2014 | Volume 17| Issue 1|

Современные подходы к построению систем метаданных и поисковых сервисов для решения прикладных задач в области наук о Земле

М.А. Попов, Е.Б. Кудашев, С.Ю. Марков, С.А. Станкевич

Аннотация

Рассматриваются различные аспекты организации сервисов управления пространственными метаданными. Представлена модель использования гетерогенной пространственной информации. Поддержка интероперабельности обеспечена на основе преобразования метаданных для пространственной и непространственной информации. Предложена архитектура, используемая для актуализации метаданных об обновленных пространственных данных.

Ключевые слова: организация эффективного поиска, метаданные, пространственная и непространственная информация, гетерогенность, поддержка интероперабельности, сервисы управления метаданными.

Введение

Практика решения прикладных задач в области наук о Земле предполагает широкое использование различной пространственной и непространственной информации. При этом одной из основных проблем является организация эффективного поиска необходимых данных в огромном многообразии информационных ресурсов, прямо или косвенно связанных с решаемой задачей. Ключевую роль в обеспечении поиска необходимых данных играют пространственные и непространственные метаданные. Их правильное использование позволяет в значительной мере упростить решение проблемы гетерогенности данных, которая является одной из самых серьезных, особенно при решении комплексных задач. Учитывая то, что большая часть практических задач в области наук о Земле являются комплексными, становится понятной важность обеспечения эффективной работы с метаданными, особенно с точки зрения обеспечения интероперабельности информации в рамках решения этих задач.

Метаданные и гетерогенность информации

В [1] описана модель использования гетерогенных данных при решении задач устойчивого развития территорий (рис. 1). В этой модели предложен способ решения проблемы гетерогенности данных на основе сервис-ориентированной архитектуры. Одним из важнейших видов сервисов, от которых в значительной степени зависит общая эффективность представленной модели, являются сервисы согласования структуры метаданных. И это не случайно, ибо метаданные

являются основной информационной основой поиска. Метаданные позволяют кратко определить основные характеристики данных, а также оценить их пригодность для решаемой задачи. Поэтому современные подходы к сбору, хранению и распространению данных основаны на создании и параллельном сопровождении фактически двух наборов данных: собственно данных и метаданных им соответствующих. В связи с этим возникает несколько проблем, в частности:

- Структуры метаданных, используемых специалистами прикладных областей, формировались независимо от стандартов метаданных на пространственную информацию, что привело к их несоответствию, а во многих прикладных областях стандарты на метаинформацию могут вообще отсутствовать);
- Даже существующие стандарты на метаданные для пространственной и непространственной информации весьма многочисленны и не всегда совместимы между собой;
- Процедуры актуализации пространственных данных не синхронизируются с процедурами актуализации соответствующих им метаданных;
- Как пространственным, так и непространственным данным, которые используются при решении прикладных задач, присуща гетерогенность различного рода, что значительно затрудняет их совместное использование;
- Подходы к преодолению разного рода гетерогенности не стандартизированы, что ведет к несоответствию данных, полученных в результате различных преобразований данных;
- Отсутствие стандартизованных процедур подбора данных, необходимых для решения конкретных классов задач, обычно на практике это решается по разумению того, кто эту задачу решает;
- Отсутствие формализованных процедур трансляции запросов, ориентированных на решение прикладных задач, в запросы на получение необходимых пространственных данных;
- Метаданные о сервисах, необходимых пользователям при решении задач в области наук о Земле, не стандартизированы;
- Нет программных решений, позволяющих выполнять совместный поиск пространственных и непространственных данных и сервисов при решении прикладных задач в области наук о Земле.

