Структура сервисов управления метаданными для разнородных информационных систем♣

О.Л. Жижимов, И.А. Пестунов, А.М. Федотов

Институт вычислительных технологий СО РАН, Новосибирск
zhizhim@sbras.ru, pestunov@ict.nsc.ru, fedotov@sbras.ru

Аннотация

Рассматриваются различные аспекты организации сервисов управления метаданными в распределенных информационных системах, интегрирующих разнородную информацию для обеспечения научных исследований. Приводятся примеры реализации информационных систем.

В работе рассматриваются различные аспекты организации сервисов управления метаданными в распределенных информационных системах, интегрирующих разнородную информацию для обеспечения научных исследований, в частности, для задач исследования природных экосистем.

Список необходимых для управления метаданными сервисов включает:

- сервисы управления данными: сервисы каталогизации данных, т.е. создание вторичных информационных массивов, сервисы пакетной загрузки метаданных, сервисы заимствования метаданных из других информационных систем, сервисы синхронизации метаданных между разными информационными системами;
- сервисы распределенного поиска в массивах разнородных метаданных;
- сервисы извлечения метаданных в различных схемах и форматах;
- сервисы просмотра индексов;
- сервисы информирования о деталях конфигурации информационной системы (ИС) и всех ее компонентов;
- сервисы контроля доступа к данным и метаданным;
- сервисы предоставления доступа к контенту.

Ниже приводится краткая характеристика каждой группы сервисов и способов их реализации.

1. Сервисы управления данными

Эти сервисы необходимы для формирования информационных ресурсов с использованием и без использования диалоговых пользовательских интерфейсов. Сервисы позволяют использовать метаданные других информационных систем в диалоговом и пакетном режимах. Их функциональность должна обеспечивать поиск и извлечение метаданных из других систем, конвертирование полученных метаданных в схемы и структуры локальной системы. Обычно эти сервисы связаны с сервисами каталогизации и сервисами синхронизации метаданных [1].

Высокая степень интероперабельности информационных систем позволяет организовывать обмен информацией даже между гетерогенными системами.

В настоящее время активно применяется технология OAI [2], предполагающая поддержку специального сервиса, основанного на простом (поверх HTTP) протоколе и передаче XML-структур, содержащих инструкции и структурированные данные.

Реализация технологий OAI в ИС, например, обмен данными по протоколу OAI-PMH, позволяет строить отказоустойчивые информационные кластеры с реплицируемыми данными (см. рис. 1).

Заметим, что наделение ИС упомянутой функциональностью не связано с большими затратами, т.к. сервисы могут быть реализованы с помощью готовых программных компонентов. Однако эта функциональность переводит ИС на качественно новый уровень.

2. Сервисы распределенного поиска в массивах разнородных метаданных

Обычно разработчики обеспечивают поиск информации в ИС посредством визуальных графических интерфейсов. Это хорошо для пользователя-человека, но плохо для пользователя-системы. Для обеспечения функций поиска вне графических интерфейсов требуется поддержка специальных сетевых сервисов и языков запросов. В идеальном случае все информационные системы должны поддерживать единый поисковый профиль и единый язык запросов.

Для этой задачи широко распространенная модель поиска, основанная на SQL, мало подходит, так как SQL оперирует локально определенными терминами реляционных таблиц и полей.

Труды XIV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2011), Санкт-Петербург, Россия, 2011.

