Построение информационной системы научного сообщества на основе интеграции разнородных коллекций ресурсов♣

© А.М.Федотов, В.Б.Барахнин, А.Е.Гуськов, Ю.В.Леонова

Институт вычислительных технологий СО РАН
fedotov@ict.nsc.ru, bar@ict.nsc.ru, guskov@ict.nsc.ru, juli@ict.nsc.ru

Аннотация

В работе описана технология создания информационной системы «База данных организаций и сотрудников СО РАН». В основе представленной технологии лежит принцип децентрализованного хранения информации при наличии единого каталога ресурсов. Такой подход обеспечивает интероперабельность, т.е. возможность интеграции разнородных ресурсов как внутри системы, так и с внешними системами, а также позволяет оперативно, с использованием средств автоматизации, осуществлять актуализацию информации.

1 Введение

В настоящее время научные сообщества наиболее развитых стран и регионов мира обладают достаточно мощными информационными системами (ИС). Так, в Европе функционирует интегрированная система ERGO [1]. Среди американских разработок своими масштабами выделяется информационная система Библиотеки Конгресса США [2]. К числу наиболее крупных и востребованных научным сообществом отечественных информационных систем относятся ЕНИП РАН [3], ИС «База данных организаций и сотрудников СО РАН» [4], Информика [5], ИС-Россия [6], eLIBRARY [7], Соционет [8]. Эти системы в той или иной степени удовлетворяют потребностям исследователей в информации, однако каждая из них страдает определенными недостатками, основными из которых являются недостаточно своевременная актуализация информации (этот недостаток не относится к библиотечным системам) и ограниченность возможностей обеспечения интеграции ресурсов как внутри каждой из систем, так и с внешними системами (иными словами, низкая интероперабельность).

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

Актуализация информации является слабым местом практически всех информационных систем некоммерческой направленности (за исключением, разумеется, систем, поддерживаемых государственными органами), предназначенных для функционирования в течение неопределенно долгого времени. Причина этого очевидна – недостаток средств, прежде всего, для оплаты труда лиц, которые должны отслеживать изменения информации.

Низкая интероперабельность – также весьма существенный недостаток, негативно отражающийся на важнейшей функции информационных систем: организации поиска документов по их атрибутам. Различают два уровня интероперабельности: семантический и технический [9], причем в последнем выделяют иногда синтаксический уровень [10]. Семантическая интероперабельность, заключающаяся в использовании согласованных стандартов метаданных (обзор которых применительно к научным информационным системам приведен в [10]), как правило, соблюдается. Проблемы возникают на уровне технической интероперабельности, точнее, согласования моделей данных и форматов их представления (что относится к синтаксической интероперабельности).

В настоящей работе мы опишем технологию создания ИС «База данных организаций и сотрудников СО РАН» [4], позволяющую в той или иной степени устранить перечисленные выше недостатки.

2 Особенности построения информационной системы СО РАН

Научные центры СО РАН, расположенные на территории трех федеральных округов, включают около 120 организаций. При этом каждая организация, являясь самостоятельным субъектом научной деятельности, традиционно обладает широкой самостоятельностью в выборе форм научно-организационной работы, включая информационное обеспечение. Это делает невозможным жесткую стандартизацию частных

аспектов корпоративных решений в области информационных технологий. Поэтому для построения информационной системы СО РАН выбран путь интеграции информационных систем институтов в распределенную систему с единым каталогом ресурсов (точнее, набором каталогов, каждый из которых включает объединяющих однородные ресурсы той или иной тематической направленности).

