Преобразование данных в библиографических системах: на примере форматов ISO-2709 и XML

Data Transformation in Bibliographic Systems: Based on the Example of ISO-2709 and XML Formats

Перетворення даних у бібліографічних системах: на прикладі форматів ISO-2709 і XML

Кузнецова Л. В., Мазов Н. А.
Объединенный Институт геологии, геофизики и минералогии им. академика А. А. Трофимука Сибирского Отделения РАН, Новосибирск, Россия

L. V. Kuznetsova and N. A. Mazov
The Academician Trofimuk United Institute of Geology, Geophysics and Mineralogy of the Siberian Branch of Russian Academy of Science, Novosibirsk, Russia

Кузнєцова Л. В., Мазов Н. А.
Об'єднаний Інститут геології, геофізики і мінералогії ім. академіка А. А. Трофімука Сибірського Відділення РАН, Новосибірськ, Росія

В настоящее время многочисленные информационные органы и библиотеки как в России, так и за рубежом используют различные СУБД, основу которых составляют файлы в структуре стандарта ISO-2709 (или ему подобных). Этот стандарт лежит в основе таких национальных обменных форматов для библиографических записей, как USMARC, UNIMARC, RUSMARC и др. Это обусловлено тем, что библиографическая информация является свободнотекстовой и слабо структурированной, что не позволяет эффективно использовать для ее обработки реляционные СУБД.
Бурное развитие в последнее время XML-технологий, а также рост программных продуктов, оперирующих с данными в формате XML, предоставляет стандартную возможность кодирования содержания информационных документов, обеспечивая при этом гибкость в создании структур данных. При этом иерархическая структура библиографической записи хорошо согласуется с моделью XML-документа. Использование XML в качестве формата обмена и хранения библиографических данных позволяет осуществлять контроль корректности записей на уровне проверки XML-документа.
В докладе рассматривается разработанное программное приложение, предназначенное для межформатного преобразования данных ISO-2709 и XML. Акцентируется внимание на средствах форматирования, позволяющих не только преобразовывать данные между этими форматами, но и гибко изменять представление данных в процессе обработки.

At present numerous information centers and libraries either in Russia or abroad use various DBMS based on the files structured in accordance with ISO-2709 or similar standards. This standard underlies such national exchange formats for bibliographic records, as USMARC, UNIMARC, RUSMARC, etc. The reason for that is that the bibliographic information is created as a free text and poorly structured, that does not allow relational DBMS to be used use effectively for its processing.
Rapid development of XML-technologies, and also growth of the software products operating on the data in XML format, gives a standard opportunity to encode the contents of documents, providing thus flexibility in creation of structures of the data. Thus the hierarchical structure of bibliographic record will well be coordinated with the model of the XML-document. Use XML as a format of an exchange and storage of the bibliographic data allows to carry out the control of a correctness of records over a level of check of the XML-document.
The developed program appendix intended for interformatted transformation of data ISO-2709 and XML is considered in the report. The focus is made on the of formatting means allowing not just data transformation between these formats, but also slight changes in data presentation during processing.

На сьогодні численні інформаційні органи і бібліотеки як у Росії, так і за її межами використовують різні СУБД, основу яких складають файли у структурі стандарту ISO-2709 (або подібними до нього). Цей стандарт взято за основу таких національних обмінних форматах для бібліографічних записів, як USMARC, UNIMARC, RUSMARC та ін. Це зумовлено тим, що бібліографічна інформація є вільно текстовою і слабко структурованою, що не дозволяє ефективно використовувати для її обробки реляційні СУБД.
Бурхливий розвиток протягом останнього часу XML-технологій, а також збільшення кількості програмних продуктів, що оперують з даними у форматі XML, надає стандартну можливість кодування змісту інформаційних документів, забезпечуючи при цьому гнучкість у створенні структур даних. При цьому ієрархічна структура бібліографічного запису добре узгоджується з моделлю XML-документа. Використання XML у якості формату обміну і зберігання бібліографічних даних дозволяє здійснювати контроль коректності записів на рівні перевірки XML-документа.
У доповіді розглянуто розроблений програмний додаток, призначений для міжформатного перетворення даних ISO-2709 і XML. Акцентується увага на засобах форматування, що дозволяють не тільки перетворювати дані між цими форматами, а й гнучко змінювати представлення даних у процесі обробки.

Большинство библиотек России и мира используют различные базы данных, основу которых составляют файлы в структуре ISO-2709 [1]. Этот формат лежит в основе таких обменных форматов для библиографических записей, как USMARC [2], UNIMARC [3], RUSMARC [4] и др. Широкое использование формата ISO-2709 в библиотечных системах обусловлено тем, что библиографическая информация является свободнотекстовой и слабо структурированной, что не позволяет эффективно использовать для ее обработки реляционные СУБД [5]. Несмотря на широкое распространение, следует отметить, что формат ISO-2709 имеет ряд существенных недостатков. Например, его записи имеют ограничение на длину, уровень иерархии и сложночитаемым для пользователя.

С другой стороны, можно сказать, что наиболее перспективным и универсальным средством для представления структурированных данных в настоящее время является язык XML [6-7]. Иерархическая структура библиографической записи хорошо согласуется с моделью XML-документа. 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 читаемый формат для человека и легко документируемый. В отличие от большого разнообразия используемых MARC-форматов, XML стандартизирован и поддерживается большим количеством производителей программного обеспечения. В стандарт XML включена поддержка страниц Unicode, что упрощает создание многоязычных документов.