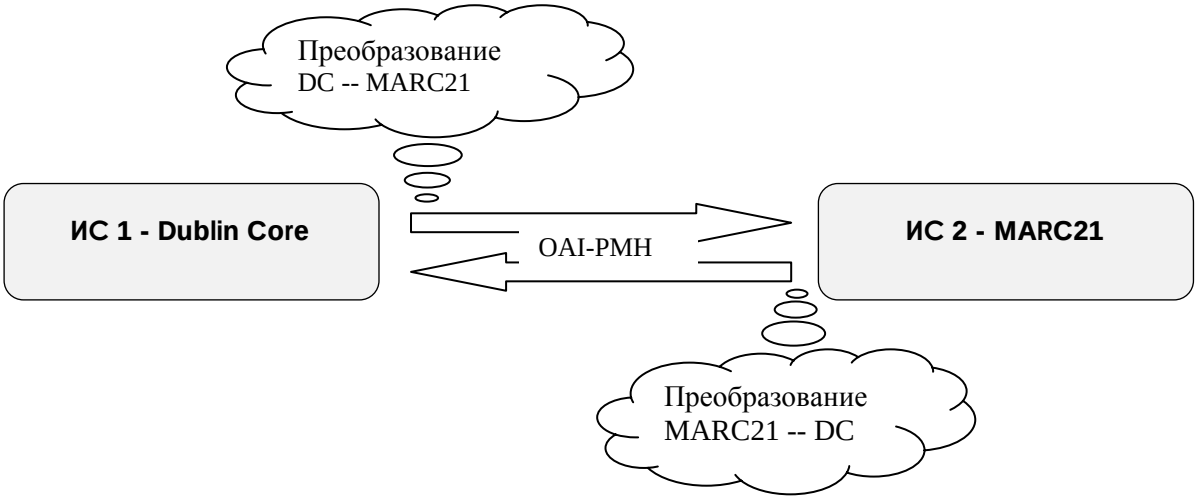

Рис. 1. Схема обмена данными по OAI-PMH

В настоящее время реализация парадигмы абстрактного поиска существует в виде нескольких моделей организации поисковых сервисов, например, модель Z39.50 [3,4] и более простая модель SRW/SRU [4,5]. У каждой из них есть свои достоинства и недостатки, но их эффективность подтверждается существованием больших распределенных гетерогенных информационных систем, в которых возможен сквозной поиск информации на основе абстрактных поисковых атрибутов.

Практическая реализация, например, основанная на технологиях XML сервисов SRW/SRU, не представляет для разработчиков большого труда, но в результате достигается новое качество ИС – возможность включения ее ресурсов в глобальные поисковые системы на более высоком уровне, чем уровень внешней индексации статических WEB-страниц другими системами.

Другие возможные типы поиска связаны с поиском по заданным шаблонам и поиском с привлечением онтологии. Поиск с привлечением онтологии является более интеллектуальным. Для его реализации требуется дополнительная информации о предметной области, включающая определения терминов, сущностей и связей. Следует отметить, что представление этой дополнительной информации должно соответствовать глобальным договоренностям – международным стандартам, иначе, поиск с привлечением словарей, тезаурусов и онтологии всегда будет ограничен текущей системой, а интероперабельность не будет реализована [6].

Сервисы извлечения метаданных в различных схемах и форматах

Информация, содержащаяся в ИС, при взаимодействии последней с другими информационными системами должна быть представлена в виде, в котором она может быть обработана. Для этой цели представление информации в виде WEB-страниц не является оправданным, т.к. для систем нужна структурированная информация. Переход от HTML-разметки документов к XML-документам существенно увеличивает эффективность дальнейшей их обработки.

Более строгие правила должны накладываться на представление вторичной информации – метаданных. Информационная система должна предоставлять метаданные в стандартных схемах (ISO-19115, METS, MARC21, DC, RUSmarc и т.п.) и стандартных форматах (XML, ISO2709, GRS-1 и др.).

