Система ретроконверсии каталогов Национальной библиотеки Беларуси

The System of Catalog Retroconversion of the National Library of Belarus, Minsk, Belarus

Система ретроконверсії каталогів Національної бібліотеки Білорусі

М. П. Раковчук, Е. Л. Крейдик
Унитарное предприятие «Агат-Систем», Минск, Беларусь

Mikhail Rakovchuk and Evgeny Kreydik
«Agat-System» Research and Production Republican Unitary Enterprise, Minsk, Belarus

М. П. Раковчук, Є. Л. Крейдік
НП РУП «Агат-Сістем», Мінськ, Білорусь

Для решения задачи ретроконверсии каталогов Национальной библиотеки Беларуси НП РУП «Агат-Систем» была разработана специальная Информационно-технологическая система (ИТСР). В докладе приведен краткий анализ существующих методов проведения ретроконверсии и обоснование принятых технологических решений. Содержатся общие сведения об особенностях реализации основных технологических процессов, технических средств и программных продуктов, предназначенных для решения поставленной задачи.

To accomplish retroconversion of Belarus National Library's catalogs, Agat-System company has designed a special Information Technological System (ITSR). The methods of retroconversion are analyzed in brief, and the technological solutions are justified. General information on the features of main technological processes, hard- and software intended to fulfil the task.

Для вирішення завдань ретроконверсії каталогів Національної бібліотеки Білорусі НП РУП «Агат-Сістем» була розроблена спеціальна Інформаційно-технологічна система (ІТСР). В доповіді приведено короткий аналіз існуючих методів для проведення ретроконверсії та узагальнення прийнятих технологічних рішень. Також містяться свідчення про особливості реалізації основних технологічних процесів, технічних засобів і програмних продуктів, що призначені для вирішення поставленого завдання.

1. Предпосылки разработки системы

Национальная библиотека Беларуси является главной универсальной библиотекой страны, ведущим научно-исследовательским учреждением в области библиотековедения, библиографоведения и книговедения, координационным центром внутригосударственного и международного книгообмена, центром корпоративной каталогизации, а также центром национальной библиографии и библиографической деятельности в области общественных наук.

В связи с вводом в эксплуатацию нового здания Национальной библиотеки Беларуси и внедрения комплекса информационно-технологических систем, составной частью которых является Автоматизированная библиотечная информационная система (АБИС), встал вопрос о создании электронного каталога.

Автоматизация библиотечной деятельности невозможна без наличия такого важного элемента, как электронный каталог, который является ядром библиотечной технологии. Перевод уже имеющейся библиографической информации на традиционном носителе (карточных каталогов Национальной библиотеки Беларуси) в более удобную для использования машиночитаемую форму, т. е. ретроспективная конверсия карточных каталогов, позволит создать принципиально новую среду для информационного и документального обеспечения читателей-пользователей (в том числе и через сеть Internet), повысит полноту и оперативность их обслуживания.

В рамках задания на создание электронного каталога необходимо обработать:

- Генеральный служебный алфавитный каталог (ГСАК) – более 3 млн. карточек;

- Генеральный служебный алфавитный каталог на украинском языке – около 40 тыс. карточек;
- Алфавитный каталог микрофиш диссертаций – 110 тыс.
- Генеральный алфавитный каталог нот (ГАКН) – 120 тыс.
- Генеральный алфавитный каталог изодокументов – более 25 тыс.
- Систематический каталог картографических изданий – более 25 тыс.

В целом ≈ 3, 5 млн. карточек.

Срок выполнения задания – 2 года (ноябрь 2005 г. – октябрь 2007 г.).

Задача ретроконверсионной обработки карточных каталогов Национальной библиотеки Беларуси была возложена на НП РУП «Агат-Систем» вследствие имеющегося большого опыта в области преобразования информации, содержащейся на бумажных носителях в электронный вид, в частности в области дактилоскопической регистрации. Однако задача ретроконверсии карточных каталогов, вследствие специфики предметной области, потребовала разработки специальной информационно-технологической системы. Был проведен анализ известных методов выполнения ретроконверсии в открытых источниках, проведена большая работа с представителями заказчика по определению наиболее сложных и важных аспектов формирования выходных данных требуемого качества, загружаемых в Автоматизированную библиотечную информационную систему (АБИС).

