Использование метаданных в электронных библиотеках по науке. Ведение прикладных профилей метаданных по науке. Пример интеграции данных в распределенной Интегрированной Системе Информационных Ресурсов РАН

ЛопатенкоА.С.(ЦНТК РАН, andrei@derpi.tuwien.ac.at.,
Серебряков В.А. (ВЦ РАН, serebr@ccas.ru),
Филиппова А.А. (ВЦ РАН, filipova@ccas.ru)

Аннотация. В работе анализируются потребности в доступе к информации пользователей, работающих с научно-значимой информацией. Излагается, какие требования должны быть предъявлены к распределенным электронным библиотекам доступа к научным знаниям. Предлагаются методы создания федераций независимых электронных библиотек на основе технологий Semantic Web (SW).

Введение

Решение всего круга задач доступа к научным данным путем создания централизованных НИС (Научных Информационных Систем) невозможно, как показано в [1]. Развитие НИС демонстрирует, что во многих случаях научные данные будут храниться в различных информационных системах. Тем не менее, эти данные интересны для широконо круга научной общественности и пользователь системы должен получать ответ на запросы независимо от физического расположения данных.

Работы по созданию национальных и всевропейских систем доступа к научной информации показали, что в настоящий момент только в исключительных случаях (Дания, Беларусь, Исландия) возможно создания единых информационных систем, созданных на основе одной базы данных или на основе одной технологии. В других случаях создание национального единого интегрированного информационного пространства для науки требует доступа к данным независимых друг от друга административно и технологически информационных систем.

В настоящее время в России существует несколько приложений для науки и технологий. Среди них отметим следующие:

- Интегрированная Система Информационных Ресурсов РАН (ИСИР РАН) [2]. ИСИР РАН – портал научной информации Российской Академии Наук;
- Информационная система Сибирского Отделения Академии Наук;
- Информационные системы РФФИ. В РФФИ используются как минимум два приложения: 1) “гранты РФФИ” – приложения доступа к информации о грантах, спонсированных РФФИ; 2) Система Управления Инновационными Разработками (СУИР ИСИР) РФФИ, созданная в проекте ИСИР РАН;
- Приложение для доступа инвесторов к информации о перспективных разработках российских ученых Международного Научно-Технического Центра (МНТЦ).

Все эти системы существуют независимо друг от друга, принадлежат различным ведомствам, используют различные схемы данных, и методы работы пользователей с информацией в них различны.

Цель этой статьи - ответить на вопросы: возможно ли создание единого информационного пространства, объемлющего различные системы, и на основе каких методов это может быть сделано. В качестве основы создания единого информационного пространства рассматривается система ИСИР РАН.

Постановка задачи

При проектировании распределенной сети НИС в рамках концепции ИСИР был проанализирован опыт создания распределенных информационных систем для науки в рамках следующих проектов:

- ERGO (European Research Gateways Online) - проект CORDIS. Проект создания всевропейской сети доступа к научной информации.[3];
- NBOI (Netherlands Agency for Research Information);
- Работы Касельского Университета (Германия) [4];
- Работы по создания Финской Национальной Научной Сети [5];
- Ряд CRIS по инновационной деятельности в Европе;
- Работы проекта Desire (http://www.Desire.org) - проекта по созданию сети доступа к научной информации в Европе.

Анализ имеющихся проектов показал, что существует ряд типовых задач, с которыми приходится сталкиваться во всех проектах создания единой информационной системы для науки. В итоге, при проектировании распределенной информационной структуры задача была сформулирована следующим образом:

необходимо предоставить доступ к научной информации

- без устранения локального доступа к информации, а с добавлением нового типов доступа к уже имеющийся и опубликованной информации;
- уважая принцип независимости информационных систем, использовать распределенный подход вместо централизованного;
- использовать технологии интернет и открытые стандарты доступа к данным и создания программного обеспечения;

- пользователь должен иметь возможность получить доступ в распределенной системе к информации, представленной в каноническом виде, общем для всей сети (распределенной системы). Для получения более подробной и специфичной информации пользователь может обратиться к источнику данных. Если это возможно, система-источник данных должна предоставлять информацию о семантике и структуре специфичной информации;
- сеть должна быть эволюционной, изменения модели данных узла, значений данных не должны влиять на участие этого узла в сети.