Ниже приведен пример SRU-запроса на поиск термина «unix» в заголовках в базе данных DBTEST1 и извлечения найденных записей в количестве 1, начиная с первой, в схеме DC:

```
http://server:port/DBTEST1э?version=1.1&
   operation=searchRetrieve&
   query=dc.title=unix&startRecord=1&
   maximumRecords=1&recordSchema=dc
```

Сервисы просмотра индексов

Для формирования поисковых запросов может потребоваться информация о содержании текущих индексов в ИС. Для предоставления такой информации в ИС должен поддерживаться специальный сервис – сервис просмотра индексов. В технологии Z39.50 такой сервис обеспечивается обработкой специальных запросов (ScanRequest). Аналогичный сервис может предоставляться и в технологиях SRW/SRU. Пример запроса SRU на просмотр индекса dc.title с позиции «unix» для базы данных DBTEST выглядит следующим образом:

```
http://server:port/DBTEST1?version=1.1&
   operation=scan&scanClause=dc.title=unix
```

Сервисы информирования о деталях конфигурации информационной системы и всех ее компонент

Эволюция мировой информационной инфраструктуры имеет тенденцию к интеграции разрозненных информационных систем в единую, но распределенную систему. Однако заставить совокуп-

ность отдельных информационных систем функционировать как нечто связанное можно лишь на основе их полной интероперабельности. Эта интероперабельность должна, кроме всего прочего (стандарты, протоколы, запросы, схемы, форматы и т.п.), включать возможность взаимного информирования систем о своих функциональных возможностях и о своем информационном наполнении. Без этого информирования невозможно обеспечить свойство адаптивности информационной системы при интеграции ее в какой-либо гетерогенный кластер. Необходимы специальные сервисы, реализующие требуемую функциональность.

Следует заметить, что на обеспечение функциональной адаптивности информационных систем в части WEB-сервисов направлена технология на основе WSDL [7], в части описания программных интерфейсов – IDL. На обеспечение информационной адаптивности информационных систем ориентированы технологии на основе XML, RDF, OWL. Каждая из упомянутых технологий решает частную задачу обеспечения адаптивности для специальных (на основе XML) систем. Однако задача обеспечения адаптивности информационных систем, несомненно, намного шире.

Что касается специальных информационных систем, то существуют упоминавшиеся выше решения для систем на основе Z39.50 с использованием специальных стандартизованных системных сервисов (Explain). Через сервис Explain системы на основе Z39.50 могут обмениваться информацией о деталях своих конфигураций. Идеология систем на основе Z39.50 была частично сохранена при попытке реализовать функциональность Z39.50 в WEB-системах в виде технологий XML/SOAP/SRW – ZeeRex (см. http://explain.z3950.org).

Сервисы контроля доступа к данным и метаданным

Для распределенных информационных систем актуальными являются сервисы управления доступом к информационным ресурсам. Одним из сервисов этой группы является сервис аутентификации пользователей. В распределенной среде этот сервис проще всего реализовать на основе технологий LDAP, которые содержат технологию глобальной идентификации пользователей и методы их аутентификации. Несмотря на то, что сервисы LDAP основаны на специальном протоколе, их использование не представляет большой проблемы, т.к. сегодня поддержка LDAP встроена на уровне специальных библиотек в любую операционную систему. Для любителей технологий WEB-сервисов существуют решения на основе DSML (подмножество XML, ориентированное на структуры LDAP).

Сервисы предоставления доступа к контенту

Сервисы доступа к первичным информационным ресурсам зависят от типа этих ресурсов. Их функциональность должна обеспечивать информационные потребности пользователей конкретной информационной системы. Список этих сервисов может быть достаточно широк – от сервисов просмотра простых документов, до сервисов интеллектуальной обработки данных и предоставления пользователю результатов этой обработки.

В качестве примеров демонстрируются решения, реализованные полностью или частично в информационных системах СО РАН.

Пример 1

В целях обеспечения доступа потенциальных пользователей к спутниковым данным на базе Института вычислительных технологий (ИВТ) СО РАН создается Новосибирский узел сбора, хранения и обработки данных дистанционного зондирования [8]. К основным функциям узла относятся: телекоммуникационное обеспечение сбора данных, архивирование «сырых» данных, предварительная обработка данных, каталогизация обработанных данных, обеспечение оперативного и долговременного хранения обработанных данных, предоставление доступа к данным и тематическая обработка данных.