2. Технологии ретроспективной конверсии и их анализ

Известны различные методы ретроконверсии каталогов, существующих в карточной форме, но в целом можно выделить следующие:

1. перевод в машиночитаемую форму (MARC-формат) библиографических записей непосредственно с карточек или из книжного варианта либо ввод текстовой информации с клавиатуры в поля MARC-формата на основе образов карточек на экране дисплея;

2. 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.;

3. сканирование карточек с целью создания их машинных (цифровых) образов, оптическое распознавание символов и приведение записей к стандартному MARC-формату вручную или автоматически.

Метод № 1

Ручной набор записей с клавиатуры требует огромных затрат труда и времени, особенно при ретроконверсии каталогов больших библиотек. Для качественного выполнения работ данным методом требуется привлечение в качестве исполнителей опытных библиографов. Возможно, конечно, использование более дешевой рабочей силы, при условии «разбиения» библиографического описания на отдельные поля и ввод информации одной карточки несколькими «узко специализированными» операторами. Однако существенного увеличения производительности при этом не происходит.

Обычные нормы при технологии ручного ввода – 40–50 каталожных карточек в день на одного человека. Несложно рассчитать время на проведение ретроконверсии при условии использования десяти операторов:

3 500 000 / (10 * 50) = 7000 человекодней (примерно 25 лет).

Естественно абсурдно использовать данный метод для ретроконверсии всех каталогов Национальной библиотеки.

Метод № 2

Технология заимствования библиографических записей из электронных каталогов других библиотек представляется малоперспективной, т. к. анализ информации Web-страниц ведущих библиотек Российской Федерации, где процесс ретроконверсии уже прошел стадию начального этапа, показал, что количе-

ство отраженных в электронных каталогах библиотечных документов не превышает 10% от общих объемов фондов. В то же время, уже имеющиеся электронные каталоги можно рассматривать как коммерческий продукт, стоимость которого значительна, а если учесть необходимость «адаптации», которая также требует определенных затрат, то этот путь экономически нецелесообразен.

Метод № 3

Сканирование карточных каталогов с использованием специальных высокопроизводительных сканеров, приспособленных для автоматической подачи каталожных карточек из пачек, формирование массивов факсимильных образов карточек, использование сильного программного обеспечения с тщательно продуманными справочниками и словарями для распознавания образов карточек и маркировки «расшифрованного» текста маркерами MARC-формата дает очень серьезную экономию времени при проведении ретроконверсии. Данная технология является достаточно сложной, относительно дорогой, но наименее трудоемкой.

3. Выбор технологии ретроспективной конверсии

Выбор той или иной технологии для ретроконверсии каталогов зависит преимущественно от качества последних. Учитывая качество каталогов Национальной библиотеки Беларуси, где на карточках есть и машинописный, и рукописный текст, с помарками, нечеткий, следовало использовать смешанные технологии. Карточные каталоги относительно приличного качества, выполненные печатным способом целесообразно сканировать и выполнять **программное** распознавание текста и маркировку полей (использовать метод № 3), неудовлетворительно распознанный текст обрабатывать вручную, а маркировку полей – программно (метод № 1 и частично метод №3). Все рукописные каталожные карточки обрабатывать вручную (метод № 1).

Таким образом, в результате проведенного анализа, специалистами НП РУП «Агат-Систем» был разработан единый комплекс организационных мероприятий, технологических процессов, технических средств и программных продуктов, предназначенных для решения поставленной задачи, т. е. создана специальная Информационно-технологическая система ретроконверсии (ИТСР).

