Internet, можно рассчитывать на его рекламный характер для получения финансирования для продолжения работ. Вначале предполагалось и в РГБ отобрать по алфавиту, например, записи произведений русских писателей 19 века, но такое решение оказалось неактуальным, поскольку выяснилось, что выигравшая конкурс фирма "ПроСофт-М" планирует сразу же продолжить проект и инвестировать ретроконверсию всего сводного каталога русской книги 1826-1917 гг.

7. Подготовка карточных каталогов заключается в отборе карточек, их обработке.

В частности, следует:

- 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.);

- проверить, не дублируется ли та же самая информация на карточке в нескольких местах (например, разные формы имени автора и т.д.);
- сгруппировать на одной карточке все экземпляры одного издания.

В печатных каталогах следует:

- 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.

8. Управление ретроспективной конверсией включает подготовку запроса на деловые предложения от потенциальных подрядчиков, контроль качества и стандартизацию.

В принципе, полагается провести конкурс (тендер) на лучшие предложения от возможных исполнителей, но и без всякого конкурса нужно непременно убедиться, что будущий исполнитель понимает задачу и способен ее выполнить в приемлемые сроки, с требуемым качеством и с оптимальным соотношением "цена – качество".

9. В документе, который представляется потенциальным подрядчиком в качестве постановки задачи, требуется описать:

- свою библиотеку как объект автоматизации (статистические данные о фондах, читателях, информационных потоках и т.д.);
- каталоги, подлежащие ретроконверсии (количество карточек в каждом, физическая и логическая структура);
- метод конверсии, который выбран (или можно попросить подрядчика обосновать метод который он предпочтет использовать в свете предъявляемых библиотекой требований);
- библиографическую информацию, представленную на карточках, способ ее представления и стандартизации всех полей;
- 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 (например, удаление продублированной информации, стандартизация авторских записей и связей, штрих-кодирование);
- обязательные процессы: тестирование загрузки, тестирование ввода и т.д.;
- требуемую систему и стандарты (ISO 2709, ISO 1001 и т.д.).

К документу необходимо приложить **образцы карточек**.

Полезно также предложить подрядчику **провести тестовую обработку массива данных из каталога библиотеки**.

При ретроконверсии больших массивов данных невозможно проверить визуально каждую запись, поэтому **контроль качества** должен производиться **на основании случайной выборки**.

В ISO 2859 (части 0,2,3,4) описывается план проведения выборки для проверки массивов данных (массивы выделяются на основе объема отконвертированных данных и времени работы, например, количество записей за месяц), а также принципы определения допустимого уровня качества (процент записей с ошибками от общего количества записей).

10. Таким образом и была организована работа по пилотному проекту ретроконверсии в РГБ.

Для проведения ретроконверсии были подготовлены **четкие и детальные инструкции, называемые спецификациями**. Для каждого поля давалось описание преобразуемых данных, их представление на карточках и правила их перевода в электронную форму.

Спецификация содержит:

- правила ввода символов (например, прописные и строчные буквы);
- 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 ввода информации о разных экземплярах издания и их местонахождении;
- правила работы в специфических случаях (периодические издания; многотомники и т.д.).

В инструкции приведены **примеры** для иллюстрации заполнения разных полей и **несколько образцов полных записей с приложением оригиналов соответствующих карточек**.

11. После ретроконверсии рекомендуется провести контроль по нормативным (авторитетным) файлам и простановку штрих-кодов на единицах хранения.

12. Итак, на основании предпроектного обследования и постановки задачи осуществляется **выбор методов ретроконверсии и фирм - исполнителей**.

Существуют различные **сочетания известных методов ретроконверсии каталогов, существующих** в карточной и книжной форме.

Выделим основные из них:

- 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);
- в случае обработки отпечатанных каталожных карточек хорошего качества - оптическое распознавание символов (каталог должен быть заранее переведен в цифровую форму сканированием) и приведение записей к стандартному формату базы данных вручную или автоматически;
- использование библиографических записей из других электронных каталогов;
- предварительное изготовление копий библиографических карточек на микрофишах, их сканирование и их последующая ручная или программная обработка, как описано выше.

Отечественные библиотеки, которые уже приступили к длительному и трудоемкому процессу ретроконверсии, в подавляющем большинстве используют, казалось бы, наименее эффективную и наиболее трудоемкую методику непосредственного ввода текста библиографической карточки в соответствующую базу данных. Информация вводится или непосредственно с бумажных каталожных карточек или реже с графического изображения каталожной карточки, полученной в результате предварительного сканирования каталога, подвергающегося ретроконверсии. Однако, опыт работы с фирмами показал нам, что этот метод непродуктивен в тех случаях, когда, во-первых, тексты на карточках не настолько ужасно выглядят, чтобы их нерентабельно было сканировать с оптическим распознаванием образов знаков и, во-вторых, когда ввод осуществляется мало подготовленными к этой работе людьми без необходимой организации технологического процесса и за очень низкую плату.

