Унификация построения и организации доступа к тезаурусам и классификационным схемам в распределенных информационных системах по протоколу Z39.50

Мазов Н.А.
ОИГГиМ СО РАН, Новосибирск, Россия
mazov@uiggm.nsc.ru

Жижимов О.Л.
ОИГГиМ СО РАН, Новосибирск, Россия
zhizhim@uiggm.nsc.ru

В настоящее время, как в России, так и за рубежом существует значительное количество различных тезаурусов и классификационных схем, используемых в информационно-библиотечной практике. Отраслевой тезаурус по сельскому хозяйству, используемый в ЦНСХБ, русско-английская версия известнейшего медицинского тезауруса MeSH, используемого в ГЦНМБ, Государственный рубрикатор НТИ, используемый так или иначе практически во всех информационных органах и библиотеках России - это лишь незначительный перечень из огромного количества тезаурусов и классификаторов, используемых в настоящее время. Как правило, во всех организациях, в которых используются ресурсы этого класса, они оформлены в машиночитаемые базы данных, поддерживаемые собственным программным обеспечением и как следствие этому - затруднена интеграция и совместное использование таких баз данных в распределенных информационных системах.

Насколько известно авторам, ряд организаций, располагающих такими базами данных (далее БДТК - базы данных тезаурусов и классификаторов), в настоящее время ведут работы над тем, чтобы обеспечить унифицированный сетевой доступ к ним. Цель этих работ – обеспечение возможности не только просмотра тезаурусов и классификаторов в удобном интерфейсе, но и активное использование их при поиске соответствующей информации в базах данных (далее БД), в том числе и в электронных каталогах.

Если для интеграции ведения, поиска и отображения библиографической информации в настоящее время наметился определенный сдвиг в области форматов, стандартных протоколов передачи и схем данных, то в области тезаурусов и классификационных схем к этому только подходят, о чем свидетельствует факт появления в середине 1999 года бэта-версии схемы данных Zthes, для работы с тезаурусом по протоколу Z39.50 [1].

Действительно, с развитием технологий построения больших распределенных информационных систем, включающих в себя множество различных баз данных, достаточно актуальным становится вопрос поиска информации в БД с использованием тезаурусов и классификационных схем. Более того, в распределенной информационной системе логично обеспечить доступ к БДТК в той же самой технологии, в которой осуществляется доступ к БД, т.е. в технологии "клиент-сервер" с использованием единого протокола Z39.50 [2].

Все вышеизложенное, актуальность этого вопроса для информационно-библиотечного сообщества России, а также предшествующий опыт работы авторов по разработке программного обеспечения доступа к БД по протоколу [3]–[4], позволило применить протокол Z39.50 для работы с тезаурусом по Наукам о Земле и рубрикатором БД ВИНИТИ в рамках информационной системыОИГГиМ СО РАН.

Исходя из общей идеологии Z39.50, доступ к любой базе данных, в том числе и к БДТК, должен осуществляться через единую стандартную схему данных, на которую должны быть корректно отображены все частные структуры БДТК. Проект такой схемы сегодня уже существует - Zthes (OID 1.2.840.10003.13.8) [1] и она активно обсуждается.

Авторами была предпринята попытка использования схемы данных Zthes для предоставления доступа к БДТК по протоколу Z39.50. Ниже изложены основные результаты этой работы.

В таблице 1 представлена схема Zthes, которая согласно [1] определяет абстрактную структуру записи

© Вторая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
26-28 сентября 2000г., Протвино

БДТК в иерархической схеме (колонка TagPath определяет полный путь в терминах tagsetM (1), tagsetG (2) и tagsetZthes (4)):