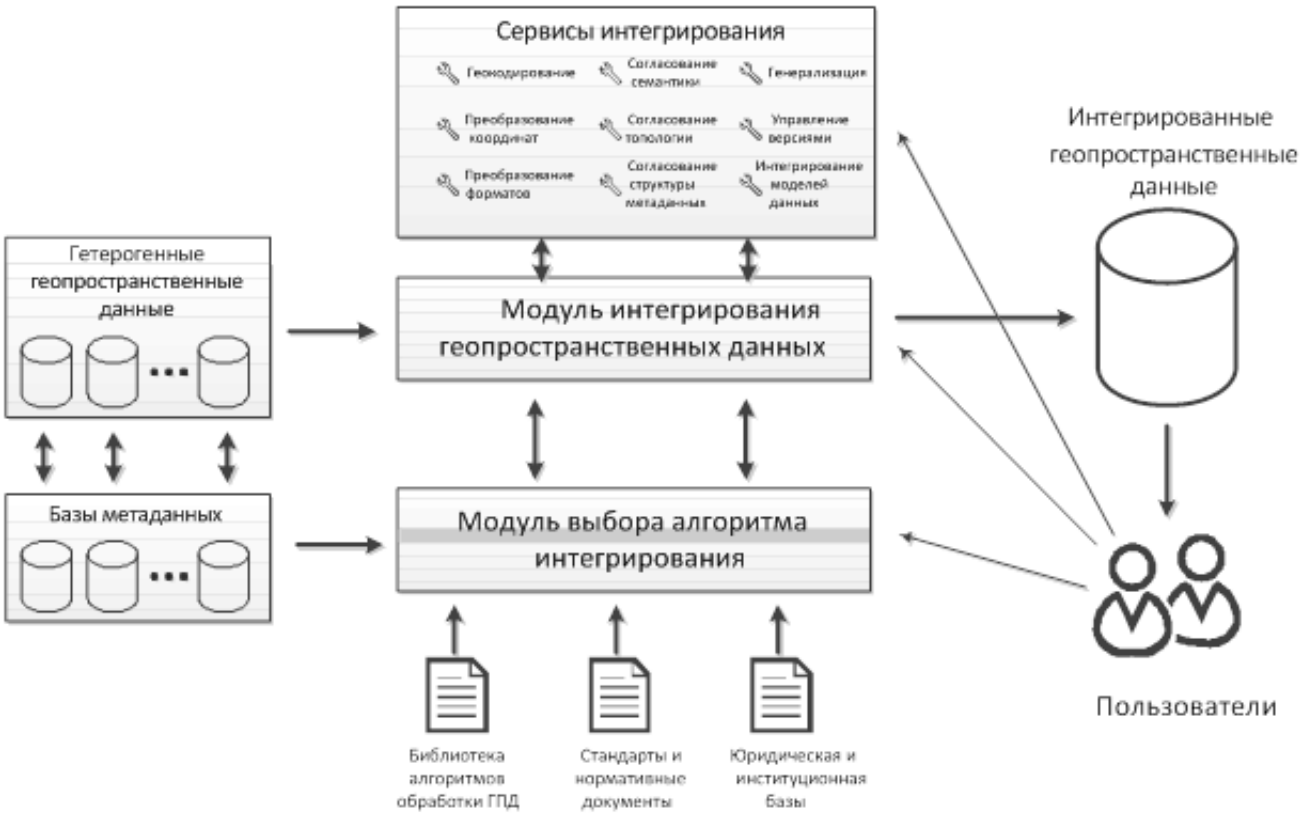

Рис. 1. Модель использования гетерогенной пространственной информации

Таким образом, можно выделить несколько основных направлений исследований, связанных с повышением эффективности поиска информации и работы с метаданными, а именно, подходы к стандартизации метаданных, а также методы реализации поисковых процедур в современной веб-среде.

Стандартизация метаданных

Стандартизация метаданных предполагает существование некоторой договоренности (де-юре или де-факто) об использовании некоторого стандартного подхода к описанию информационных ресурсов (стандарта метаданных) в некотором сообществе (например, сообществе исследователей наук о Земле либо любом другом профессиональном объединении ученых).

В настоящее время для реализации этого подхода могут быть использованы стандарты Dublin Core [2] – для непространственых данных, ISO 19115 [3] – для пространственных данных и WSDL (Web Service Definition Language) [4] – для веб-сервисов.

Однако совершенно очевидна утопичность этой идеи, поскольку в различных научных школах и организациях существуют некоторые традиции, в том числе, описания данных и сервисов, от которых не так легко отказаться. Кроме того, существуют огромные массивы уже накопленных данных и, соответственно, метаданных, которые хранятся в старых форматах, преобразование их в новые форматы потребует гигантских затрат ресурсов.

Поэтому в данном контексте с практической точки зрения лучше ориентироваться

на решение задачи интероперабельности (или информационной совместимости) данных и метаданных различных организаций.