Основной поставщик спутниковых данных – Сибирский центр Государственного учреждения «Научно-исследовательский центр космической гидрометеорологии «Планета» (СЦ ГУ "НИЦ «Планета»"), который является крупнейшим за Уралом центром, занимающимся приемом и обработкой спутниковой информации.

Для передачи данных из СЦ ГУ "НИЦ «Планета»" в ИВТ СО РАН организована подсеть сбора данных сети передачи данных Сибирского отделения РАН, состоящая из двух сегментов – ГИС-сегмента локальной сети ИВТ СО РАН и сегмента СЦ ГУ "НИЦ «Планета»".

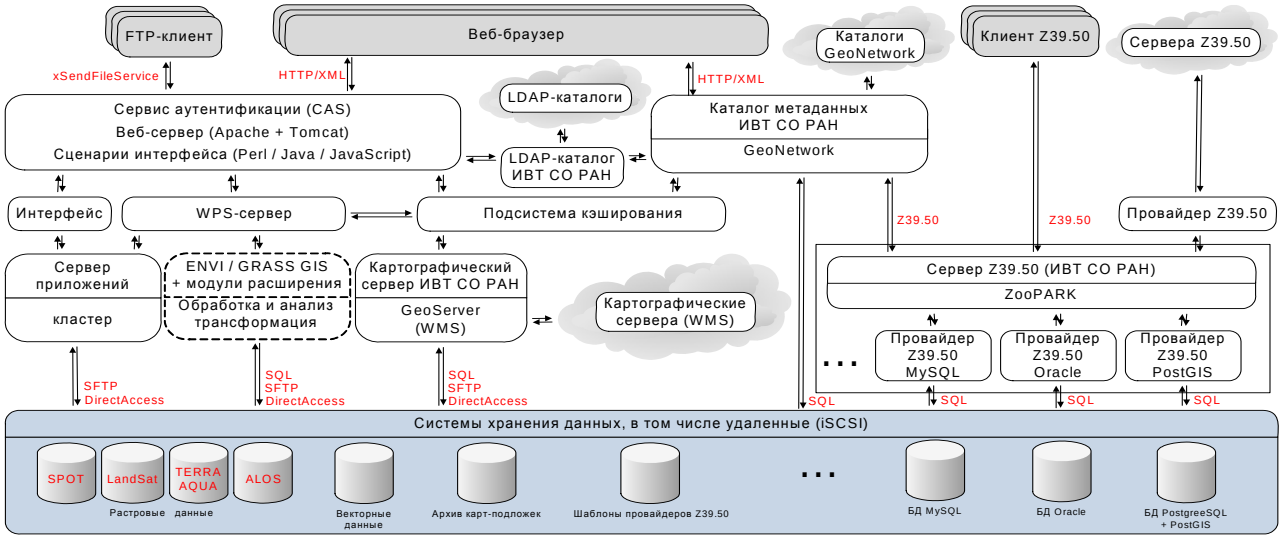

Рис. 2. Структура картографической информационной системы

В 2008 году на базе системы хранения данных ИВТ СО РАН создан каталог, который регулярно пополняется оперативными данными SPOT 4 [9]. Каталог включает также архивные данные со спутников серии LandSat на территорию РФ за 1982-2002 гг. На базе этого каталога создается сервис-ориентированная система, реализованная в виде базового набора приложений, работающих в среде сервера приложений Tomcat [10]. Подсистема пользовательских интерфейсов разработана с использованием технологий PHP/JavaScript. Доступ к системе реализован посредством модуля Central Authentication Service (CAS). Он позволяет организовать многоуровневую систему разграничения прав доступа с централизованной базой пользователей на основе LDAP-каталога Сибирского отделения РАН и реализовать практически индивидуальные настройки доступа к любому защищаемому ресурсу.

