Доклад конференции EVA-2000

"Электронная конвергенция: новые технологии в музеях, галереях, библиотеках и архивах"

Видео версия доклада(http://www.a-z.ru/eva/4/12.ram)

Дальнейшее продолжение тема получила в работах автора по созданию прикладных профилей метаданных для науки, в создание RDF схемы для CERIF и в работах по применению семантических описаний метаданных для науки.

См. метаданные для науки (http://derpi.tuwien.ac.at/~andrei/Metadata_Science.htm)

Метаданные ИСИР: определение и использование.

А. Н. Бездушный, А. М. Серебряков, А. А.Филиппова
Вычислительный центр РАН
Адрес: Москва, ул. Вавилова, 40
Тел. (095)1355471
А. С. Лопатенко
Центр научных телекоммуникаций и информационных технологий РАН
Адрес: Москва, ул. Губкина, 8
Тел. (095)9383938

Для современных сложных информационных систем необходима поддержка метаданных – информации, предназначенной для анализа, дизайна, развития и использования информационных систем[1]. Для систем класса цифровых библиотек, каковой является ИСИР РАН (Интегрированная система информационных ресурсов РАН), метаданные особо необходимы для задач управления системой, корректного использования и понимания данных системы.

С точки зрения ИСИР РАН метаданные делятся на два типа:

- метаданные, описывающие информационную, функциональную и другие схемы системы;
- метаданные, описывающие информационные ресурсы.

Остановимся подробнее на каждом из типов метаданных.

Метаданные схемы системы.

Для описания **информационной модели** ИСИР РАН используется *стандарт Object Management Group XML Metadata Interchange (OMG XMI)*. Этот стандарт позволяет описывать классы объектов, представляющие типы информационных ресурсов ИСИР РАН, атрибуты и отношения между различными типами ресурсов.

Этот стандарт используется в ИСИР РАН для следующих целей:

- система экспорта/импорта данных использует XMI-представление для анализа и понимания информационной модели ИСИР, создания объектного API к ресурсам ИСИР (то есть API, реализующего функции извлечения данных о ресурсах ИСИР и изменения этих ресурсов);
- в будущем возможно использование CASE-средств, понимающих XMI. Это позволит автоматически настраивать на изменения в модели данных ИСИР другие информационные системы, работающие с ИСИР.
- Стандарт OMG XMI был выбран по следующим причинам:
- стандарт является открытым, получил активную поддержку среди разработчиков средств дизайна и CASE средств, в научной общественности;
- стандарт базируется в качестве средства синтаксического представления на XML – открытом часто используемом стандарте. Разработчики ИСИР имеют большой опыт работы с XML-данными;
- элементы OMG XMI имеют определенную семантику, близкую к семантике элементов ИСИР РАН. Мощность метамодели стандарта достаточна для описания схемы данных и функциональной модели ИСИР РАН

Для описания **схемы данных** ИСИР РАН используется формат *OMG Common Warehouse Model Interchange*, базирующийся на XMI. Этот формат позволяет описывать многие аспекты как структуры системы, так и операций трансформации, загрузки и выгрузки данных из них.

Этот формат используется:

-

для настройки системы импорта/экспорта данных на схему БД РАН;
- для описания трансформации данных из схемы ИСИР РАН в другие схемы.

В дальнейшем его использование позволит не менять код систем, работающих с ИСИР, а менять только описания БД. Кроме того, многие современные производители и разработчики РСУБД и решений хранилищ данных начинают использовать этот стандарт, что позволит использовать эти технологии в качестве средств проектирования и разработки ИСИР РАН.

Данный стандарт был выбран по следующим причинам:

- стандарт получил широкую поддержку среди разработчиков РСУБД и производителей ПО в близких к РСУБД областях (Date Warehousing, OLAP);
- стандарт имеет метамодель для представления данных о реляционных схемах, достаточную для полного описания базы данных ИСИР РАН, за исключением особенностей, свойственных реализациям на конкретных РСУБД. Для описания таких возможностей планируется создавать расширения CWM;
- стандарт имеет метамодель, достаточную для решения базовых задач обмена данными ИСИР РАН. Для решения сложных задач разработаны его расширения;
- стандарт базируется на XML. Преимущества использования XML описаны выше;
- модель данного стандарта близка к модели другого популярного стандарта для описания моделей данных, реляционных схем, схем обмена данными и др. – Metadata Coalition Open Informational Model (MDC OIM). В будущем это позволит выражать схему ИСИР в MDC OIM и использовать для хранения метаданных распространенные репозитории.