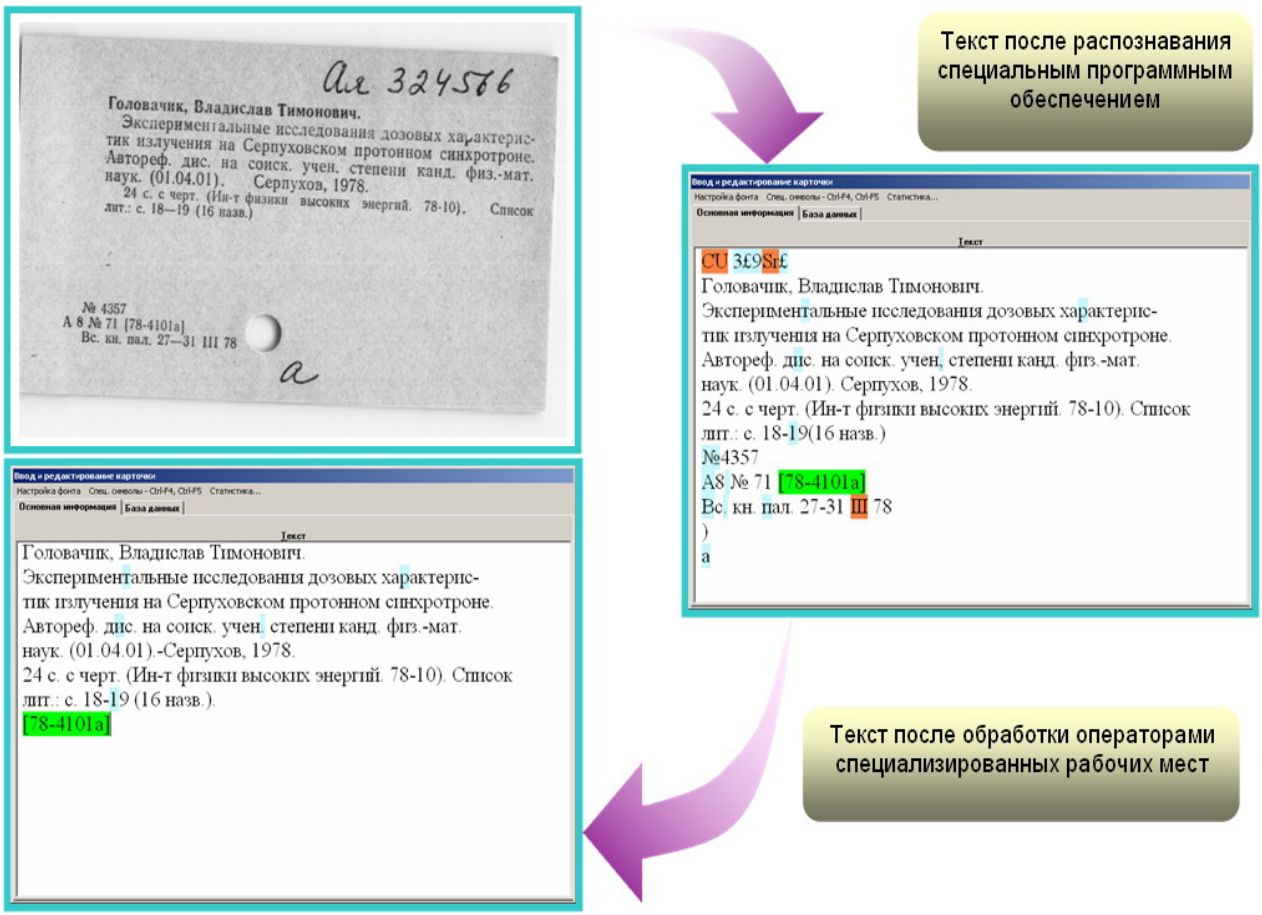
4. Основы информационно-технологической системы ретроконверсии

Информационно-технологическая система ретроконверсии (ИТСР) предназначена для обработки миллионных массивов графических данных, полученных в результате сканирования карточных каталогов Национальной библиотеки Беларуси и формирования выходной информации в формате BELMARC. Информация, полученная в результате ретроконверсии, предназначена для загрузки в Автоматизированную библиотечную информационную систему (АБИС).