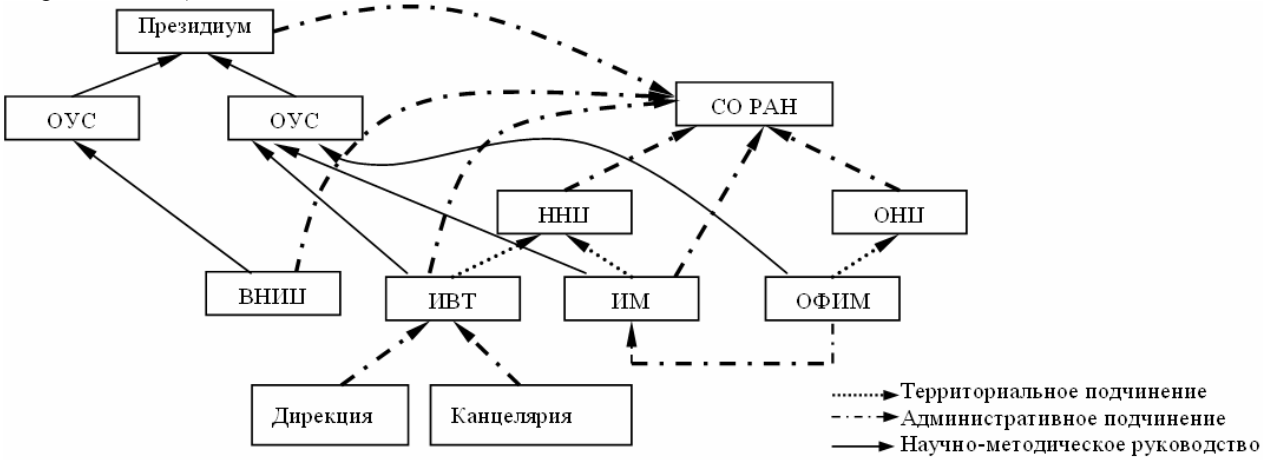

Рис. 1. Пример навигационной структуры «подчинения» в СО РАН

Наличие единого каталога имеет большое значение для решения проблем актуализации информации и обеспечения интероперабельности. Действительно, наиболее эффективным средством обеспечения процесса актуализации информации является его максимальная автоматизация. Она обеспечивается посредством регулярной автоматической проверки актуальности зарегистрированных в каталоге документов и пополнения его информацией о новых документах. Некоторые технические аспекты решения этой задачи рассмотрены в [11]. Одним из основных требований к информационной системе, процесс пополнения которой автоматизирован, является наличие в системе каталога. В противном случае, при отсутствии явно созданного множества унифицированных структурированных описаний документов, автоматическое добавление в систему новых документов становится, очевидно, крайне проблематичным. С другой стороны, каталог является наиболее естественной формой унификации представления данных, обеспечивающей интероперабельность ресурсов, является каталог.

Основными компонентами ИС СО РАН являются каталоги «Организации» и «Персоны».

Каталог «Организации» содержит:

- основные данные об организации – полное и сокращенное названия, почтовые и электронные адреса, телефоны, список головных организаций, фотографию, время существования организации (дата открытия и дата закрытия (переименования) организации);
- дополнительные сведения – описание деятельности, историческую справку, ключевые слова, список информационных ресурсов, описание инновационной деятельности.

Между организациями присутствуют иерархические связи подчинения следующих типов.

- территориальное подчинение, т.е. институты, входящие в региональный научный центр, подчинены Президиуму научного центра;
- научно-методическое подчинение, т.е. организации научно-методически подчинены Объединенному ученому совету по направлениям наук либо Научно-координационному совету. Объединенные ученые советы и Научно-координационные советы научно-методически подчинены Президиуму СО РАН;
- административное (юридическое) подчинение, т.е. региональные институты подчинены ассоциативному или объединенному институту, филиалы или подразделения институтов подчинены институтам СО РАН.

Иерархические связи используются при построении навигатора (рис. 1) по иерархической структуре подчинения организаций.

Важная особенность каталога «Персоны» состоит в том, что персоны могут работать в нескольких организациях, поэтому информация о персоне разделяется на две части: личную (связанную с самой персоной) и служебную (связанную с должностью персоны).

Личная информация содержит:

- Общие сведения – анкетные данные персоны: имя, фамилия, отчество, пол, место рождения, время жизни персоны (дата рождения и дата смерти), ученое звание и ученая степень, фотография
- Контактные данные: полный почтовый адрес и телефон (домашние)

- Дополнительную информацию: область интересов и деятельности, образование, специальность, награды.

Служебная информация содержит:

- Связь персоны с организацией, должность
- Рабочий телефон, факс, адрес электронной почты, www-страница
- Время работы персоны в организации на данной должности (дата принятия на должность и дата освобождения должности).