Два последних требования к распределенной НИС приводят к тому, что в сети должны содержатся семантические описания элементов данных.

Для решения задачи создания национальной НИС требуется разрешить ряд более частных задач:

- Проанализировать и создать структурные и другие модели существующих или разрабатываемых баз данных о науке или технологиях;
- Определить и формализовать потребности пользователей НИС;
- Идентифицировать проблемы и возможности соединения существующих баз данных;
- Определить критерии выбора баз данных, которые могут стать частью сети ИСИР РАН;
- Выполнить согласование схем данных этих БД со схемой SW ИСИР РАН.

Соответственно, для решения задачи создания национальной НИС необходимо создать средства решения вышеупомянутых задач.

Прикладные профили

Для решении задач интероперабельности данных необходимо создать схемы метаданных, описывающих типы информационных ресурсов, их свойства. Каждая прикладная область оперирует собственными терминами, отношениями между информационными ресурсами, вследствие чего требуется создавать схему метаданных для каждой прикладной области.

В то же время ряд схем и определений метаданных могут быть использованы, возможно, с переопределениями, во многих прикладных областях, и требуется разработать методы использования разработанных схем метаданных для новых прикладных областей. Работы по созданию механизмов создания схем метаданных для конкретных прикладных областей, использующих имеющиеся схемы метаданных, ведутся в области прикладных профилей – схем метаданных, собранных из элементов различных схем (или пространств имен) и оптимизированных для использования в конкретной прикладной области [6].

Под прикладными профилями, оптимизированными для прикладной области понимается следующее. Значения терминов словарей метаданных (элементов схемы или пространств имен) прикладной области и используемых в профиле схем могут отличаться. В прикладной области могут использоваться отличные от принятых в исходной схеме словари классификации значений свойств данных. Разработка прикладного профиля для прикладной области (приложения) может включать переопределение терминов словарей метаданных, формальные или неформальные инструкции по использованию метаданных в прикладной области, пополнения исходных схем новыми словарями значений свойств, изменение или переопределение имеющихся. Такие прикладные профили называются оптимизированными.

Удачный опыт применения концепции прикладных профилей в проектах SCHEMAS, UKOLN DESIRE, Z39.50 application profiles, их близость к концепциям электронных библиотек (Warwick Framework) являются аргументами к использованию прикладных профилей в качестве метолодогии для создания схем метаданных Научных Информационных Систем.

Существует несколько методов создания прикладных профилей метаданных. В своей работе мы следовали работам Лагозе, Хантера [7], в которых прикладные профили определяются как комбинации RDF схем, выражающих семантику элементов, и XML схем, выражающих синтаксис. В работе Лагозе, Хантера разработана схема для комбинирования RDF Schema и XML Schema описания метаданых. Кроме того, в этой работе указано, что RDF схема может нести информацию о синтаксисе, а XML Schema может выражать информацию о семантике. При комбинировании этих двух схем возможны противоречеия и необходимы методы их разрешения, которые авторами не предлагаются.

В наших работах было разработано расширение прикладных профилей с целью формализовать описания схем, позволяя сделать их более машино-понимаемыми и облегчить работу пользователя с ними.

Работы по созданию распределеных схем в России и Австрии показали недостаток подхода проекта Harmony. В частности, в этом проекте в определение прикладного профиля не включены

- определения словарей и соотвественно межсловарных отношений (пояснение и объяснение необходимости включения словарей подробно изложено в главе тезаурусы);
- правила целостности на комбинированную схему (которые могут отличаться от суммы правил целостности на исходные схемы);
- описания конфликтов и методы их разрешения.

Прикладной профиль ISIR-SW определяет

- множество включенных пространств имен;
- определение семантических отношений между элементами различных пространств имен;
- определение новых значений элементов – переопределение семантики элементов. В прикладных профилях ISIR-SW может вводиться новое определение семантики элементов, например, вместо “примитивного” в определении OIL класса CERIF “проект”, в прикладном профиле проект может быть определен как деятельность, имеющая цели и начатая в определеный момент времени с планируемым завершением;
- определения словарей и описание содержимого словарей;
- определение семантических отношений между терминами различных словарей;
- правила целостности на комбинированную схему;
- описания конфликтов;
- правило поведения в конфликтной ситуации;
- Для совместного использования данных при интеграции различных схем необходимо согласовать схемы или форматы метаданных - определить семантические отношения между терминами словарей метаданных и словарей классификации значений данных, описать конфликты и правила их разрешения. Такие механизмы согласования данных предусмотрены в профилях ISIR-SW. Существует множество методов описания и нахождения семантических отношений между элементами схем (терминами словарей) [8,9].