Анализ немногочисленных данных выполнения ретроконверсии различными исполнителями показал, что ключевым моментом преобразования каталогов в машиночитаемую форму является задача представления информации с каталожной карточки в виде текста и последующая обработка полученного текста с целью выделения областей библиографического описания и формирования записей в MARC-формате.

Операцию получения текста с карточки можно считать типовой, т. к. для этого существуют специальные пакеты программ, использующие метод распознавания образов и преобразования (или расшифровывания) текста с графического источника.

Но какое бы совершенное программное обеспечение не использовалось для распознавания, встречаются символы, которые распознаются неуверенно (или не распознаются вовсе, т. к. выполнены рукописным способом или качество самой каталожной карточки не позволило получить после сканирования удовлетворительный образ). Поэтому все тексты, полученные после программной «расшифровки» должны проходить стадию ручной обработки. Суть этой обработки заключается в корректировке текстов операторами с помощью специального программного обеспечения.



Дальнейшая обработка текстов с целью выделения областей библиографического описания с помощью специальных меток-маркеров – самая трудоемкая и ответственная стадия. После этой обработки формирование записи в заданном MARC-формате – чисто техническая задача.

Исходя из вышеизложенных особенностей, и была построена Информационно-технологическая система ретроконверсии.

НП РУП «Агат-Систем» имеет серьезный опыт построения сложных информационных систем республиканского значения. Учитывая преимущества хранения данных и результатов их обработки в единой базе данных, ИТСР была построена на базе СУБД Oracle с использованием архитектуры «клиент-сервер», что обеспечивает хранение и обработку значительных объемов информации, интенсивное совместное использование ресурсов, выполнение многоступенчатых циклов по обработке каталожных карточек, выполнения контроля результатов обработки на каждой стадии. Доступ к информации организован в соответствии с полномочиями пользователей.

ИТСР состоит из:

1. Базы данных (БД) для хранения и накопления информации.
2. Комплекса программ, предназначенных для сопровождения базы данных, контроля завершения технологических операций, а также администрирования системы и учета работы персонала.
3. Нескольких видов автоматизированных рабочих мест (АРМ), функционально различных в зависимости от выполняемых работ на соответствующей стадии технологического процесса.

- АРМ корректировки (нормализации) текста;
- АРМ ОТК текста;
- АРМ маркировки текста;
- АРМ ОТК маркировки текста;
- АРМ ОТК назначений (проверка отнесения карточек различного вида в соответствующие информационные потоки);
- АРМ администратора системы.

4. Комплекса программ для автоматизированной обработки текстов каталожных карточек с целью выделения отдельных областей библиографического описания, соответствующих полям Белорусского коммуникативного формата BELMARC. Данный комплекс программ включен в состав АРМ маркировки и существенно снижает трудоемкость данной технологической операции.

5. Комплекса программ для контроля полей библиографических описаний и выгрузки данных в выходные файлы для информационного обмена

6. Программа контроля библиографических записей. Данный модуль не имеет интерфейса с базой данных ИТСР и предназначен для проведения выборочного или сплошного контроля выходных файлов (настраивается опционально).

5. Особенности обработки, реализованные в системе

Основная особенность обработки информации в ИТСР – верификация, т. е. проверка правильности выполнения основных операций (корректировка и маркировка текста). Режим верификации заключается в выполнении двукратной обработки, анализе рассогласования в данных, обработанных двумя исполнителями, и принятия решения для устранения т. н. «возмущающего» воздействия (исправления ошибок).

Следует отметить также заложенные в системе принципы распределения информации при обработке: при корректировке текста информация на рабочие места поступает в случайном порядке, а на самой ответственной операции – маркировке текста (библиографической разметке) – информация на рабочие места поступает в виде ограниченных по количеству карточек блоков данных. В этом случае блок данных представляет собой последовательный набор каталожных карточек соответствующий физическому размещению в ящиках хранения.

