Некоторые проблемы доступа к библиографической информации по протоколу Z39.50

Some Problems of Access to Bibliographic Information via Z39.50 Protocol

Деякі проблеми доступу до бібліографічної інформації за протоколом Z39.50

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

O. L. Zhizhimov 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

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

В последние годы в библиотечном сообществе России активно развивается организации распределенного доступа к библиографическим информационным ресурсам с применением протокола Z39.50, проделана большая работа по стандартизации процессов поиска и представления информации в соответствии с мировыми правилами и международными стандартами, создана национальная инфраструктура библиографических ресурсов доступных по протоколу Z39.50, появились серверные программные продукты, не уступающие мировым аналогам, а иногда и превосходящие их. Однако, существует ряд проблем, которые препятствуют интеграции библиографических ресурсов с другими информационными ресурсами и которые усложняют и без того не простые процедуры доступа к библиографическим ресурсам по протоколу Z39.50. В докладе рассмотрены противоречия содержания стандарта в части библиографической информации модели данных Z39.50.

In recent years the library community in Russia actively develops distributed access to bibliographic information resources with application of Z39.50 protocol. A tremendous job on search and information representation processes standardization is done in accordance with global rules and international standards. The national infrastructure of bibliographic resources accessible via Z39.50 protocol is created, the server software products which are not conceding to world analogues, and sometimes and surpassing them have appeared. However, there are some problems, which interfere with bibliographic resources integration with other information resources and complicate not easy procedures of accessing bibliographic resources via Z39.50 protocol even more. Contradictions in the standard in part of bibliographic information within the model of Z39.50 data are considered.

Протягом останніх років у бібліотечній спільноті Росії активно розвиваються організації розподіленого доступу до бібліографічних інформаційних ресурсів із застосуванням протоколу Z39.50, виконано велику роботу із стандартизації процесів пошуку і представлення інформації згідно зі світовими правилами і міжнародними стандартами, створено національну інфраструктуру бібліографічних ресурсів, доступних за протоколом Z39.50, з'явились серверні програмні продукти, що не поступаються світовим аналогам, а іноді й кращі за них. Однак, існує ряд проблем, які заважають інтеграції бібліографічних ресурсів з іншими інформаційними ресурсами і які ускладнюють і без того непрості процедури доступу до бібліографічних ресурсів за протоколом Z39.50. У доповіді розглянуто протиріччя змісту стандарту у частині бібліографічної інформації моделі даних Z39.50.

В последние годы в библиотечном сообществе России активно эксплуатируется парадигма организации распределенного доступа к библиографическим информационным ресурсам с применением протокола Z39.50 [1]. Сам по себе этот факт благоприятно повлиял и продолжает влиять на общую концепцию построения библиотечных информационных систем и упорядочивание собственно библиографических ресурсов. За сравнительно небольшой срок была проделана колоссальная работа по стандартизации процессов поиска и представления информации в соответствии с мировыми правилами и международными стандартами, создана национальная инфраструктура библиографических ресурсов доступных по протоколу Z39.50, появились серверные программные продукты, не уступающие мировым аналогам, а иногда и превосходящие их [2].

Тем не менее, на наш взгляд, не все так прекрасно в общей концепции организации доступа к библиографической информации по протоколу Z39.50, как зачастую преподносится специалистами библиотечному сообществу. Существует ряд проблем, которые препятствуют интеграции библиографических ресурсов с другими информационными ресурсами, проблем, которые усложняют и без того не простые процедуры доступа к библиографическим ресурсам по протоколу Z39.50. Некоторые из этих проблем рассмотрены ниже.

Прежде всего, возьмем на себя смелость сделать следующее утверждение:

Несмотря на то, что протокол Z39.50 разрабатывался для организации сетевого доступа к библиографическим информационным ресурсам, на сегодняшний день практика организации этого доступа и содержание стандарта в части библиографической информации противоречит модели данных Z39.50.

Модель данных Z39.50 основана на абстрактных иерархических схемах (schema), которые формулируются в терминах меток (тегов) из стандартизованных наборов меток (tagSet). Схема в свою очередь представляет собой набор семантических правил, по которым формируется абстрактная структура записи (abstract record structure), являющаяся в Z39.50 основой обработки данных при извлечении информации. Идеология максимального абстрагирования от структур реальных баз данных приводит к весьма изощренной схеме извлечения данных, описанной в стандарте Z39.50: записи из результирующих наборов отображаются в записи абстрактной базы данных через схему, определяющую абстрактную структуру записи в виде дерева элементов, специфицируемых метками (tag) из стандартных наборов (tagSet); затребованные элементы выбираются в нужной альтернативной форме (variant) из абстрактной записи и отображаются в экспортируемую структуру, определяемую форматом внешнего представления (recordSyntax). Все объекты описанной процедуры (schema, tagSet, elementSpec, variantSet, recordSyntax) определены в соответствующих классах с присвоением OID.