Система состоит из следующих функциональных блоков (см. рис. 2).

Центральным блоком системы является подсистема картографических сервисов, созданная на основе пакета GeoServer. Подсистема обеспечивает доступ к картографической информации, хранящейся в системе (базовые подложки, векторные слои, построенные по базам данных и др.). Для публикации динамических данных используется пакет UMN MapServer, который обеспечивает доступ к данным, формируемым в оперативном режиме, а также к пользовательским наборам данных.

Для расширения функциональности системы используется подсистема сервисов [11,12]. На основе WPS-сервера deegree, распространяемого по лицензии GPL, разработан модуль для интеграции в систему алгоритмов обработки пространственных данных. Он осуществляет интерпретацию входных и выходных данных согласно спецификации протокола WPS и выполняет функции контейнера для неограниченного числа WPS-процессов. Архитектура модуля представлена на рисунке 3.

Для обработки данных с помощью WPS-процесса пользователь вводит в клиентском приложении адрес WPS-сервера, после чего ему предоставляется список доступных процессов и их описания (метаданные на естественном языке). Выбрав необходимый алгоритм, пользователь указывает значения входных параметров в соответствии со спецификацией протокола WPS. Эта технология позволяет обеспечить широкому кругу потенциальных пользователей доступ к хранилищу современных наукоемких алгоритмов и вычислительным ресурсам, необходимым для оперативной обработки больших массивов разнородных данных.

Для обеспечения функционирования системы в распределенном режиме и интероперабельности по протоколам доступа к метаданным и их представлению в нее интегрированы модули поддержки протокола Z39.50. Поисковая система позволяет не только находить данные по метаданным, но и выполнять комплексные запросы.

С апреля 2010 г. к системе подключен комплекс по приему и обработке данных, принимаемых с платформ Terra/Aqua. Инфраструктура СЦ ГУ "НИЦ «Планета»" обеспечивает бесперебойный прием данных в режиме реального времени. Для приема данных с платформ Terra/Aqua задействованы две станции MEOS-POLAR, ведущие прием и аппаратную распаковку потока данных в автоматическом режиме. Расположение приемного комплекса обеспечивает прием данных, покрывающих Сибирь, часть Дальнего Востока и Якутии, а также территории Урала и Центральной России. Имеется возможность приема данных и с других активных в настоящее время платформ. Для решения задачи обработки поступающего потока данных на базе информационно-вычислительной инфраструктуры ИВТ СО РАН развернут вычислительный комплекс обработки потока «сырых» данных MODIS до продуктов уровня L2G/L3. Сборка и валидация вычислительного комплекса обработки данных MODIS

Рис. 3. Структурная схема подсистемы сервисов

была проведена в Центре космического мониторинга Алтайского государственного университета [13]. При адаптации комплекса сотрудниками ИВТ СО РАН созданы программные компоненты для обеспечения режима потоковой обработки и автоматической архивации данных. Для хранения поступающей информации и продуктов ее обработки используется промышленная система хранения данных EMC Clariion с подсистемой параллельного файлового доступа. Объем ежедневного «продукта» составляет ~ 35-50 Gb информации. Работа комплекса полностью автоматизирована и не требует вмешательства оператора, за исключением функций управления расписанием приема, корректировки параметров алгоритмов и контроля работы комплекса.

Функциональность системы постоянно расширяется как за счет расширения списка продуктов обработки данных MODIS, так и за счет подключения к системе новых потоков данных. Так, с октября 2010 г. развернута обработка данных гиперспектрометра AIRS, поступающих с платформы Aqua. Доступ к продуктам обработки обеспечивается с помощью подсистемы сервисов. Эта подсистема включает набор HTTP/FTP ресурсов, а также интерфейсы доступа к данным с использованием технологий геосервисов (KML/KMZ, WMS).

В настоящее время пользователями спутниковых данных являются более 30 организаций и институтов СО РАН. Получаемые данные используются для выполнения крупных интеграционных проектов.