5.1. Верификация при нормализации текста

Для обеспечения минимизации ошибок в тексте после распознавания, а также для обеспечения максимальной степени готовности к операции маркировки, корректировка текста выполняется двумя разными операторами с последующим автоматическим посимвольным сравнением данных. При полном сравнении информация записывается в БД и автоматически передается на следующую технологическую операцию, а при наличии рассогласования – решение принимает администратор:

- возврат текста на повторную обработку оператору (операторам) с возможностью отправки дополнительного сообщения или указаний по устранению ошибок;
- самостоятельное устранение рассогласования ввиду его неоднозначности.

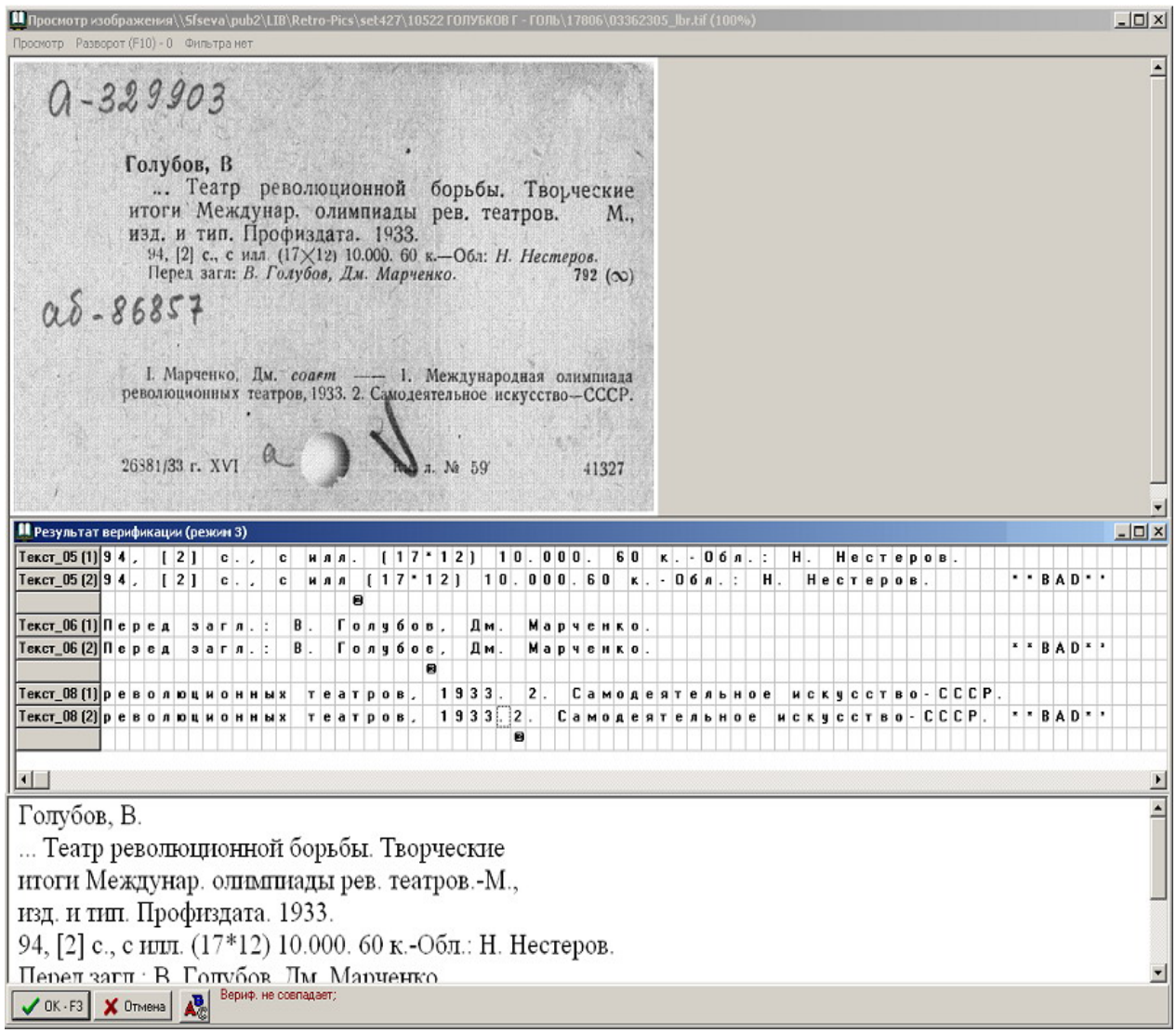
5.2. Модуль автоматической расстановки маркеров