Прототип системы обмена данными выполняет задачи настраиваемого представления данных ИСИР в виде объектного API, экспорта данных ИСИР из БД в XML/RDF-представление (база данных – источник, XML/RDF - приемник данных). Данный прототип в настоящее время реализован на языке Java с использованием технологии Enterprise JavaBeans.

Данный прототип позволяет экспортировать информационные ресурсы в следующие модели данных:

- разработанное специально для ИСИР XML-представление информационных ресурсов;
- разработанное специально для ИСИР RDF-представление информационных ресурсов (это представление согласованно между БЕН и ИСИР РАН).

При экспорте для выражения семантики могут быть использованы различные стандарты.

К настоящему моменту система настроена на:

- Open Archives Metadata Set (OAMS) представления данных о e-print (Santa Fe convention Open Archives Initiative);
- DC (Dublin Core – международная инициатива представления данных о публикациях);
- vCard (электронная визитная карточка);
- CERIF (европейская инициатива формата описания проектов);
- собственное пространство имен.

Метаданные ресурсов.

При разработке модели данных ИСИР РАН одними из основных требований к ней были:

- возможность адекватного представления информации, относящейся к Российской Академии Наук;
- интероперабельность с другими системами, содержащими аналогичную информацию.

С учетом этих требований, в модель были включены ресурсы следующих типов: “организация”, “персона”, “проект” и “публикация”. Разработка структуры ресурсов требует определения характерных свойств каждой информационной сущности, достаточно полно описывающих ее. Эта задача усложняется требованием интероперабельности. Обеспечение интероперабельности является одной из наиболее важных задач, так как значительные информационные ресурсы уже представлены в цифровом виде, и основная проблема, возникающая при их использовании, заключается в отсутствии связей между ними. Для обеспечения взаимодействия с широким кругом информационных систем необходимо, чтобы модель была совместима с общепринятыми стандартами представления данных. Нами были изучены предложения и стандарты метаописания информационных ресурсов, представленных в ИСИР РАН, например: Dublin Core, MARC, GILS, CERIF, vCard.

В качестве основы для реализации ресурса “публикация” из множества имеющихся вариантов был выбран Dublin Core. Этот выбор обусловлен следующими преимуществами этого стандарта:

- набор основных семантических элементов компактен и, в то же время, позволяет задавать практически все требуемые атрибуты;
- семантика каждого элемента может быть уточнена с помощью квалификаторов, как стандартных, известных и понятных всем, так и специально разработанных для точной спецификации семантического смысла определенного атрибута при обмене данными внутри небольшого сообщества;
- в стандарте заложена возможность использования различных семантических схем, словарей и т. п.
- определен механизм, позволяющий извлечь информацию из описания, использующего нестандартные расширения пространства имен;
- стандарт получает все более широкое распространение в мировом сообществе.

Модель данных публикации ИСИР РАН позволяет задавать любой базовый элемент Dublin Core (Заглавие, Автор и т. д.). Предусмотрена возможность использования квалификаторов, уточняющих семантику базовых элементов (например, Параллельное заглавие, Участник-Редактор и т. п.). Это позволяет достаточно легко обмениваться библиографической информацией на основе этого стандарта. Однако серьезным препятствием для интероперабельности этой подсхемы модели ИСИР РАН с другими системами является то обстоятельство, что большинство этих систем рассматривает отдельные свойства публикаций, такие как “автор”, “издатель”, “источник”, как обычные текстовые атрибуты, в то время как они, по сути, являются связями с другими сущностями (персонами, организациями, публикациями). Такие модели не противоречит Dublin Core, но приводит к определенной несовместимости их с моделью ИСИР РАН и, в частности, к неоднозначности при интеграции данных в систему.

Для представления информации о научных проектах предназначен стандарт CERIF. Он основан на модели данных, которая включает сущности “проект”, “организация” и “персона”, связи между ними, а также атрибуты этих сущностей. Стандартом определяется три уровня детализации при описании ресурсов:

1. Полное описание ресурсов – содержит расширенный набор атрибутов, позволяющий описывать различные схемы ресурсов;
2. Набор атрибутов для осуществления обмена данными между различными системами;
3. Сокращенный набор атрибутов – метаописание ресурсов.

В модели ИСИР РАН ресурс “проект” имеет следующие свойства: даты начала и окончания проекта, код проекта, тип, ключевые слова, описание. Также в нее включена связь “участник” между проектом и организацией/персоной, имеющая атрибут “тип участия”. Таким образом, взаимодействие с другими системами на основе CERIF реализуется достаточно легко. Тем не менее, в настоящий момент стандарт реализован в ограниченном варианте (так, в ИСИР РАН не определено понятие “результат проекта”), и модель ресурса “проект” будет развиваться. Примечательно, что модель данных, предлагаемая CERIF, явно определяет персоналии и организации как самостоятельные сущности, что хорошо согласуется с моделью ИСИР РАН и делает информационный обмен еще проще.