Пример 2

Для иллюстрации обсуждаемых технологий можно привести схему организации электронной библиотеки (ЭБ) ИВТ СО РАН в части построения подсистемы научных публикаций сотрудников.

В качестве цифрового репозитория, интегрирующей данные различного типа, была выбрана система DSpace [14]. Информационная система DSpace обладает широкими возможностями по управлению цифровым контентом, обеспечивает поддержку обмена по OAI-PMH с расширяемым списком допустимых схем и форматов данных, но не содержит интерфейсов Z39.50, SRW/SRU.

Для реализации поддержки Z39.50 и SRW/SRU был использован программный комплекс ZooPARK [4], позволяющий работать с различными данными из различных источников на основе применения только стандартных схем и форматов данных. Сервер ZooPARK обеспечивает доступ к гетерогенной информации по протоколам Z39.50 и HTTP, поддерживает стандартные модели поиска Z39.50 и SRW/SRU, содержит встроенный шлюз для преобразований протоколов обмена и форматов данных.

На рисунке 4 показана общая схема информационной системы, реализующей электронную библиотеку. Наряду с компонентой, управляющим цифровым контентом (DSpace), система включает СУБД (PostgreSQL) для хранения библиографических метаданных. При этом библиографическая база метаданных синхронизируется с цифровым репозиторием по протоколу OAI-PMH, используя XML (MARCXML) представление схемы данных ГОСТ 7.19-2001, в которой научные публикации могут быть представлены более адекватно, чем в традиционной для России библиографической схеме RUSMARC.

