МЕЖФОРМАТНОЕ ПРЕОБРАЗОВАНИЕ ДАННЫХ В БИБЛИОГРАФИЧЕСКИХ СИСТЕМАХ НА ПРИМЕРЕ ФОРМАТОВ ISO-2709 И XML

Л. В. Кузнецова, Н. А. Мазов
Объединенный институт геологии, геофизики и минералогии СО РАН им. А. А. Трофимука, Новосибирск, Россия
e-mail: lioubov@ngs.ru, mazov@uiggm.nsc.ru

Nowadays numerous information bodies and libraries in Russia and abroad use various Database Management Systems that are based on ISO-2709-files. This standard underlies national exchange formats for bibliographic records such as USMARC, UNIMARC, RUSMARC, etc. However ISO-2709 format has a number of essential restrictions. Currently XML is the most universal tool for coding and representation of informational documents. In this article a software application intended for the inter-format transformation of data between ISO-2709 and XML is considered. The attention is focused on the means of formatting that allow not only to transform the data between these formats, but also to flexibly change presentation of data during processing.

Введение

В настоящее время многочисленные информационные органы и библиотеки как в России, так и за рубежом используют различные СУБД [1], основу которых составляют файлы в структуре стандарта ISO-2709 [2]. Широкое использование формата ISO-2709 в библиотечных системах обусловлено тем, что библиографическая информация является свободнотекстовой и слабоструктурированной, что не позволяет эффективно использовать для ее обработки реляционные СУБД. Стандарт ISO-2709 лежит в основе всех форматов для библиографических записей семейства MARC (MAchine Readable Cataloguing) [3], таких как USMARC [4], UNIMARC [5], RUSMARC [6] и др. Однако формат ISO-2709 имеет ряд существенных ограничений (например, на длину записи и уровень иерархии) и является сложночитаемым для пользователя (рис. 1).

В настоящее время наиболее универсальным средством кодирования и отображения содержания информационных документов является язык XML [7–9]. Иерархическая структура библиографической записи хорошо согласуется с моделью XML-документа (рис. 2). Использование XML в качестве формата обмена и хранения библиографических данных позволяет осуществлять контроль корректности записей на уровне проверки XML-документа. В отличие от формата ISO-2709, XML — это формат, читаемый для человека и легко документируемый.

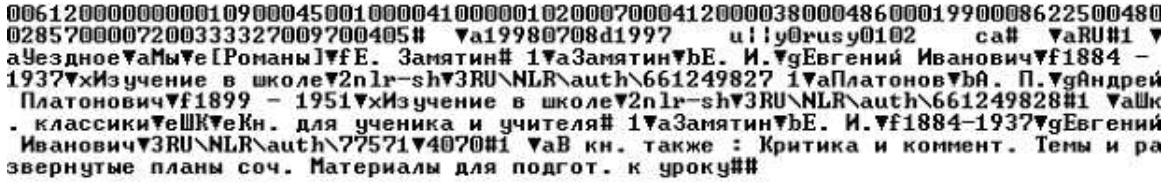


Рис. 1. Пример библиографической записи в формате ISO-2709.

Рис. 2. Пример библиографической записи в формате XML.

В отличие от большого разнообразия используемых MARC-форматов, XML стандартизирован и поддерживается большим количеством производителей программного обеспечения. В стандарт XML включена поддержка Unicode, что позволяет создавать многоязычные документы, а также использовать расширенный набор символов.

1. Общая характеристика разработанного приложения

Для межформатного преобразования (конвертирования) данных ISO-2709 и XML было разработано специализированное программное приложение. Отличительной чертой данного приложения от аналогичных конвертеров [10, 11], доступных в сети Интернет, является уникальная возможность преобразовывать внешнее представление данных в процессе конвертирования. Это необходимо, прежде всего, для полноценной работы с библиографической информацией, поскольку в повседневной практике часто требуются разнообразные способы отображения документа как целиком, так и его отдельных частей. Так, например, данные, содержащиеся в библиографической записи, могут быть использованы: для формирования карточки библиографического описания, требования для заказа книги в библиотеке, для изменения или добавления новой записи и т.д. То есть внешнее представление данных должно быть гибким и разнообразным. При использовании конвертеров, свободно распространяемых в сети Интернет, чтобы получить запись в нужном представлении в формате XML, необходимо воспользоваться как минимум двумя конвертерами, один из которых изменяет формат данных (между ISO-2709 и XML), а второй меняет внешнее представление данных. Основной отличительной чертой этого программного приложения является возможность осуществлять эти преобразования одновременно.