Ввиду простоты реализации и невысоким требованиям к наполнению источников данных, в работах ISIR-SW выбран метод создания онтологии, общей для всех систем (ядра, канонической модели), и выражения семантики элементов остальных схем через каноническую. Кроме того, анализ имеющися НИС показывает, что этот метод подходит для Научных Информационных Систем. Многие из этих систем оперируют близкими терминами. В качестве канонической модели решено использовать модель данных CERIF (http://www.cordis.lu/cerif), которая одобрена Еврокомиссией и используется в некоторых европейских CRIS. Модель CERIF концептуально близка к модели ИСИР. CERIF предлагает достаточно богатое множество терминов для построения онтологии, используемой для НИС (в нашем прототипе онтологии CERIF 54 класса, 191 слово, в этих терминах почти полностью описываются известные нам отечественные и европейские системы). Для работы с прикладными профилями разработана RDF схема, позволяющая описывать прикладные профили со всеми их свойствами упомянутыми выше

Для работы с прикладными профилями разработано веб-приложение, позволяющее создавать, редактировать, использовать RDF схемы и словари и комбинировать их в прикладной профиль.

Между элементами RDF схем могут быть установлены семантические отношения. В качестве метода описания отношений между элементами RDF схемы используется Ontology Inference Layer [10] и MetaNet.

В качестве прототипа была выполнена интеграция данных систем ИСИР РАН (http://isir.ras.ru) и узла СУИР ИСИР РФФИ. ИСИР РАН хранит науко-значимую информацию о персонах, публикациях, проектах, организациях РАН[2]. СУИР ИСИР РФФИ хранит информацию о научных разработках по грантам РФФИ, имеющим потенциал коммерциализации. Так как ИСИР РАН - портал для доступа к информации о научных проектах в России и к информации о научной деятельности ученых и организаций РАН (70% проектов РФФИ), то данные РФФИ – потенциально часть информационного пространства ИСИР. Для интеграции данных был создан прикладной профиль ИСИР-СУИР РФФИ, состоящий из

- прототипов онтологий ИСИР и СУИР РФФИ, выраженных в терминах канонической модели CERIF;
- описания словарей CERIF, РФФИ (используется и в ИСИР РАН);
- RDF схемы CERIF;
- XML схемы для проектов CERIF.

Созданы RDF описания проектов СУИР РФФИ и части проектов ИСИР РАН. RDF описания информационных ресурсов были интегрированы в одну RDF базу данных (распределенную, ссылки между ресурсами по URL). Создан ориентированый на потребности пользователей ИСИР механизм запросов на поиск информации к распределенной базе данных. Использовались средства RDF запросов Squish (http://swordfish.rdfweb.org/rdfquery/), RDF Gateway(http://www.intellidimension.com/RDFGateway/gateway.asp).

Синдикация

Важным примером распределенных служб для Электронных Библиотек по науке является синдикация - поставка данных пользователю. Возможности композиции и трансформации содержимого из различных электронных библиотек делают возможным предоставить пользователю новую информацию, более соответствующую его информационным потребностям. Невозможность охвата всевозможных типов организаций коллекций над информацией требует разделения сервисов хранения информации, представления информации и представления коллекций информации.