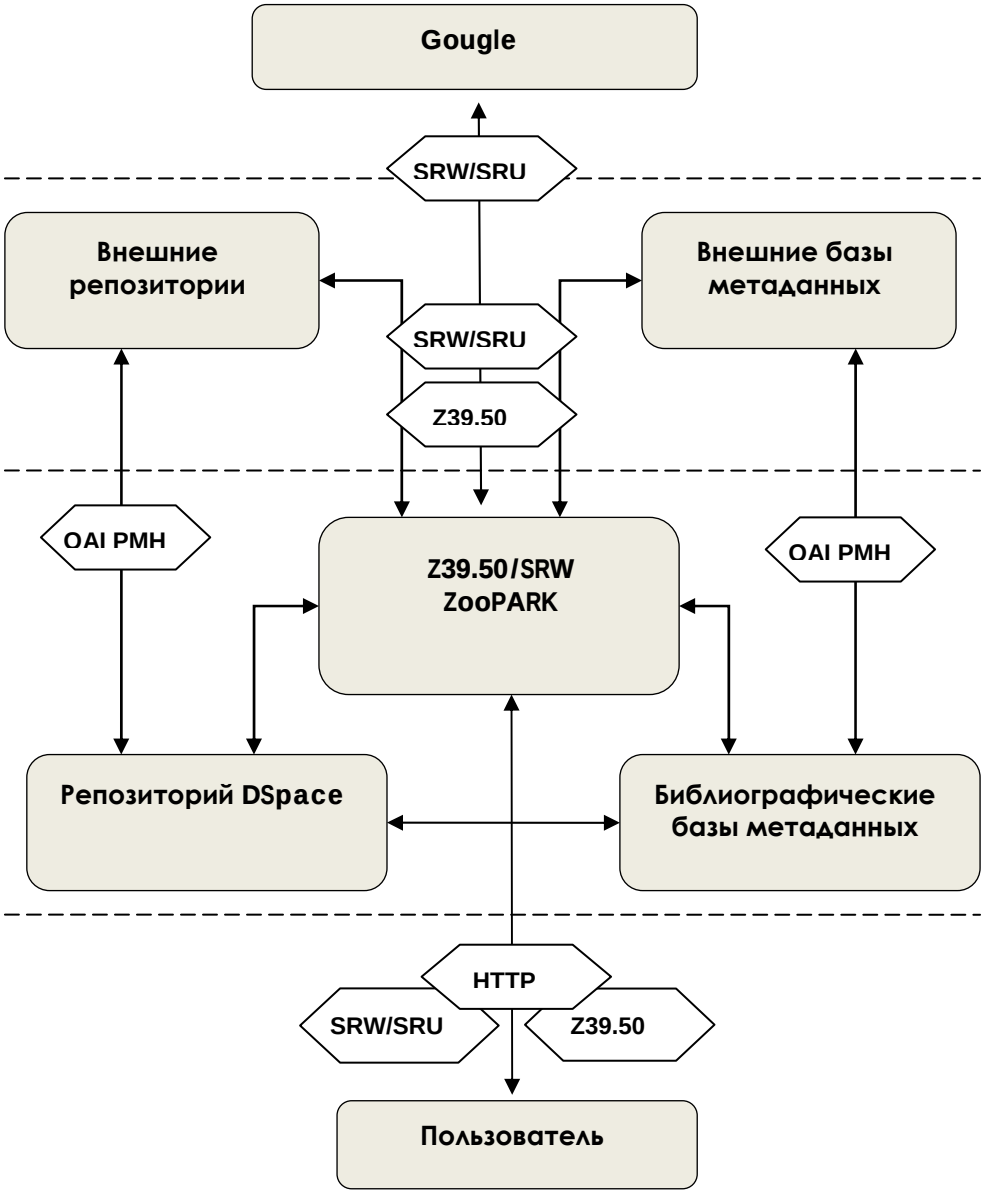

Рис. 4. Структура взаимодействие подсистем ЭБ СО РАН и внешних источников данных

Поскольку описания цифровых объектов в репозитории DSpace представлены в схеме Dublin Core, при синхронизации по OAI-PMH с библиографической базой публикаций выполняются преобразования DC → ГОСТ-7.19 и ГОСТ-7.19 → DC при прямом и обратном потоках соответственно.

Пополнение библиографической базы данных может осуществляться не только через интерактивные WEB-интерфейсы, но и другими методами:

– пакетной загрузки данных в формате XML и ISO2709 в схемах ГОСТ 7.19, MARC21, RUSMARC;
– прямого заимствования библиографических записей из внешних источников по протоколам Z39.50, SRW/SRU или HTTP. При этом различные внешние источники могут предоставлять записи в разных схемах, но чаще всего – в MARC21 и RUSMARC.

Поиск информации по различным критериям осуществляется через интерфейсы ZooPARK, который напрямую связан с метаданными DSpace и другими базами данных. Существенно, что одновременно поиск может происходить по разным информационным источникам. При этом поисковые запросы формулируются в терминах SRW/SRU, Z39.50 или CIP [15] (для географической информации). Это обеспечивает единый язык запросов для разных информационных систем, не привязанный к схемам и структурам данных конкретных целевых систем.

Возможность доступа к данным со стороны внешних поисковых систем, например, Google, обеспечивается поддержкой сервером ZooPARK протоколов SRW и SRU. Это делает доступной информацию, содержащуюся во внутренних СУБД, которая скрыта при обычном доступе через WEB для индексации внешними системами.

Литература

[1] Жижимов, О.Л. Некоторые заметки об эволюции цифровых репозиториев традиционных библиотек к полнофункциональным электронным библиотекам / Жижимов О.Л., Мазов