Tag Path	Occurrence	Element
(1,14)	M	termId
(2,1)	M	termName
(4,1)	O	termQualifier
(4,2)	O	termType
(2,17)	O	termNote
(2,20)	O	termLanguage
(1,15)	O	termCreatedDate
(1,27)	O	termCreatedBy
(1,16)	O	termModifiedDat
(1,28)	O	termModifiedBy
(4,4)	O, R	postings
(4,4)(2,36)	M	sourceDb
(4,4)(4,5)	O	fieldName
(4,4)(4,6)	M	hitCount
(2,30)	O, R	relation
(2,30)(4,3)	M	relationType
(2,30)(2,36)	O	sourceDb
(2,30)(1,14)	M	termId
(2,30)(2,1)	M	termName
(2,30)(4,1)	O	termQualifier
(2,30)(4,2)	O	termType
(2,30)(2,20)	O	termLanguage

Таблица 1: Абстрактная структура записи Zthes. Элементы: M - обязательный, O - факультативный, R – повторяемый

Каждая запись БДТК должна быть отображена на эту структуру. В частности, одна из статей тезауруса по Наукам о Земле (bromine) выглядит следующим образом (представление XML):

```
<zthes>
        <termId> 549BC38E </termId>
        <termName> bromine </termName>
        <termType> TT </termType>
        <termNote> Chemical element.
                Use bromine deposits for bromine
                as a commodity. </termNote >

        <relation>
                <relationType> UF </relationType>
                <termName> Br </termName>
                <termId> F90CF05F </termId>
        </relation>

        <relation>
                <relationType> BT </relationType>
                <termName> halogens </termName>
                <termId> CCEACFE6 </termId>
        </relation>

        <relation>
                <relationType> NT </relationType>
                <termName> bromide ion </termName>
                <termId> F8BB2A67 </termId>
        </relation>

        <relation>
                <relationType> RT </relationType>
                <termName> brines  </termName>
                <termId> AD0A65E3 </termId>
        </relation>

        <relation>
                <relationType> RT </relationType>
                <termName>  bromine deposits
                </termName>
                <termId> 4510AE36 </termId>
        </relation>
</zthes>
```

Аналогичное отображение можно осуществить и для записей БД классификационных схем.

Для апробации реальной работы были выбраны следующие БДТК, хранящиеся в СУБД CDS/ISIS (в скобках - имя базы данных):

- Тезаурус по Наукам о Земле (grefth)
- Тезаурус по сельскому хозяйству ЦНСХБ (agrith)
- Рубрикатор ГРНТИ (версия ГПНТБ России) (grnti)

Последние две БДТК находятся в настоящее в промышленной эксплуатации и были любезно предосталенны руководством соответствующих библиотек авторам для тестирования. Записи этих БДТК были отображены на схему Zthes штатными средствами сервера ZooPARK (v2.34) [5] и Z-ISIS - провайдера данных CDS/ISIS. С результатом работы можно ознакомиться через Интернет по протоколу Z39.50 на сервере geolibr.uiggm.nsc.ru:210 (имена баз данных указаны выше). Кроме того, доступ к этим БДТК может быть осуществлен через шлюз Z39.50 (http://geolibr.uiggm.nsc.ru/zgwn/) с удобным графическим интерфейсом для навигации. На рис.1 и рис.2 представлен интерфейс пользователя при работе с тезаурусом и рубрикатором соответственно.

Следует отметить, что сам по себе сетевой доступ к БДТК хотя и предоставляет интерес, но, на наш взгляд, не несет особого смысла без возможности одновременного выхода в поисковую систему по БД. Иными словами, просматривая статьи тезауруса или классификационной схемы, логично иметь возможность

проведения одновременного поиска в БД по соответствующим ключевым словам или кодам рубрик. Именно здесь протокол Z39.50, ввиду стандартизации поискового механизма, дает уникальную возможность подключать к параллельному поиску совершенно различные БД. На вышеуказанном шлюзе продемонстрирована возможность реализации этого механизма на примере рубрикатора ГРНТИ, коды которого сегодня присутствуют во многих БД (каталог ГПНТБ России, ГПНТБ СО РАН, БД ВИНИТИ и др.).