Более грубая схема извлечения данных выглядит так:

1. Запись извлекается из базы данных в своей физической структуре.
2. Запись конвертируется в иерархическую теговую абстрактную структуру (abstract record structure) в соответствии с правилами затребованной схемы данных (schema).
3. Абстрактная структура записи модифицируется в соответствии с затребованными вариантами (variant) и элементами (elements).
4. Результирующая абстрактная структура записи преобразуется в физическую структуру внешнего представления в соответствии с затребованным форматом (recordSyntax).
5. Результирующая структура внешнего представления отсылается клиенту.

При этом следует напомнить, что в запросе на представление данных со стороны клиента могут присутствовать:

- указание на схему данных (schema);
- указание на вариант представления (variant);
- требуемые элементы (elementSpec);
- формат представления (recordSyntax).

Очень важным моментом в этой модели является тот факт, что семантическое наполнение записи происходит только на шаге 2. Только на этом шаге возможна конвертация записи из схемы хранения в базе данных в затребованную схему. При этом преобразование должно происходить только для абстрактной структуры записи. Модель не подразумевает конвертации конечных форматов записей.

Если применить описанную модель извлечения данных к библиографической информации, мы сразу получим противоречие: *в глобальном реестре Z39.50 не существует ни одного объекта в классе определения схем данных (OID 1.2.840.10003.13 — database schema definitions) для библиографической информации*. Иными словами, отсутствует стандартная схема структуры и семантики библиографической записи.

Возникает вопрос, а как тогда мы вообще умудряемся извлекать библиографические записи в Z39.50. Очень просто, мы никогда не требуем схемы (ее, как мы выяснили, нет), а требуем только формат (recordSyntax), например, RUSmarc или USmarc, ожидая семантическое наполнение извлекаемых записей в соответствии с затребованным форматом. Действительно так можно поступать, т. к. в классе *record syntax definitions (OID 1.2.840.10003.5)* определены RUSmarc, USmarc и множество других диалектов MARC, которые определяют не столько структуру записи, которая у них у всех одна – ISO-2709, сколько семантику записи, которая должна определяться в другом классе. Таким образом, налицо второе противоречие: *используемые в Z39.50 форматы представления библиографических записей семейства MARC, вообще говоря, форматами не являются, т. к. все имеют один и тот же формат – ISO-2709*. Но самое интересное состоит в том, что объект, соответствующий настоящему формату ISO-2709, по непонятной причине отсутствует в глобальном реестре Z39.50, что вносит дополнительную путаницу. Сразу возникает третье противоречие: *все национальные MARC-форматы, являющиеся по своей сути схемами данных, не определены в классе «database schema definition»*.

Наличие этих противоречий в Z39.50 носит исторический характер. Все уже привыкли называть, например, RUSmarc форматом данных, а не схемой данных, в то время как он является именно схемой данных, а форматом является ISO-2709. Тем не менее, отсутствие в Z39.50 для MARC четкого разграничения понятий формата (recordSyntax) и схемы данных приводит к некоторым трудностям при извлечении данных и при автоматической конвертации данных между различными схемами.

Приведем несколько примеров.

Пример 1.

Клиент желает получить библиографическую запись в семантике RUSmarc в формате XML. К сожалению, при существующих спецификациях это может быть выполнено только в том случае, если записи хранятся в базе данных в соответствии с семантикой RUSmarc, т. к. сформулировать в Z39.50 запрос на извлечение данных в формате XML в схеме RUSmarc невозможно по причине отсутствия схемы RUSmarc.

Пример 2.

Клиент желает получить одну и ту же библиографическую запись в формате XML сначала в семантике RUSmarc, а затем в семантике USmarc. Это, казалось бы, простое естественное желание не может быть сформулировано как запрос на представление данных в Z39.50 по причине отсутствия схем RUSmarc и USmarc.

Приведенные примеры демонстрируют отсутствие в Z39.50 необходимых объектов для работы с библиографическими данными, но ни в коем случае не компрометируют саму идеологию Z39.50. Для других типов информационных ресурсов, для которых созданы и зарегистрированы схемы данных, такой проблемы не возникает.