Н.А., Федотов А.М. // Вестник Владивостокского государственного университета экономики и сервиса. Территория новых возможностей. 2010. №3 (7). С. 55-63.
[2] The Open Archives Initiative Protocol for Metadata Harvesting // Protocol Ver. 2.0 2002-06-14, Doc. Ver. 2008-12-07 [Электронный ресурс]. – Режим доступа: http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.
[3] ANSI/NISO Z39.50-2003. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification // NISO Press, Bethesda, Maryland, U.S.A. ISBN 1-880124-55-6. 267 p.
[4] Жижимов, О.Л. Принципы построения распределенных информационных систем на основе протокола Z39.50 / Жижимов О.Л., Мазов Н.А. Новосибирск: ОИГГМ СО РАН; ИВТ СО РАН, 2004. 361 с.
[5] SRU – Search/Retrieval via URL // The Library of Congress - USA [Электронный ресурс]. – Режим доступа: http://www.loc.gov/standards/sru
[6] Шокин, Ю.И. Проблемы поиска информации / Шокин Ю.И., Федотов А.М., Барахнин В.Б. Новосибирск: Наука, 2010. 198 с.
[7] Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language // W3C Recommendation 26 June 2007 [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/wsdl20.
[8] Шокин, Ю.И. Корпоративная информационная система СО РАН сбора, хранения и обработки спутниковых данных / Шокин Ю.И., Пестунов И.А., Смирнов В.В., Синявский Ю.Н., Добротворский Д.И., Скачкова А.П. // Горный информационно-аналитический бюллетень. 2009. Отд. вып. “Кузбасс 2”. С. 3-10.
[9] Пестунов, И.А. Каталог пространственных данных для решения задач регионального мониторинга / Пестунов И.А., Смирнов В.В., Жижимов О.Л., Синявский Ю.Н., Скачкова А.П., Дубров И.С. // Вычисл. технологии. 2008. Т. 13. Вестн. КазНУ им. аль-Фараби. Серия: Математика, механика, информатика. 2008. № 4 (59). Совм. вып. Ч. III. С. 71-76.
[10] Жижимов, О.Л. Интеграция разнородных данных в задачах исследования природных экосистем / Жижимов О.Л., Молородов Ю.И., Пестунов И.А., Смирнов В.В., Федотов А.М. // Вестник НГУ. Серия: Информационные технологии. 2011. Т. 9. Вып. 1. С. 67-74
[11] Смирнов, В.В. Корпоративные картографические сервисы Сибирского отделения РАН / Смирнов В.В., Пестунов И.А., Добротворский Д.И. Синявский Ю.Н. // Горный информационно-аналитический бюллетень. 2009. Отд. вып. “Кузбасс 3”. С. 61-67.
[12] Добротворский, Д.И. Веб-сервисы для непараметрической классификации спутниковых данных / Добротворский Д.И., Куликова Е.А., Пестунов И.А., Синявский Ю.Н. // Сб. матер. VI Междунар. научн. конгресса «ГЕО-Сибирь-2010». Новосибирск: СГГА. 2010. Т. 1. Ч. 2. С. 171-175.
[13] Лагутин, А.А. Математические технологии оперативного регионального спутникового мониторинга характеристик атмосферы и подстилающей поверхности. Ч. 1. MODIS / Лагутин А.А., Никулин Ю.А., Жуков А.П. и др. // Вычислительные технологии. 2007. Т. 12. № 2. С. 67-89.
[14] Система DSpace [Электронный ресурс]. Режим доступа: http:// www.dspace.org.
[15] Catalogue Interoperability Protocol (CIP) Specification - Release B // CEOS/WGISS/ ICS/CIP-B, Is. 2.4.75. April 2005.

Structure of metadata services management for heterogeneous information systems

O.L. Zhizhimov, I.A. Pestunov, A.M. Fedotov

Various aspects of the organisation of services of management by metadata in the distributed information systems integrating the heterogeneous information for maintenance of scientifical researches are considered. Examples of realisation of information systems are presented.

* РФФИ гранты № 09-07-00277-а «Разработка технологий построения распределенных интегрируемых систем обработки, хранения и передачи информационных ресурсов на основе открытых спецификаций моделей данных»; № 10-07-00302-а «Разработка и анализ модели построения электронных библиотек на основе международных стандартов».