Приложение имеет классический многодокументный интерфейс, предоставляет возможность просматривать библиографические записи (рис. 3), а также просматривать и

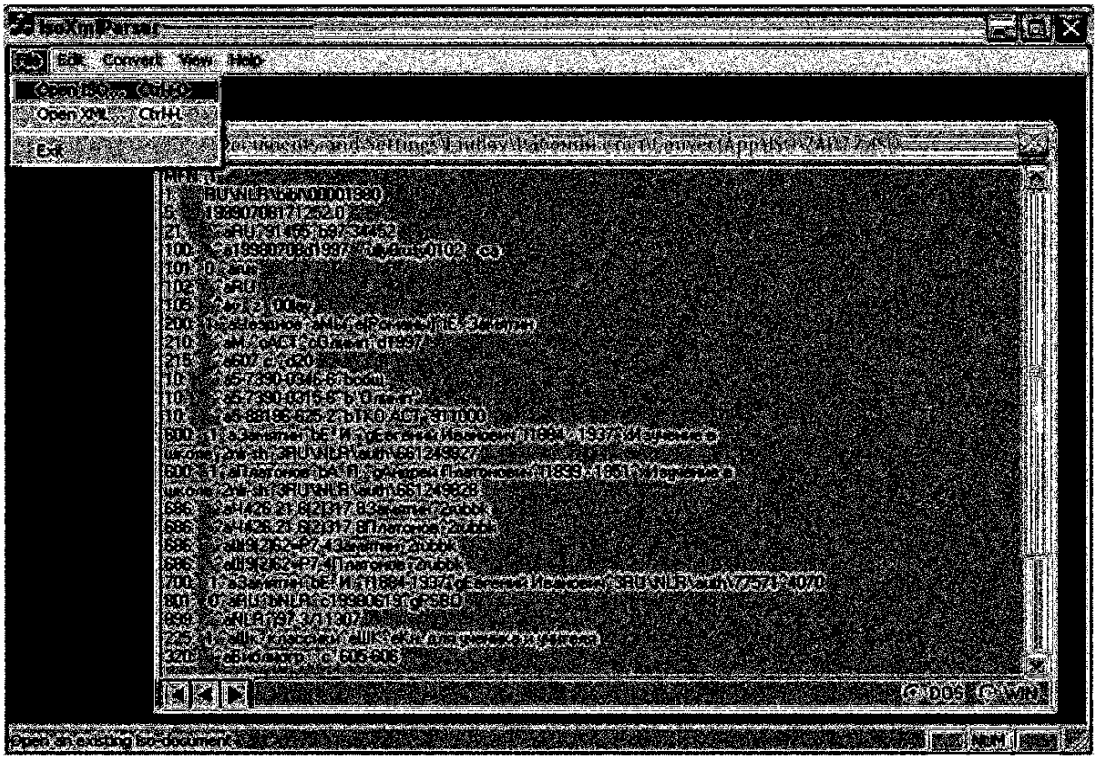


Рис. 3. Пример окна программного приложения с библиографической записью в формате ISO-2709.