Результатом проведения вышеописанных работ явилось не только появление вполне работоспособного интерфейса доступа к БДТК с их интеграцией с БД (см. ссылку на шлюз выше), но и накопление опыта, суть которого может быть выражена в следующих тезисах:

- применение протокола Z39.50 для доступа к БДТК дает богатые возможности для построения распределенных информационных систем - интеграция БДТК и БД с обеспечением единого сетевого доступа по стандартным глобальным схемам;
- использование стандартных схем данных позволяет скрыть частные различия структур различных БДТК и обеспечить единый интерфейс без перегенерации последних;
- в распределенных информационных системах можно и нужно организовывать специализированные сервера для хранения БДТК с предоставлением к ним доступа по Z39.50 (чем меньше копий данных, тем проще поддерживать их синхронность);
- при построении конкретных библиографических БД следует как можно шире использовать привязку записей к различным тезаурусам и классификационным схемам. Отсутствие этой информации в БД исключает последние из единого информационного пространства в распределенной информационной системе;
- глобальная схема Zthes требует расширения для более адекватного отображения информации БДТК.

Последний тезис связан с тем фактом, что в схеме Zthes отсутствует корректная ссылка на базу данных БДТК, из которой выбирается терм. Поле sourceDB, предназначенное для этой цели, содержит лишь имя БД, но не содержит имени сервера и порта. Но в распределенной системе указание только имени БД недостаточно для ее однозначной идентификации. Расширение схемы позволит снять это ограничение при построении распределенной системы взаимосвязанных БДТК. Помимо этого, существует еще ряд замечаний к схеме Zthes. Поскольку в настоящее время происходит активное обсуждение готовящегося стандарта схемы Zthes, в котором авторы принимают участие, есть надежда, что в следующих версиях схемы эти ограничения будут сняты.

В заключение отметим, что изложенный в настоящей работе подход для доступа к БДТК позволяет также снять ряд ограничений по поиску информации в БД, доступных по протоколу Z39.50, с которыми иногда сталкиваются пользователи, а именно пользователь может и не знать расшифровок конкретных рубрикационных шифров.

Насколько известно авторам, настоящая разработка доступа к БДТК по протоколу Z39.50 является в России уникальной, хотя подобные работы в мире в последнее время проводятся [6]. Это лишний раз подтверждает ее актуальность не только для организаций Сибирского Отделения РАН, но и для других организаций, использующих в своей работе БДТК.

Библиография

[1] Mike Taylor. Zthes: A Z39.50 Profile for Thesaurus Navigation.Version 0.3b. http://lcweb.loc.gov/z3950/agency/profiles/zthes-03.

[2] 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.

[3] Жижимов О.Л., Мазов Н.А., Болванов А.Ю. Опыт построения распределенной информационной системы на базе протокола Z39.50. Матер. 6 Междунар. Конф. "Крым-99", т.1, стр. 249-252.

[4] Мазов Н.А., Жижимов О.Л. Интеграция Z39.50 и CDS/ISIS: состояние и перспективы развития. Матер. 6 Междунар. Конф."Крым-99", т.2, стр. 249-251.

[5] ZooPARK модульный сервер Z39.50. Версия 2.36. ОИГГиМ СО РАН. http://geolibr.uiggm.nsc.ru/doklads/Z-docs/ZooPARK.doc

[6] Index Data's thesaurus WWW-Z39.50 gateway. http://muffin.indexdata.dk/zthes/tbrowse.zap

