Технологии XML – как основа интеграции данных в информационно-библиотечных системах

XML-technologies – a Basis for Data Integration within Library Information Systems

Технології XML – як основа інтеграції даних в інформаційно-бібліотечних системах

Н. А. Мазов
Институт нефтегазовой геологии и геофизики им. Академика А. А. Трофимука Сибирского Отделения РАН, Новосибирск, Россия

О. Л. Жижимов
Институт вычислительных технологий Сибирского Отделения РАН, Новосибирск, Россия

Nikolay Mazov
Ac. A. A. Trofimuk Institute of Oil and Gas Geology and Geophysics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

Oleg Zhizhimov
Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia

М. О. Мазов
Інститут нафтогазової геології та геофізики ім. Академіка А. А. Трофимука Сибірського відділення РАН, Новосибірськ, Росія

О. Л. Жижимов
Інститут обчислювальних технологій Сибірського відділення РАН, Новосибірськ, Росія

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, переносимость на различные аппаратно-программные платформы. 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 роль XML и перспективы его использования в библиотечно-информационных системах, а также применительно к основным функциям информационных систем – хранению, поиску, представлению и передаче данных.

The recent rapid development of XML-technologies, and advance of software products for XML-format data enables standard coding of document contents while providing flexibility of data organization, and compatibility with hardware & software platforms. The hierarchical structure of bibliographic records conforms with XML-document model. Using XML – format for bibliographic data exchange and storage enables record correctness control records at the stage of XML-document checking. The role of XML and prospects for of its use in library and information systems, and also with reference to the basic functions of information systems – to storage, search, browse and data transmission is considered.

Бурхливий розвиток останнім часом Xml-Технологій, а також збільшення кількості програмних продуктів, що оперують із даними у форматі XML, надає стандартну можливість кодування змісту інформаційних документів, забезпечуючи при цьому гнучкість у створенні структур даних, можливість перенесення на різні апаратно-програмні платформи. Ієрархічна структура бібліографічного запису добре узгодиться з моделлю XML- документа. Використання XML у якості формату обміну та зберігання бібліографічних даних дозволяє здійснювати контроль коректності записів на рівні перевірки Xml-Документа. У доповіді розглянуто роль XML та перспективи його використання в бібліотечно-інформаційних системах, а також відносно основних функцій інформаційних систем – зберігання, пошуку, подання та передачі даних.

Информационно-библиотечные системы, используемые в настоящее время в российском библиотечном сообществе основаны на различных СУБД, основу которых, в той или иной степени, составляют файлы в структуре стандарта ISO-2709 или подобных ему [1]. Кроме этого, структура ISO-2709 лежит в основе таких обменных форматов для библиографических данных, как MARC21 [2], UNIMARC [3], RUSMARC [4] и др. Широкое использование стандарта ISO-2709 в библиотечной практике обусловлено исторически. В связи с бурным развитием информационных и web технологий в последние годы, делает малооправданным использование форматов на основе ISO-2709, несмотря на его широкое распространение. Ряд его существенных недостатков, например, ограничение на длину, уровень иерархии и сложночитаемость становятся все более заметнее для пользователя.

С другой стороны, можно сказать, что наиболее перспективным и универсальным средством для представления структурированных данных в настоящее время является язык XML [5-9]. С момента своего появления язык XML зарекомендовал себя с самой лучшей стороны, поэтому довольно быстро получил широкое распространение. Он оказался чрезвычайно полезной технологией, а по сравнению с ISO-2709 язык XML имеет ряд очевидных преимуществ. 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 легкочитаемый формат для пользователя и легко документируемый, а также в нем отсутствуют ограничения на длину документа. В отличие от ISO-2709 и разнообразием MARC-форматов, порожденных на его основе, XML более динамичен и позволяет легко порождать новые схемы данных и правила перехода между ними, которые формулируются на том же самом языке (XSLT – преобразования), а также поддерживается большим количеством производителей программного обеспечения. В стандарт XML включена поддержка кодировок Unicode, что упрощает создание многоязычных документов. Кроме этого следует заметить, что ISO-2709 в отличие от XML не предусматривает передачу двоичных данных, таких как графическое изображение или аудио- или видеоматериалы, без которых сегодня немыслима информационная среда. Особо важным достоинством языка XML является его интеграция с web-средой, а также платформенная независимость. Все это де-факто делает XML стандартом для обмена данными.

Таким образом, очевидна актуальность проблемы взаимного преобразования данных в форматах ISO-2709 и XML. В настоящее время предприняты попытки построения MARC – XML – конвертеров (парсеров), известно несколько разработок европейских и американских университетов и библиотек [10-11]. 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.

Следует заметить, что существующая практика представления библиографических записей в формате XML, основано на эмуляции структур ISO-2709 средствами XML, является половинчатым решением, в рамках которого невозможно использовать все достоинства XML-технологии. Иными словами, проблема не решается, а, образно говоря, прячется за красивые слова об XML. В качестве иллюстрации ограничения, используемых подходов можно сказать, что структура описания документа наследуется полностью и остается двухуровневой. В качестве различных интерпретаций XML-представлений ISO-2709 записей можно привести следующие эмуляции записей: SimpleXML, OAI, MarcXML, XChange и др. (см. http://z3950.nsc.ru:210).

На наш взгляд, более перспективным применением XML-технологий для представления библиографических записей является применение схем данных, не связанных с ISO-2709. В качестве

примера такой схемы может выступать схема MODS (http://www.loc. gov/mods), однако ее применение потребует изменения правил каталогизации и дополнительной стандартизации библиографических описаний, отличных от идеологии MARC-форматов и в будущем именно эти технологии будут занимать лидирующее место.

Наряду со средствами преобразования, следует отметить полезное ПО для создания и редактирования XML документов в принципе достаточно самого простого текстового редактора, однако разработаны и существуют более «продвинутые» XML – редакторы, например, XED, XML Notepad, VisualXML и др. [12-14]. Для анализа и разбора (проверки корректности и/или состоятельности) XML документов доступны следующие анализаторы: XP, AElfred и др. [15-16].

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.

В настоящее время авторами ведутся работы по разработке XML-сервера библиографических БД, позволяющего осуществлять полнофункциональную обработку библиографических данных, представленных в формате XML. Следует отметить, что основная цель данной работы заключается не в детальном рассмотрении MARC-форматов и XML-технологий, а в том, чтобы подчеркнуть актуальность проблемы преобразования данных между форматами ISO-2709 и XML.

В заключение отметим, что переход на XML стандарт описания библиографической информации позволит независимым пользователям, использующим различные аппаратно-программные платформы легко использовать информацию, получаемую друг у друга, поскольку документы XML можно использовать как минимум в двух ипостасях: во-первых, как информацию для восприятия человеком и, во-вторых, как данные для обработки приложениями.

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. Валиков А. Н. Технология XSLT. СПб.: БХВ-Петербург, 2002
8. Грейвс М. Проектирование баз данных на основе XML. М: Издательский дом «Вильямс»,2002
9. XML 1. 0 (Second Edition) – http://www.w3.org/TR/2000/REC-xml-20001006
10. Конвертер MARC-записей в XML-документы http://xmlmarc.stanford.edu/
11. Конвертер UNESCO http://www.unesco.org/webworld/isis/xml2isis.htm/
12. XED www.ltd.ed.ac.uk/`ht/xed.html
13. XML Notepad www.microsoft.com/xml
14. VisualXML www.pierlou.com
15. XP www.jclark.com
16. AElfred www.microstar.com