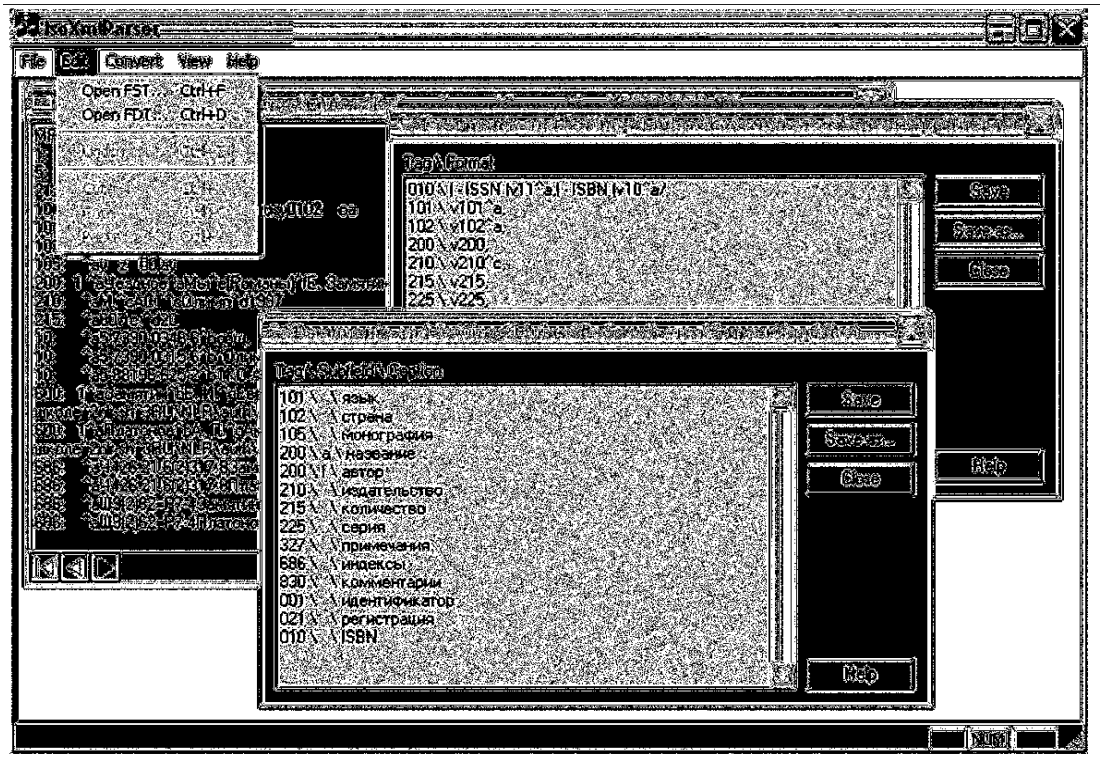


Рис. 4. Пример окон программного приложения с таблицами, применяемыми при преобразовании данных.

редактировать таблицы (рис. 4), используемые в процессе преобразования данных, структура которых будет описана ниже.

Кроме того, приложение предусматривает различные кодировки данных, в частности кириллицу (отсутствие данной возможности не позволяет эффективно эксплуатировать разработки зарубежных авторов, представленные в сети Интернет), и имеет достаточно подробную справочную систему.

2. Реализация программного приложения

Процесс конвертирования данных из формата ISO-2709 в формат XML в разработанном приложении можно представить блок-схемой, изображенной на рис. 5.

Преобразование данных осуществляется с использованием таблицы выбора полей и таблицы описания полей, определяемых пользователем на расширенном языке форматирования CDS/ISIS. Обе таблицы имеют простую структуру и могут заполняться как в любом текстовом редакторе, так и через интерфейс представленного программного приложения (см. рис. 4). Таблица описания полей (с расширением FDT — File Definition Table) предназначена для связи цифровых меток формата ISO-2709 и мнемонических названий XML-тегов. Структура FDT-файла представлена на рис. 6.

Другой пример таблицы описания полей (в интерфейсе приложения) представлен на рис. 4. Следует сказать, что для функционирования данного приложения не предусматривается никакой конкретной XML-схемы, а структура конечного XML-документа определяется только таблицей описания полей, задаваемой пользователем. Более того, пре-