Независимым коллективом авторов создан стандарт синдикации электронных ресурсов - RSS (RDF Site Summary). RSS – стандарт описания каналов поставки информации. Каждый RSS файл описывает один канал

информации в следующем виде: URI канала, название канала, описание канала, рисунок представления канала, набор меньших каналов информации, входящих в данных (items) . Для каждого подканала (item), указывается его URI, наименование, описание. В силу того, что в качестве источника информации указывается URI, возможно хранить описание канала независимо от источника информации. Указывая в качестве каналов или подканалов другие RSS ресурсы, возможно организовывать представления информации в виде графов, в частности деревьев (см., например, http://www.xmltree.com/).
В силу того, что RSS основывается на технология Semantic Web, в работах SW ИСИР было решено использовать именно эту технологию. Кроме того для RSS уже создан ряд средств по созданию RSS описаний и использованию их. RSS планируется использовать в научных системах Австрии и Британии.
RSS позволяет организовывать коллекции и точки доступа к ресурсам и коллекциям электронных библиотек в виде, независимом от точек доступа самих этих библиотек. Например, веб узел ИСИР РАН предоставляет головной странице доступ к сервисам поиска объектов ИСИР РАН и просмотра структуры организаций РАН. В RSS возможно организовать новую точку доступа в ИСИР, например, с возможностью непосредственного доступа к информации о проекте ИСИР и другим проектам авторов ИСИР.
Пример RSS точки доступа в ИСИР.

```
<!-- created  by Andrei Lopatenko (CSTIT RAS) -->
<rdf:RDF xmlns="http://purl.org/rss/1.0/" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    <channel rdf:about="http://isir.ras.ru">
        <title>ISIR RAS</title>
        <link>http://isir.ras.ru</link>
        <description>Portal of Russian Academy of Science. Access to information about persons, organization units,
projects, publications of RAS</description>
        <image rdf:resource="http://isir.ras.ru/img/start.gif"/>
        <items>
            <rdf:Seq>
                <rdf:li resource="http://isir.ras.ru/win/db/browse_adm.asp?P=.oi-.vi-.fi-.id-121"/>
                <rdf:li resource="http://isir.ras.ru/win/db/search_org.asp?P=.oi-.vi-.fi-.id-121"/>....
            </rdf:Seq>
        </items>
    </channel>...
    <item rdf:about="http://isir.ras.ru/win/db/browse_adm.asp?P=.oi-.vi-.fi-.id-121">
        <title>RAS Organization structure</title>
        <link>http://isir.ras.ru/win/db/browse_adm.asp?P=.oi-.vi-.fi-.id-121</link>
        <description> Organization structure of Russian Academy of Science </description>
    </item>...
</rdf:RDF>
```

Недостаток RSS - отсутствие формального описания содержимого каналов и методов доступа к информации в них. Отсутствие таких описаний не позволяет ввести каталоги каналов, искать каналы по их содержимому (например. “все каналы, которые описывают научные проекты или публикации, но не персоналии”). Кроме того, наличие формальных описаний содержимого канала позволит программным агентам и поисковым системам выполнять более точные операции классификации и поиска содержимого в интернет.
Проблему недостатка формального описание содержимого каналов возможно решить путем использования RDFS-OIL файлов для описания семантики содержимого каналов, и привязки RDFS-OIL описания к описанию канала.
Для того, чтобы описать семантику содержимого канала надо присвоить определению канала уникальный идентификатор. Это делается в файле канала использованием rdf:id для элемента определения канала.
Семантическое наполнение канала описывается в OIL или OIL-RDFS. В качестве канонической онтологии используется разработанная онтология для CERIF.
Например, сайт предоставляющий информацию о проектах, финансируемых РФФИ с датой начала проекта от 2000 года может быть описан в OIL следующим выражением:

```
class-def  RFBRProject
   subclass-of ResearchProject
   slot-constraint funded-by
          value-type rfbr
class-def NewRFBRProject
          subclass-of RFBRProject
   slot-constraint begin-date
```

value-type (min 2000)

или его эквивалент в OIL-RDFS

Для связывания описания канала с описанием его содержимого либо используется элемент описания канала, либо в RSS описаниях канала могут использоватся элементы RDFS-OIL или связывающие RDFS-OIL с RSS:

```
<channel rdf:about="http://isir.ras.ru">
<title>ISIR RAS</title>
…
<isir-sw:contentchannel rdf:resource="http://127.0.0.1/isir/sw/channeldef.rdf@NewRFBR"/>
</channel>
```

Ниже в качестве примера проводится простое расширение описания канала в соотвествии с механизмами расширения, описанными в [6]. Для описания семантики канала вводится элемент simple-content, который содержит описание содержимого в сответствии со словарем:

```
<item rdf:about="http://isir.ras.ru/win/db/browse_adm.asp?P=.oi-.vi-.fi-.id-121">
<title>RAS Organization structure</title>
                <link>http://isir.ras.ru/win/db/browse_adm.asp?P=.oi-.vi-.fi-.id-121
 </link>
<description>
  Organization structure of Russian Academy of Science
 </description>
<isir-sw:simple-content isir-sw:about="Organization"/>
</item>
```