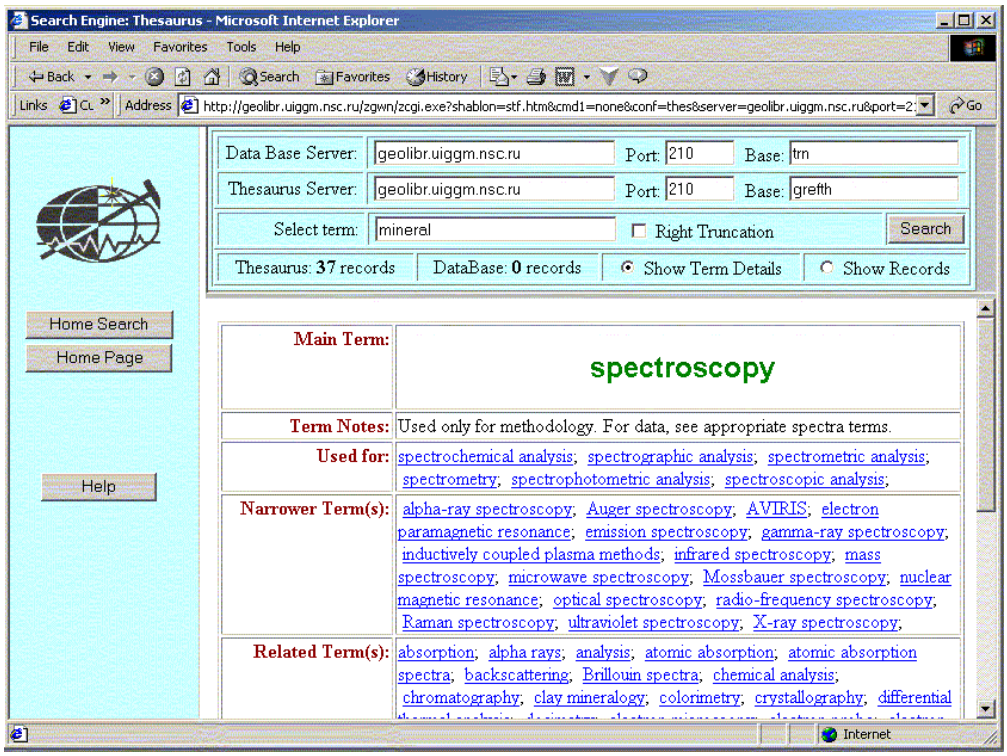

Рис.1: Интерфейс пользователя при навигации по тезаурусу

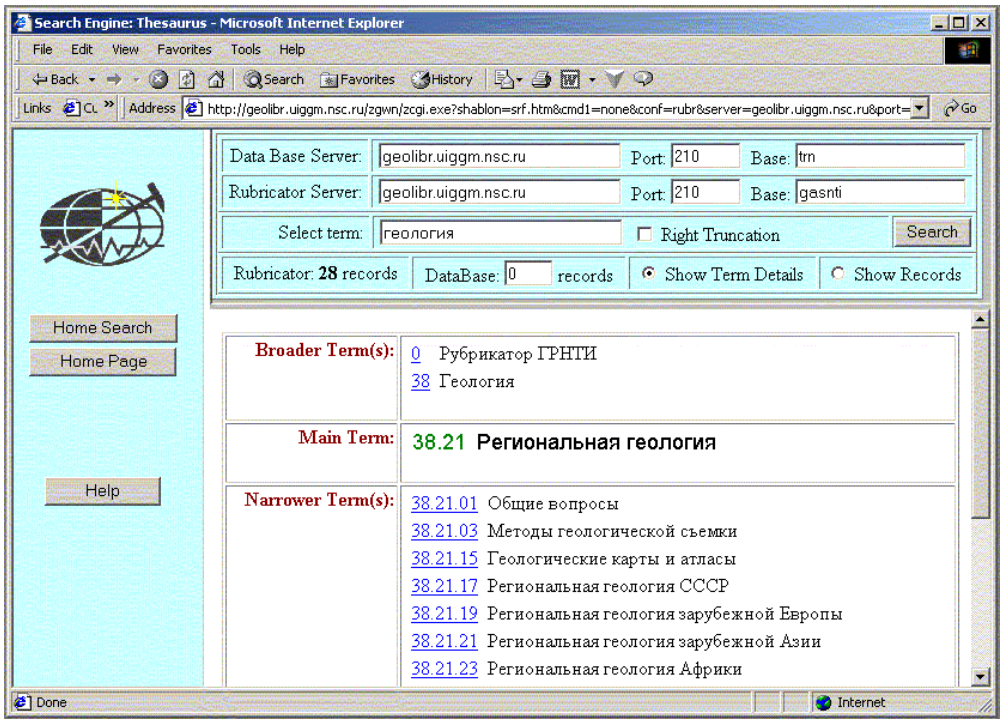

Рис.2: Интерфейс пользователя при навигации по рубрикатору