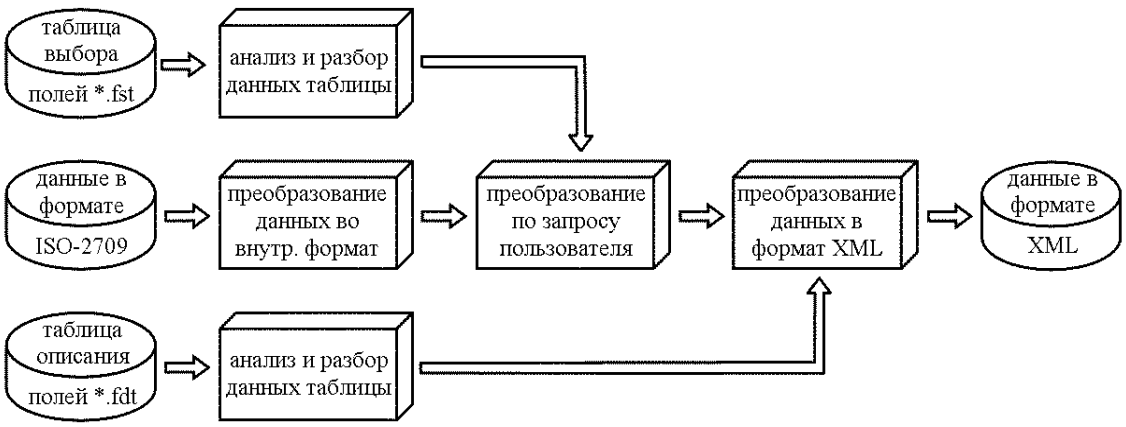


Рис. 5. Блок-схема конвертирования данных из формата ISO-2709 в формат XML.

Рис. 6. Структура таблицы описания полей.

Таблица 1
Пример FDT-файла

Поле	Подполе	Наименование
010	a	ISBN
101	a	язык
102	a	страна
200	a	название
200	f	автор

образование из формата ISO-2709 в формат XML может осуществляться при неполной таблице описания полей или вообще при ее отсутствии. В этом случае XML-тегами будут цифровые метки формата ISO-2709 (например, метка v200 на рис. 7). Таблица выбора полей (с расширением FST — File Selection Table) служит для преобразования внешнего представления данных и определяет содержимое элементов конечного XML-файла (или ISO-файла в зависимости от направления преобразования). Строки этой таблицы содер-

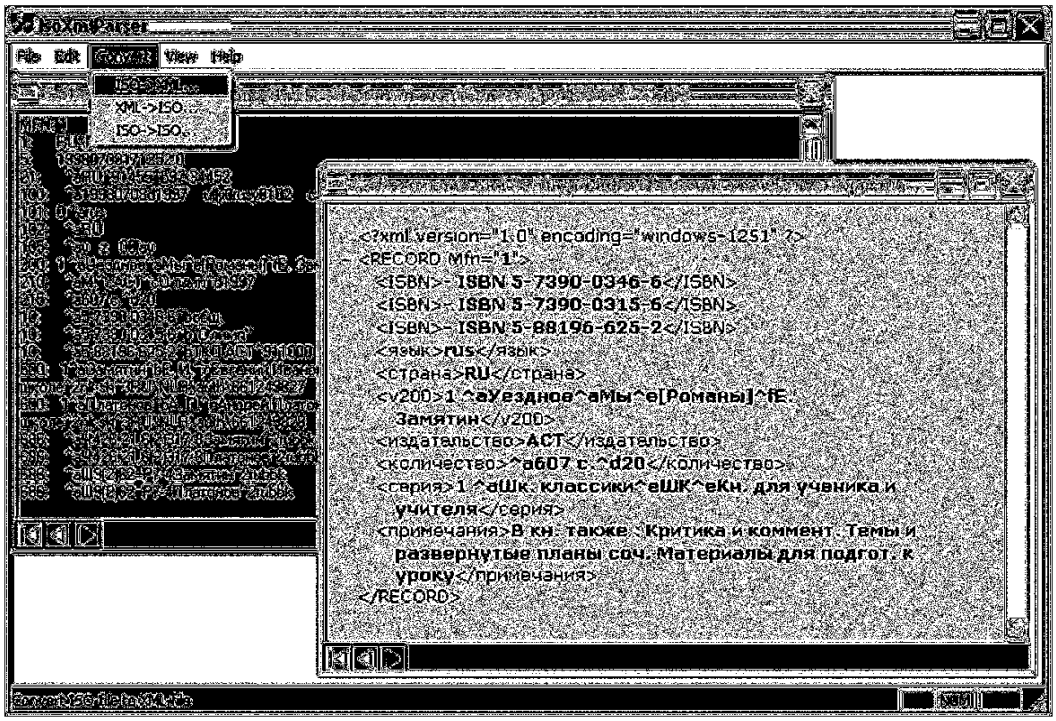


Рис. 7. Пример окон программного приложения с библиографическими записями: исходная — в формате ISO-2709, преобразованная — на языке XML.