Такой подход к решению проблемы даже при экономии на ручном труде требует огромных затрат средств. Низкая производительность ручного ввода информации при описанных условиях доказана и зарубежным, и отечественным опытом ретроконверсии. При этом может быть достаточно высокий процент ошибок. В противовес ручному вводу существует **сканирование карточек с последующим распознаванием графических образов знаков** в тексте. При обработке карточек относительно приличного качества или, например, печатных 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 разным, естественно, почерком, причем

сбоку, наискосок и с помарками, затертый сотнями пальцев, то нетрудно представить 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 на неплохом материале обычно составляет не менее 2-5 %. Поэтому суммарные затраты на обработку и последующую проверку и коррекцию ошибок могут превышать затраты на обработку массива **по технологии ручного формирования машинных записей по на основе изображения карточки** при тех же временных затратах.

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

Два в каком-то смысле противоположных **подхода к технологии ретроконверсии** библиотечных каталогов **были представлены участниками конкурса**, объявленного для пилотного проекта РГБ: **отечественными фирмами "Гипер" и "ПроСофт-М"**.

Итак, с одной стороны, предлагалась 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", имея опыт такой работы. С другой стороны, фирма "ПроСофт-М" при ретроконверсии карточных каталогов наших библиотек ориентируется на высококвалифицированный персонал, формирующий библиографические записи путем ручного заполнения полей на основе изображений карточек на экране, тотальный автоматизированный контроль технологии, автоматизацию отдельных (рутинных) процессов, максимально комфортные условия труда и его специализацию. Обе фирмы, которые были выбраны для участия в конкурсе, не использовали технологию прямого клавиатурного ввода данных с бумажных карточек. **Промежуточным этапом их работы было получение машинных (цифровых) изображений карточек путем сканирования.** Только потом формировались машиночитаемые записи по различным технологиям.

Кроме того, **для сравнения** использовались **данные о технологиях, используемых фирмой "Жув" (Франция).**

После формирования спецификации для выбора компании в мае 1999 года проведена **тестовая конверсия** на массиве из 500 карточек. Тестирование позволило потенциальным подрядчикам продемонстрировать свои методы создания остальных 9500 записей в требуемом формате, а также сообщить примерную стоимость работ. Библиотека и зарубежные эксперты оценивали качество и технологии проведения работ. Для проведения ретроконверсии каталога книг 19-го века с низким качеством текстов на карточках более приемлемыми оказались методы фирмы "ПроСофт-М".

13. Итак, библиотека передает фирме библиографические записи на каталожных карточках (это называется "информацией **на входе** системы"). **На выходе технологического процесса**, обеспечиваемого фирмой, она рассчитывает получить **не картинки (графические образы) карточек, а файлы (массивы) библиографических записей в машиночитаемой форме**, в которых каждый знак как бы набран с клавиатуры. И это еще не все: **элементы библиографической записи должны быть распределены по полям определенного формата представления элементов таких записей**. Преимущественно библиотеки работают в форматах **типа MARC**. В частности, РГБ потребовала от фирм перевести записи в формат USMARC.

Надо полагать, **большинство старых каталогов выглядят так**, как описано в документации для нашего каталога русской книги 19 века:

- **часть информации на карточках напечатана**, а часть **написана от руки** простым или цветным карандашом;
- информация содержится как **на лицевой, так и на оборотной стороне** карточек;
- физические размеры карточек - 75х125 мм, в нижней части карточек находится отверстие.

Было поставлено условие: оригиналы карточек не могут покидать здание РГБ и должны сканироваться на территории библиотеки.

При сканировании образ каждой карточки должен был получить отдельный идентификатор/контрольный номер, чтобы можно **было связать образы лицевой и обратной стороны карточки**.

Весь **технологический цикл ретроспективной конверсии** карточного каталога фирма условно делит на:

- цикл сканирования массива карточек каталога;
- цикл обработки массива графических образов карточек;
- цикл генерации библиографических записей формата MARC.

14. Проводился ежемесячный **контроль** на основе случайной выборки:

- полноты и качества сканирования;
- текстовой информации для выявления орфографических ошибок и опечаток до разбивки информации по полям формата;
- разнесения информации по полям формата MARC.

Размер выборки обычно задается в зависимости от общего количества конвертируемых записей и желаемого уровня качества.

Затем производилась **тестовая загрузка записей** в автоматизированную библиотечную систему, чтобы проверить правильность ввода информации в локальные поля. По результатам тестовой загрузки библиотека высказала свое мнение, а компания внесла соответствующие изменения в рабочие процедуры.

По окончании ввода в библиотеку направлялись **нестандартные случаи** в электронной форме и в виде распечаток.

Кроме того, **в рамках пилотного проекта ТАСИС/РГБ** было обеспечено **программное формирование нормативных записей для имен лиц и наименований коллективов на основе данных, имеющих на карточках.**

15. **Более подробно** вопросы ретроконверсии рассматриваются в материалах, подготовленных Лавреновой О.А. для публикации **в приложении к журналу "Библиотека"** – "С компьютером на "ты" (выпуск ожидается в 2000 г.).

Сведения об авторах



Лавренова Ольга Александровна, заведующая отделом развития компьютерных технологий и лингвистического обеспечения РГБ, канд. филол. наук, координатор пилотного проекта ретроконверсии каталогов Библиотеки в электронную форму в рамках проекта РГБ/TACIS. E-mail: lavr@rsl.ru, тел.: 202-7998.

Фреон Мари-Элиз, специалист французской фирмы Jouve ("Жув"), Франция, эксперт по указанному пилотному проекту.