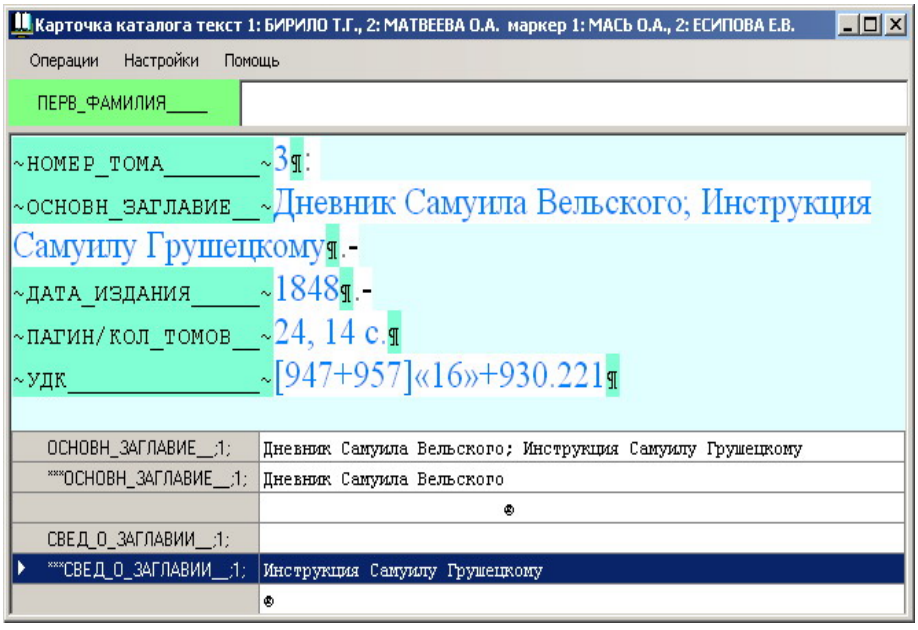
Одна из особенностей операции маркировки – подключение специального программного модуля для автоматизированной расстановки меток-маркеров. Данный модуль подключается автоматически при вызове оператором очередной каталожной карточки на обработку и существенно снижает трудоемкость (80% маркеров проставляется автоматически). Задача оператора – визуальная проверка результатов программной обработки и, если требуется, внесение необходимых корректив.

5.3. Верификация при маркировке текста

Маркировка текста – выделение в тексте карточки областей библиографического описания, соответствующих полям Белорусского коммуникативного формата BELMARC. Выполняется двумя разными операторами с последующим автоматическим сравнением результатов. Сравниваются не только типы проставленных меток-маркеров, но и информация, ограниченная этими маркерами. При полном сравнении информация записывается в БД и автоматически подготавливается к выгрузке, а при наличии рассогласования – решение принимает администратор:

- возврат информации на повторную обработку оператору (операторам) с возможностью отправки дополнительного сообщения или указаний по устранению ошибок;
- самостоятельное устранение рассогласования ввиду его неоднозначности.