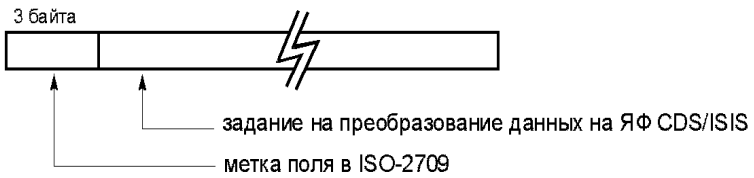


Рис. 8. Структура таблицы выбора полей.

Таблица 2

Пример FST-файла

Метка поля	Формат
010	| - ISBN |v10 ^ a/,
200	v200^ a|, |, v200^ f
801	if v801^ a='03' or v801^ a='08' then 'Депонир. рук.' fi,

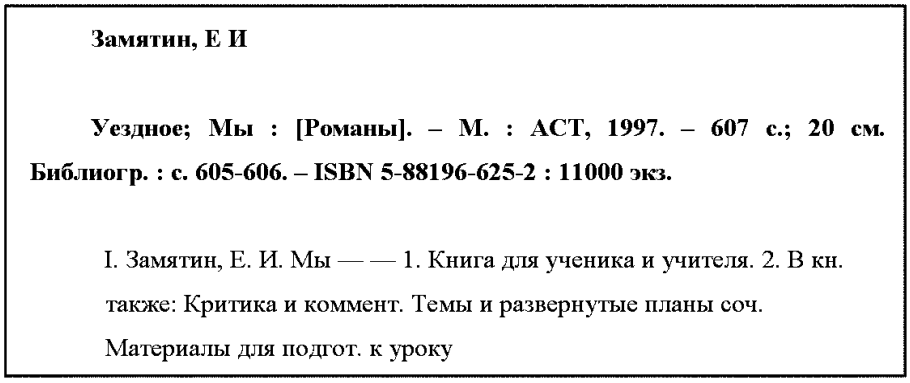
жат метку поля и задание на форматирование на расширенном языке форматирования CDS/ISIS. Структура FST-файла представлена на рис. 8.

Следует отметить, что применяемый язык форматирования представляет собой расширение языка CDS/ISIS, традиционно используемого для обработки библиографических данных в системе ISIS (UNESCO). Используемый расширенный язык форматирования является достаточно полным (например, включает в себя условный оператор, операторы циклов, множество строковых функций и др.) и имеет практически неограниченные возможности по представлению данных в различных формах. Несмотря на разнообразие конструкций, язык форматирования остается легко понимаемым и легко осваиваемым пользователями.

Стоит подробнее прокомментировать схему, представленную на рис. 5. На первом этапе работы приложения данные ISO-2709 преобразуются во внутренний формат (это структура, содержащая такие данные, как метки полей, длины полей и сами поля данных), кроме того, происходит анализ таблиц выбора полей и определения полей (если таковые имеются). Далее происходит изменение внешнего представления данных согласно таблице выбора полей. Именно возможность этого изменения и отличает данное приложение от программных продуктов данного класса. По завершении форматирования данных происходит создание XML-элементов (см. рис. 7) с учетом названий тегов, представленных в таблице описания полей (при отсутствии таблицы описания полей для создания тегов XML используются метки ISO-2709). Следует заметить, что если поле в записи ISO-2709 содержит подполя, то они оформляются более глубоким уровнем вложения (иерархии) в конечном XML-документе. Повторяющиеся поля, наличие которых допускается форматом ISO-2709, получают после преобразования одинаковые теги и одинаковый уровень иерархии.