Наиболее приемлемый путь решения данной проблемы для всех организаций – описание собственных корпоративных стандартов на метаданные в XML (eXtencible Markup Language) - совместимой форме, что даст возможность обеспечить однозначный формализованный информационный обмен с внешними системами. Учитывая факт, что все современные форматы метаданных, как для пространственных, так и непространственных данных, реализованы на основе XML, это дополнительно позволит реализовать процедуры интероперабельности данных различной природы, что крайне важно при решении комплексных задач в области наук о Земле.

С целью решения проблемы интероперабельности на основе преобразования метаданных для пространственной и непространственной информации, предложена архитектура, изображенная на рис. 2. В этой архитектуре профили корпоративных стандартов на пространственные метаданные (ПМД) подвергаются обработке в Блоке преобразования в XML с целью трансформирования данного профиля в XML-подобный вид (если, конечно, профиль изначально не был реализован на XML). В результате получаются схемы ПМД для всех корпоративных профилей стандартов на пространственные метаданные, реализованные на основе XML.

Далее полученные XML-профили обрабатываются в Блоке генерализации ПМД с целью формирования некого единого для всех ПМД профиля, основанного на стандарте ISO 19115. Для информационного обеспечения этого преобразования используются словари и классификаторы, сгенерированные Блоком анализа профилей ПМД.

Параллельно в предложенной архитектуре сами пространственные данные (по крайней мере, вновь создаваемые) трансформируются Блоком преобразования в GML в стандартный формат, общепринятый в международном сообществе геоматики, а именно GML (Geographic Markup Language) [5], поддерживаемый во всех основных программных продуктах, ориентированных на работу с пространственными данными. Кроме прочих преимуществ, данный подход позволяет реализовать процедуры автоматического извлечения метаданных из некоторого набора пространственных данных (выполняется Блоком автоматической генерации ПМД) и упростить решение проблемы синхронизации процедур актуализации данных и метаданных.

Сгенерированные в автоматическом режиме метаданные далее передаются в Блок генерализации ПМД и используются, прежде всего, для актуализации метаданных об обновленных пространственных данных.