Имея такие описания каналов возможно каталогизировать источники информации, искать нужные источники. Пример. Поиск каналов, содержащих информацию об организациях над RDF базой данных каналов. Язык SQUISH.

```
SELECT ?title, ?url
FROM http://127.0.0.1/ISIR-SW/channels.rdf
WHERE
(rss::item ?x ?y)
(rss::title ?y ?t)
(rss::link ?y ?url)
(isir-sw::about ?y ?cont)
AND ?cont ~ Organization
USING rss FOR http://purl.org/rss/1.0/,
isir-sw FOR http://127.0.0.1/ISIR-SW/adddef.rdfs
```

Выводы

Использование Semantic Web технологий позволяет

- упростить построение распределенных электронных библиотек;
- создавать распределенные структуры поиска информации в гетерогенных источниках данных;
- создавать распределенные электронные библиотеки в которых службы представления данных, организация коллекций будут независимы от служб хранения данных.

Методы прикладных профилей и использования канонического ядра

- пригодны для задач объединения распределенных научных данных;
- позволяют решать задачи интеграции имеющихся и новых источников научной информации в имеющиеся распределенные системы такие как ИСИР РАН;
- упрощают циклы разработки приложений поиска данных в гетерогенных информационных системах;
- с использованием технологий Semantic Web пригодны для задач согласования форматов метаданных (определения семантических отношений между терминами словарей метаданных и словарей классификации значений данных, описания конфликтов и правил их разрешения).

Литература

[1] М. В. Кулагин, А. С. Лопатенко "Научные информационные системы и электронные библиотеки . Потребность в интеграции", Электронные библиотеки-2001, Петрозаводск

[2] А. Н. Бездушный , А. Б. Жижченко , М. В. Кулагин , В. А. Серебряков, "Интегрированная система информационных ресурсов РАН и технология разработки цифровых библиотек". Программирование , 4 2000

[3] "Final ERGO report to the INNOVATION Programme Committee", 1996

[4] W. Adamchak, H. Begemann, S. Stefani, "Research report online as portal to a wider CRIS", EuroCRIS-2000

[5] S. Laitinen, S. Pirjo, K. Tirronen, “Development of Current Research Information Systems in Finland”, EuroCRIS-2000
[6] R. Heery, M. Patel, "Application Profiles: mixing and matching metadata schemas", Ariadne Issue 25, September 2000.
[7] J. Hunter, C. Lagoze, "Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles", WWW10, Hong Kong, May 2001
[8] M. Doerr, Semantic Problems of Thesaurus Mapping, Journal of Digital information, volume 1 issue 8
[9] M. Buckland, “VOCABULARY AS A CENTRAL CONCEPT IN LIBRARY AND INFORMATION SCIENCE”, “Digital Libraries: Interdisciplinary Concepts, Challenges, and Opportunities. Proceedings of the Third International Conference on Conceptions of Library and Information Science”, Dubrovnik, 1999
[10] Ontology Inference Layer homepage (http://www.ontoknowledge.org/oil/)

METADATA USAGE IN DIGITAL LIBRARIES FOR RESEARCH AND TECHNOLOGY. CREATING AND SUPPORT OF APPLICATION PROFILES FOR SCIENCE.

Lopatenko A. A. (CSTIT RAS), Serebryakov V. A. (CC RAS), Filipova A. A. (CC RAS)

Researchers' requirements are analyzed to access R&D data in distributed independent heterogenous systems for science. Based on these requirements demands are defined to architecture of distributed Digital Libraries for R&D. Methods are described of development metadata schema for applications based on application profiles adopted for Research Information Systems. Shortages of already used application profiles methods are shown, and a method is proposed in the article that matches requirements of developers of Digital Libraries for scince more then previosly used. Authors propose the Semantic Web-based solutions for development of national-wide architectures for access to research data. Application of RDF for knowledge representation of R&D is proposed and illustrated. Extension of RSS (RDF Site Summary) for advanced syndication in R&D is submited.