Для представления информации о персоналиях существует значительное количество форматов. Тем не менее, достаточно широко используется только стандарт vCard, разработанный для унификации обмена данными между программными приложениями и информационными системами. Основными понятиями формата являются: *поток данных* и *объект vCard*. Поток данных – это набор vCard–объектов, которым обмениваются системы. vCard-объект – это набор предопределенных атрибутов, описывающих некоторый ресурс. Атрибут определяется своим именем, возможно, набором параметров, и значением. Параметры конкретизируют значение атрибута – например, для атрибута “телефон” параметры уточняют его значение: домашний, рабочий. Набор элементов и их параметров четко определен в стандарте и не расширяем. Формат позволяет создавать “электронные визитные карточки” для объектов типа персона.

Однако после анализа имеющихся данных было установлено, что семантика vCard недостаточна для представления всех необходимых атрибутов. Например, научная степень и академический статус персоны не имеют семантически близких элементов в vCard. В связи с этим,

модели ресурсов “персона” и “организация” были разработаны так, чтобы максимально точно сохранить модель, лежащую в основе имевшихся исходных данных. Ресурс типа “организация” имеет атрибуты: название, аббревиатура, адрес, телефон, электронный адрес, вид организации, направление деятельности, историческая справка, схема проезда, ключевые слова, фотография. Ресурс типа “персона” обладает атрибутами: фамилия, имя, отчество, научная степень, научное звание, академический статус, специальность ВАК, ключевые слова, направление деятельности, фотография. Кроме того, между этими типами ресурсов установлена связь “должность”, обладающую атрибутами: название, вид, телефон, электронный адрес.

Основная цель поддержки vCard — экспорт данных в формате, понятном большому числу систем. В настоящий момент выполняется экспорт ограниченного набора атрибутов. В дальнейшем этот набор будет расширен, а также реализована возможность импорта данных из этого формата.

Следует отметить, что каждый стандарт предлагает собственную модель данных, часто и собственный синтаксис для записи информации. Подход ИСИР РАН заключается в использовании для обмена метаданными единой модели данных и синтаксиса, определяемых RDF[2]. Только семантика атрибутов тех или иных ресурсов берется из соответствующего стандарта. Если найти подходящий элемент в стандартных пространствах имен не удается, можно создать собственное пространство, определив его посредством URI, и добавлять в него элементы с требуемой семантикой. Такой подход значительно упрощает взаимодействие между ИСИР РАН и другими системами.

Заключение.

В ИСИР разработана технология обмена данными с внешними системами, основанная на международных стандартах представления метаданных, включающая в себя общую модель данных и средства импорта/экспорта данных.

---

[1] Meta Data Europe 99: Implementing, Managing and Integration Meta Data, London UK, March 1999. Technology Transfer Institute. http://www.ttiuk.co.uk

[2] Resource Description Framework (RDF) Model and Syntax", W3C Recommendation, 1999. http://www.w3.org/TR/REC-rdf-syntax

---

Бездушный Анатолий Николаевич, кандидат физико-математических наук, с.н.с. вычислительного центра РАН.

Сфера деятельности:

- ☐ Система,
- ☐ программирование,

- ☐ параллелизм,
- ☐ сети,
- ☐ информационно-поисковые технологии.

Тел. (095)1355280, E-mail: bezdushn@ccas.ru

Лопатенко Андрей Сергеевич, м. н. с. центра научных телекоммуникаций и информационных технологий РАН.

Сфера деятельности:

- ☐ программирование,
- ☐ базы данных,
- ☐ интероперабельность,
- ☐ метаданные, цифровые библиотеки.

Тел. (095)9383938, E-mail: andreyl@ccas.ru

Меденников Антон Михайлович, аспирант вычислительного центра РАН.

Сфера деятельности:

- ☐ программирование,
- ☐ базы данных,
- ☐ цифровые библиотеки.

Тел. (095)5721234, (095)1355280, E-mail: meden@ccas.ru

Серебряков Владимир Алексеевич, с.н.с., зав.отделом вычислительного центра РАН.

Сфера деятельности:

- ☐ система,
- ☐ программирование,
- ☐ параллелизм,

- сети,
- информационно-поисковые технологии.

Тел. (095)1355471, (095)3065620, E-mail: serebr@ccas.ru

Филиппова Анна Александровна, сотрудник вычислительного центра РАН.