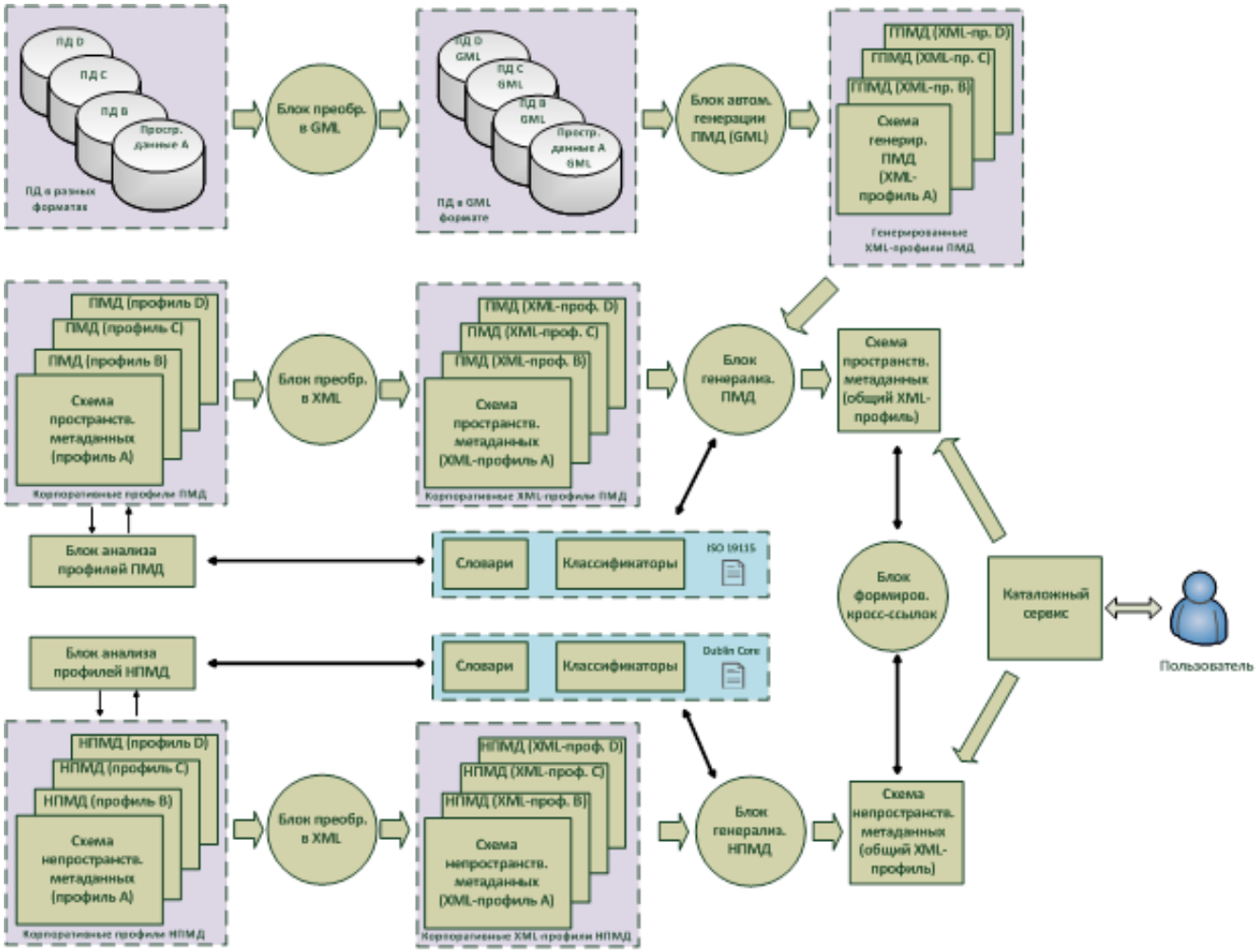

Рис. 2. Схема генерализации и актуализации метаданных

Обработка профилей непространственных метаданных (НПМД) происходит аналогично. Блок преобразования в XML трансформирует профили непространственных метаданных в стандартный XML-формат, блок генерализации НПМД формирует единую схему НПМД на основе формата Dublin Core, а также словарей и классификаторов, полученных Блоком анализа профилей НПМД. Для обеспечения связи ПМД и НПМД используется Блок формирования кросс-ссылок.

Конечный пользователь при решении задачи поиска пространственных и непространственных данных, необходимых для решения некоторой прикладной задачи, работает непосредственно в общими XML-профилями для ПМД и НПМД через каталожный сервис.

Реализация поисковых процедур данных для решения задач в области наук о Земле в WEB-среде

Как уже было отмечено, одной из основных особенностей процесса решения задач в области наук о Земле, является необходимость использования разнородных гетерогенных пространственных и непространственных данных. Поиск этих данных на практике является весьма нетривиальной проблемой, при этом основой эффективного поиска является:

- Правильное индексирование баз метаданных по необходимым атрибутам;
- Наличие стандартизированных словарей (тезаурусов) для базовых атрибутов метаданных, адаптированных под потребности конкретных предметных областей;
- Наличие кросс-ссылок не только между корпоративными стандартами на метаданные, но и между процедурами решаемых задач и требованиями к данным под эту задачу;
- Обеспечение возможности использования ресурсов уже существующих широко распространенных поисковых сервисов для потребностей прикладной задачи.

Реализация последнего из перечисленных требований на сегодня стала возможной после выхода в середине мая 2013 года платформы «Яндекс-острова» [6]. По мнению самих создателей этого веб-решения, его основное предназначение – реализация интерактивного поиска, результаты которого содержат конкретные, качественно релевантные ответы на поставленные пользователем вопросы. При этом пользователю нет нужды посещать сам сайт, на котором, возможно, находятся нужные ему ресурсы, результаты поиска ему выдаются сразу в диалоге с поисковой машиной. Таким образом, фактически при реализации поисковой процедуры выполняется интеллектуальный поиск по базе метаданных соответствующего сайта, которая описывает его информационные ресурсы.