Настоящее приложение позволяет также осуществлять преобразование данных в направлении XML-ISO-2709. В этом направлении преобразование происходит практически аналогично преобразованию из формата ISO-2709 в формат XML (вместо создания XML-элементов происходит генерация формата ISO-2709) и не требует дополнительных комментариев. Кроме этого, программное приложение позволяет осуществлять преобразование ISO-2709 — ISO-2709, т. е. в данном случае изменяется только внешнее представление данных, внутренний же формат данных сохраняется.

Ниже кратко продемонстрирована работа представленного программного приложения. В качестве примера рассмотрена библиографическая карточка. В этой карточке представлена не вся информация, имеющаяся о данном издании в библиографической записи (это один из примеров необходимости иметь различное внешнее представление данных). На рис. 9 показаны полная библиографическая запись, а также таблицы выбора и определения полей, с помощью которых осуществляется преобразование библиографической записи из формата ISO-2709 в формат XML. Как видно из рис. 9, конечные данные на языке XML содержат не полную информацию о книге (которая, возможно, избыточна в некото-



Замятин, Е И

Уездное; Мы : [Романы]. – М. : АСТ, 1997. – 607 с.; 20 см. Библиогр. : с. 605-606. – ISBN 5-88196-625-2 : 11000 экз.

I. Замятин, Е. И. Мы — — 1. Книга для ученика и учителя. 2. В кн. также: Критика и коммент. Темы и развернутые планы соч. Материалы для подгот. к уроку

Рис. 9. Пример библиографической карточки.

рых случаях), а только поля, затребованные пользователем с помощью таблицы выбора полей.

Заключение

Заметим, что основная цель данной работы заключалась не в детальном рассмотрении MARC-форматов и XML-технологий, а в том, чтобы подчеркнуть актуальность проблемы преобразования данных между форматами ISO-2709 и XML.

Поскольку формат MARC — это в первую очередь формат внешнего представления данных, а его цель — служить средством обмена данными (например, в среде сети Интернет), в настоящее время разработан опытный образец аналогичного web-ориентированного приложения, позволяющий гибко осуществлять импорт различных библиографических данных в локальные информационно-библиотечные системы. Это клиент-серверное приложение реализует все функции описанного локального приложения и позволяет обрабатывать и отображать в формате HTML записи ISO-2709, содержащие TeX-поля (это особенно актуально для работы с математическими текстами в библиографических базах данных). Программное приложение представляет собой модуль расширения PHP. Опытно-промышленный вариант эксплуатируется на web-сервере Apache 2.0.50 с PHP 4.3.7.

Список литературы

[1] Дейт К.Дж. Введение в системы баз данных. 7-е изд.: Пер. с англ. М.: Издательский дом “Вильямс”, 2001. 1072 с.

[2] International Organization for Standardization. Documentation: format for bibliographic information interchange on magnetic tape. [2nd ed.] Geneva, ISO, 1981 (ISO 2709-1981).

[3] Основные положения формата MARC для библиографических данных / Под общей ред. действительного члена постоянного Комитета по UNIMARC Я.Л. Шрайберга. М.: ГПНТБ России, 1997. 39 с.

[4] Форматы USMARC. Краткое описание: В 3 ч. М.: ГПНТБ России, 1996.

[5] Руководство по UNIMARC: Руководство по применению международного коммуникативного формата UNIMARC. М.: ГПНТБ России, 1992. 320 с.

[6] Российский коммуникативный формат представления библиографических записей в машиночитаемой форме (Рос. вариант UNIMARC). СПб.: Изд-во РНБ, 1998.

[7] Спецификации W3C: XML: общая информация. http://www.w3.org/XML;
XML 1.0 (Second Edition). http://www.w3.org/TR/2000/REC-xml-20001006

[8] Рэй Э. Изучаем XML: Пер. с англ. СПб.: Символ-Плюс, 2001. 408 с.

[9] Питц-Моултис Н., Кирк Ч. XML в подлиннике: Пер. с англ. СПб.: БХВ-Петербург, 2000. 736 с.

[10] Конвертер MARC-записей в XML-документы. http://xmlmarc.stanford.edu/

[11] Конвертер UNESCO. http://www.unesco.org/webworld/isis/xml2isis.htm/

Поступила в редакцию 18 марта 2005 г.