Перечисленные выше противоречия легко устраняются при стандартизации дополнительных объектов Z39.50 в классах «*database schema definition» и «record syntax definitions».* В частности, необходимо:

- Зарегистрировать в классе «*database schema definition» (OID 1.2.840.10003.13)* объекты, соответствующие схемам RUSmarc, USmarc и т. д.
- Зарегистрировать в классе «*record syntax definitions» (OID 1.2.840.10003.5)* объект, соответствующий формату ISO-2709.
- Определить прикладным профилем способ извлечения библиографических записей с обязательным указанием требуемой схемы и формата представления данных, т. е. пар типа

Schema	**Record Syntax**
RUSmarc	XML
RUSmarc	ISO-2709
RUSmarc	SUTRS
RUSmarc	GRS-1
USmarc	XML
USmarc	ISO-2709

Перечисленные мероприятия снимают противоречия, присущие способу работы с библиографическими данными в Z39.50, т. к. полностью переводят представление этих данных в соответствие с моделью данных Z39.50. Однако остается проблема формулирования собственно схемы данных для библиографических записей. Дело в том, что, назвав тот же RUSmarc или USmarc схемой данных, мы лишь формально удовлетворили требованиям процедуры формирования записи. Но нужно еще определить схемы данных (RUSmarc, USmarc), причем определить их в терминах Z39.50[1].

Здесь следует заметить, что ведущие разработчики программного обеспечения Z39.50, продукты которых ориентированы не только на библиотечные системы, решают по-своему перечисленные выше проблемы. В частности, компания Index Data (YAZ ToolKit, Zebra и др.) в свои продукты встраивает обработку форматов MARC как схем данных.

Однако даже если RUSmarc зарегистрировать как схему данных, останутся проблемы. Например, проблемы представления библиографической научно-технической информации (НТИ). Учитывая, что библиографических научно-технических ресурсов существует достаточно много (электронные варианты реферативных журналов ВИНИТИ, описания патентов и изобретений и др.) и они представляют значительный интерес для большой категории пользователей, при организации доступа к этим ресурсам возникает необходимость выбора типа метаданных для их описания.

Здесь следует сделать оговорку. Для поиска научно-технической информации, в том числе и библиографической, в глобальном реестре Z39.50 существует набор поисковых атрибутов STAS (**S**cientific & **T**echnical **A**ttribute **S**et) *(OID 1.2.840.10003.3.6)*, который позволяет проводить поиск информации в ресурсах НТИ в соответствии с общей моделью поиска Z39.50. Однако отсутствие схемы данных STAS не позволяет

[1] Напомним, что, в частности, RUSmarc определен с жесткой привязкой семантических элементов к структуре ISO-2709. Определение этих элементов сплошь и рядом противоречит принципам построения баз данных, изобилует повторениями и нелепостями, ориентированными не на обработку данных, а на их частное визуальное представление, именуемое «библиографической карточкой».

извлекать и просматривать найденную информацию в соответствии с моделью Z39.50. Аналогичная картина, как отмечалось выше, характерна и для обычной библиографической информации.

Учитывая, что:

- библиографическая научно-техническая информация не может быть адекватно представлена в принятых форматах семейства MARC, в частности, в RUSmarc;
- форматы семейства MARC не определены в Z39.50 как полноценные схемы данных;
- любой формат MARC определен вне Z39.50

возникает необходимость создания полноценной схемы данных для представления библиографической научно-технической информации в соответствии с моделью Z39.50.

В ОИГГМ СО РАН авторами была разработана специальная схема данных UIGGM с присвоением локальных идентификаторов схеме данных *(OID 1.2.840.10003.13.1000.155.1)*, набору меток *(OID 1.2.840.10003.14.1000.155.1)*. Эта схема полностью соответствует модели данных Z39.50. Семантика этой схемы данных соответствует семантике международного коммуникативного обменного формата МЕКОФ [3], т. к., во-первых, именно к этому формату привязана библиографическая НТИ, поставляемая ВИНИТИ и составляющая большую часть подобной информации в России, и, во-вторых, формат МЕКОФ является стандартом (ГОСТ 7.19-85, СТ СЭВ 4283-84) для подобного вида информации.

Литература

1. ANSI/NISO Z39.50-1995. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification / Z39.50 Maintenance Agency Offical Text for Z39.50-1995. — July 1995.
2. Жижимов О. Л., Мазов Н. А. Принципы построения распределенных информационных систем на основе протокола Z39.50. — ОИГГМ СО РАН, Новосибирск: ИВТ СО РАН. — 2004. — ISBN 5-9554-0017-6. — 361 с.
3. Система стандартов по информации, библиотечному и издательскому делу. Коммуникативный формат для обмена библиографическими данными на магнитной ленте. Содержание записи. ГОСТ 7.19-85 (СТ СЭВ 4283-84) // ГОСКОМСТАНДАРТ, Москва. — 1985.