Просмотр изображения\\Sfseva\pub2\LIB\Retro-Pics\set427\10522 ГОЛУБКОВ Г - ГОЛЬ\17806\03362305_lbr.tif (100%)
Просмотр Разворот (F10) - 0 Фильтра нет
а-329903
Голубов, В
... Театр революционной борьбы. Творческие итоги Междунар. олимпиады рев. театров. М., изд. и тип. Профиздата. 1933.
94, [2] с., с илл. (17×12) 10.000. 60 к.—Обл: Н. Нестеров.
Перед загл: В. Голубов, Дм. Марченко. 792 (∞)
аб-86857
I. Марченко, Дм. соавт — 1. Международная олимпиада революционных театров, 1933. 2. Самодеятельное искусство—СССР.
26381/33 г. XVI
41327
Результат верификации (режим 3)
Текст_05 (1) 94, [2] с., с илл. (17*12) 10.000. 60 к.-Обл.: Н. Нестеров.
Текст_05 (2) 94, [2] с., с илл (17*12) 10.000.60 к.-Обл.: Н. Нестеров. **BAD**
Текст_06 (1) Перед загл.: В. Голубов, Дм. Марченко.
Текст_06 (2) Перед загл.: В. Голубое, Дм. Марченко. **BAD**
Текст_08 (1) революционных театров, 1933. 2. Самодеятельное искусство-СССР.
Текст_08 (2) революционных театров, 1933. 2. Самодеятельное искусство-СССР. **BAD**
Голубов, В.
... Театр революционной борьбы. Творческие
итоги Междунар. олимпиады рев. театров.-М.,
изд. и тип. Профиздата. 1933.
94, [2] с., с илл. (17*12) 10.000. 60 к.-Обл.: Н. Нестеров.
Перед загл.: В. Голубов, Дм. Марченко
OK - F3
Отмена
Вериф. не совпадает;


Карточка каталога текст 1: БИРИЛО Т.Г., 2: МАТВЕЕВА О.А. маркер 1: МАСЬ О.А., 2: ЕСИПОВА Е.В.
Операции Настройки Помощь
ПЕРВ_ФАМИЛИЯ____
~НОМЕР_ТОМА_____~3¶:
~ОСНОВН_ЗАГЛАВИЕ__~Дневник Самуила Вельского; Инструкция Самуилу Грушецкому¶.-
~ДАТА_ИЗДАНИЯ____~1848¶.-
~ПАГИН/КОЛ_ТОМОВ__~24, 14 с.¶
~УДК_____~[947+957]«16»+930.221¶
ОСНОВН_ЗАГЛАВИЕ__;1; Дневник Самуила Вельского; Инструкция Самуилу Грушецкому
***ОСНОВН_ЗАГЛАВИЕ__;1; Дневник Самуила Вельского
СВЕД_О_ЗАГЛАВИИ__;1;
***СВЕД_О_ЗАГЛАВИИ__;1; Инструкция Самуилу Грушецкому

5.4. Особенности обработки описаний многотомных изданий

Особенности описания многотомных изданий на бумажных носителях и требования BELMARC к библиографическому описанию этих документов обусловили необходимость разработки специального пользовательского и программного интерфейса. В соответствии с требованиями стандарта, описание многотомных изданий должно состоять из общей части (библиографических сведений, общих для всех или большинства томов) и спецификаций (частных сведений, относящихся к отдельным томам). В то же время описание физически может быть выполнено как на одной каталожной карточке, так и на нескольких. В связи с этим пользовательский интерфейс должен позволять производить «слияние» данных нескольких каталожных карточек или «разбиение» данных одной карточки на логические блоки и последующую их обработку.

В ИТСР данная задача с успехом реализована, причем с возможностью привязки инвентарных номеров единиц хранения к соответствующим спецификациям, а также с возможностью подключения модуля автоматической маркировки.

6. Заключение

Описанная технология, предусматривающая режим верификации на самых ответственных операциях, позволяет выполнить довольно жесткие требования к качеству создаваемых библиографических записей. Заложенные в ИТСР технологические принципы обработки и контроля гарантируют возможность выгрузки в выходной файл только завершенных блоков данных, кратных каталожному ящику. Выгрузка информации из БД ИТСР осуществляется в файлы, структурированные в соответствии с ГОСТ 7. 14-98 (ISO 2709-96). Одновременно с выходным файлом предоставляется дополнительная информация, позволяющая контролировать полноту обработки. Достоинство хранения информации и результатов обработки в единой базе данных – возможность в любой момент оперативно получить сведения самого различного характера.