Разумеется, реализация данной концепции зависит от качества удовлетворения первых трех первых требований, описанных выше. Очевидно, необходимо либо убедить владельцев ресурсов использовать некоторые принятые в отрасли стандарты на метаданные и набор словарей для самых важных атрибутов поиска, либо, в худшем случае, разработать процедуры распознавания используемых ими атрибутов и классификаций (парсинга) и их преобразования в стандартные. Основными принципами предложенного Яндекс подхода, являются:

- Релевантность интерфейса пользователя поставленному запросу, что предполагает автоматическое отсеивание ненужной информации;
- Выделение однотипных ответов на запрос в блоки – «острова», которые позволят сразу выделить информацию, наиболее востребованную пользователем;
- Поддержка различных аппаратных платформ.

Несомненным преимуществом представленной программной платформы «Яндекс-острова» является использование интерактивных веб-технологий (или web 2.0), которые предусматривают возможность для внешнего разработчика формировать собственные поисковые фильтры и вводить собственные корпоративные системы классификаций атрибутов, т.е. практически формировать поисковые запросы, адаптированные под решение задач конкретного поиска.

Кроме того, для решения комплексных поисковых задач могут быть привлечены уже существующие мощные механизмы и информационные фильтры, разработанные Яндексом. Практически, в данном случае внешний разработчик может построить свой тематический профиль информационного запроса в среде

Яндекс-поиска.

Заключение

Исследования, выполненные в настоящей работе, позволяют сделать вывод, что направления исследований в области совершенствования сервисов работы с пространственными и непространственными метаданными являются перспективными. Они обеспечат решение проблемы интероперабельности данных и сервисов, которые используются в сложных информационных инфраструктурах, создаваемых для решения комплексных прикладных задач в области наук о Земле.

Кроме того, представляется важным обеспечение эффективного использования мощных поисковых возможностей существующих веб-сервисов (Google, Yandex, Bing и др.) для работы с метаинформацией в комплексных программах с привлечение больших массивов пространственных и непространственных данных.Сегодня, когда бурное развитие получили «облачные» информационные сервисы, описанные подходы становятся еще более актуальными.

Благодарности:

Работа выполнена при финансовой поддержке РФФИ (грант № 11-07-00006 и грант № 14-07-00032). The work was financially supported by RFBR (grant No 11-07-00006 and grant No 14-07-00032).

Литература

1. Кудашев Е.Б., Марков С.Ю., Попов М.А. Использование гетерогенной пространственной информации при решении задач устойчивого развития территорий. - Российский Электронный журнал ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ. 2011. Том 14. Вып.3 [Электронный ресурс]. – Режим доступа: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2011/part3/PMK

2. Дублинское ядро (материалы сайта Википедия) [Электронный ресурс]. – Режим доступа: http://ru.wikipedia.org/wiki/Дублинское_ядро.

3. Geographic information – Metadata. (ISO/FDIS 19115:2003(E)). ISO/FDIS 19115:2003. – [Действительный от 23.03.2003]. – Geneva: ISO, 2003. – 224 p. – (Международный стандарт).

4. Christensen E, Curbera F, Meredith G, Weerawarana S (2001). Web Services Description Language (WSDL) 1.1. W3C, W3C Note 15 March 2001 [Электронный ресурс]. – Режим доступа: http://www.w3.org/TR/2001/NOTE-wsdl-20010315.

5. Geography Markup Language [Электронный ресурс]. – Режим доступа: http://www.opengeospatial.org/standards/gml.

6. Предварительная спецификация и обзор возможностей Яндекс Островов [Электронный ресурс]. – Режим доступа:

http://xn----8sbfgf4asmaticn7o.xn--p1ai/4-preliminary_specification_yandex_islands.html.

Об авторах

Попов М.А. - доктор технич. наук, профессор, замест. директора по науке, Научный Центр аэрокосмических исследований Земли ИГН НАН Украины, Киев, Украина;

Кудашев Ефим Борисович - доктор технич. наук, академик Российской Инженерной Академии, ведущ. научн. сотрудник, Институт космических исследований РАН, профессор МГУ им. М.В. Ломоносова, Москва, Российская Федерация e-mail: kudashev@iki.rssi.ru

Марков С.Ю. - канд. техн. наук, ст.н.сотр., Научный Центр аэрокосмических исследований Земли ИГН НАН Украины, Киев, Украина;

Станкевич С.А. - доктор технич. наук, профессор, Научный Центр аэрокосмических исследований Земли ИГН НАН Украины, Киев, Украина;