Таким образом, очевидна актуальность проблемы взаимного преобразования данных в форматах ISO-2709 и XML. В настоящее время предприняты попытки построения MARC — XML — конвертеров, известно несколько разработок европейских и американских университетов и библиотек [8-9]. Наиболее удачными, на наш взгляд, являются конвертеры Стэндфордского университета и ЮНЕСКО. Но большая часть из представленных в сети Интернет программных продуктов не позволяет преобразовывать данные в процессе конвертирования. Однако для полноценной работы с документом необходимо иметь разнообразные способы отображения как документа целиком, так и его частей. Данные, содержащиеся в библиографической записи, могут быть использованы, например, для формирования карточки библиографического описания, требования для заказа книги в библиотеке, а также для изменения или добавления новой записи. Таким образом, чтобы получить запись в нужном представлении в формате XML, во-первых, необходимо воспользоваться парсером для формата ISO-2709, и только после этого возможно применять предлагаемые конвертеры. Кроме этого, можно отметить такие недостатки представленных программных продуктов, как работа только с конкретным вариантом MARC-формата и отсутствие поддержки кириллицы.

Исходя из вышесказанного, было разработано программное приложение, позволяющее одновременно с межформатным преобразованием осуществлять гибкое изменение внутреннего представления данных посредством языка форматирования, основу которого составляет язык форматирования данных системы CDS/ISIS.

Процесс конвертирования данных в созданном приложении можно представить следующей блок-схемой, представленной на рис.1.

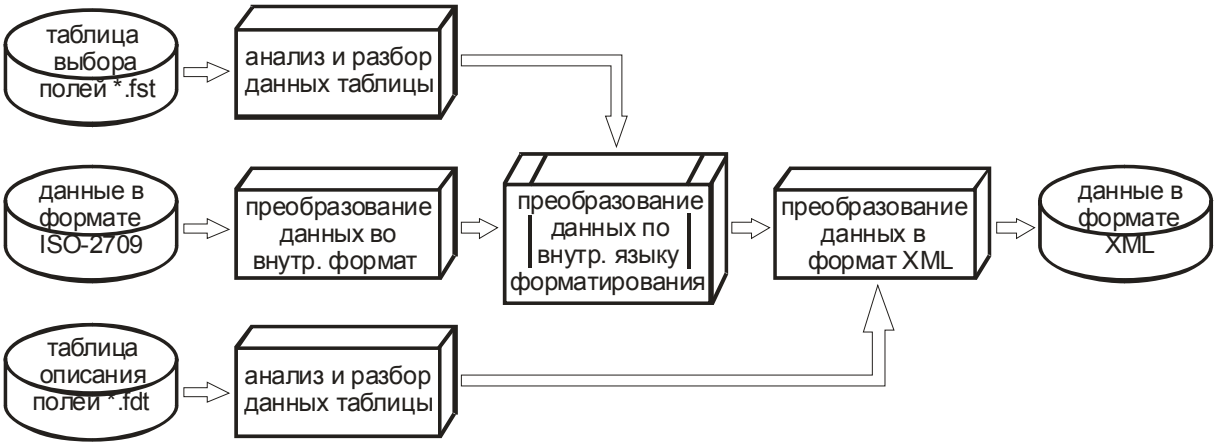

Рис.1

Для осуществления межформатного преобразования в приложении используется:

1. Таблица описания полей, предназначенная для связи цифровых меток формата ISO-2709 и мнемонических названий XML-тегов.

2. Таблица выбора полей для определения содержимого конечного файла. Строки этой таблицы содержат метку поля и задание на форматирование на расширенном языке форматирования.

После преобразования записи во внутренний формат происходит изменение представления данных согласно таблице выбора полей. (Именно возможность этого изменения и отличает данное приложение от программных продуктов данного класса, имеющихся в свободном доступе). После завершения внутреннего форматирования происходит создание XML-элементов с учетом названий тегов, представленных в таблице определения полей. Настоящее приложение позволяет получать конечные данные не только в формате XML, но и в формате ISO-2709.

Следует отметить, 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.

Учитывая тот факт, что формат MARC — это в первую очередь формат внешнего представления данных, и его цель — служить средством обмена данными (например, в среде сети Интернет), в настоящее время ведутся работы по созданию аналогичного Web-ориентированного приложения. Такое приложение позволит гибко осуществлять импорт различных библиографических данных в локальные информационно-библиотечные системы.

Литература

1. International Organization for Standardization. Documentation: format for bibliographic information interchange on magnetic tape. [2 ed.] Geneva, ISO, 1981 (ISO 2709-1981). The first edition was published in 1973.
2. Форматы USMARC. Краткое описание: В 3-х ч. М.: ГПНТБ России. — 1996.
3. Руководство по UNIMARC: Руководство по применению международного коммуникативного формата UNIMARC. — М.: ГПНТБ России. — 1992. — 320 с.
4. Российский коммуникативный формат представления библиографических записей в машиночитаемой форме: (Российский вариант UNIMARC). СПб.: Изд-во РНБ. — 1998.
5. Основные положения формата MARC для библиографических данных. / Под общей редакцией действительного члена постоянного Комитета по UNIMARC Я. Л. Шрайберга. ГПНТБ России. — М., 1997. — 39 с.
6. Питц-Моултис Н., Кирк Ч. XML: Пер. с англ. — СПб.: БХВ-Петербург. — 2000. — 736 с.
7. XML 1.0 (Second Edition) — http://www.w3.org/TR/2000/REC-xml-20001006
8. Конвертер MARC-записей в XML-документы http://xmlmarc. stanford. edu/
9. Конвертер UNESCO http://www.unesco.org/webworld/isis/xml2isis.htm/