Если персона занимает несколько должностей, то для каждой должности создается свой экземпляр служебной информации.

Ввиду того, что общие сведения об организации, ее руководстве и ведущих сотрудниках являются своеобразной визитной карточкой организации на сайте СО РАН, для ведения названных каталогов используется принцип децентрализованной актуализации, означающий, что вводом и обновлением информации занимается субъект научной деятельности, к которому эта информация относится. Актуализация информации в информационной системе выполняется непосредственно людьми, собирающими эту информацию на местах, например администраторами институтов. Администратор организации обладает всеми правами для работы с информацией, относящейся к данной организации, и может изменять информацию об организации, добавлять, изменять и удалять информацию о сотрудниках организации и их публикациях.

Заметим, что объекты реального мира – организации и персоны – являются динамическими, т.е. информация об этих объектах, предоставляемая ИС, изменяется во времени. Поскольку устаревшая информация может представлять определенный интерес для истории, в ИС осуществляется учет временных изменений. Объекту (персоне или организации) сопоставляется некоторое подмножество версий документа, соответствующих разным отметкам времени изменения объекта. Например, при изменении должности или смене фамилии порождается новая версия документа. Последняя версия документа является текущей версией объекта. Все версии документов, отображающих объект, уникальны, т.е. одновременно не существует двух разных версий документов, имеющих одну и ту же отметку времени. ИС обеспечивает доступ к информации о персонах и организациях на любой момент времени.

В настоящее время осуществляется переход от ручного наполнения каталога «Персоны» (что гарантировало занесение информации лишь о руководстве организаций) к репликации информации из каталога LDAP. Такой подход должен обеспечить занесение в каталог сведений обо всех научных сотрудниках организации. Подробнее об интеграции каталога LDAP и ИС «База данных организаций и сотрудников СО РАН» рассказано в следующем разделе.

К основным каталогам СО РАН – «Организации» и «Персоны» – присоединяются дополнительные каталоги, например «Публикации», «Инновационные предложения» и др.

Коллекция «Публикации» содержит библиографическое описание публикаций сотрудников организаций СО РАН. Для их описания используется минимально допустимый набор полей, определенных библиографическими стандартами. В информационной системе выделяется 17 типов публикаций:

1. Монографии
2. Учебные пособия с грифом УМО
3. Учебно-методическая литература
4. Центральная печать
5. Зарубежная печать
6. Труды международных конференций
7. Труды всероссийских и региональных конференций
8. Авторефераты диссертаций
9. Препринты
10. Публикации в российских изданиях
11. Тезисы конференций
12. Электронные публикации
13. Патенты или Свидетельства о регистрации программ для ЭВМ
14. Отчеты НИР
15. Депонированные издания
16. Научные издания
17. Прочие издания

Для публикации устанавливаются связи “публикация – автор публикации” и “публикация – организация”.

Коллекция «Инновационные предложения» ведется в соответствии с международными стандартами, принятыми, в частности, в Российской сети трансфера технологий. Отличительной особенностью ведения этой коллекции является необходимость ее регулярной актуализации, в том числе в плане удаления (архивации) тех инновационных предложений, которые по каким-либо причинам больше не предлагаются для коммерциализации. Для элементов этой коллекции устанавливается связь “инновационное предложение – организация”.

3 Механизмы интеграции ИС «База данных организаций и сотрудников СО РАН» с другими каталогами

3.1 Интеграция: вариативность реализации

Как уже отмечалось ранее, на текущий момент информация об одной и той же персоне может быть найдена в большом количестве распределенных хранилищ данных. Более того, такие хранилища по своей природе могут быть разнородными: одни реализованы на основе реляционных СУБД, другие – на основе постреляционых, третьи функционируют в рамках LDAP-серверов. В тоже время, персона может выступать в разных ролях, и в

разных системах будут представлены различные аспекты ее деятельности. В частности, в научно-исследовательской среде можно выделить следующие характерные сведения:

- персональная информация;
- кадровая информация;
- перечень публикаций, участие в конференциях;
- участие в проектах;
- участие в диссертационных советах, ученых советах;
- образовательная деятельность.

Ситуация усугубляется фактическим отсутствием общепринятых соглашений об используемых схемах для представления персональных данных. В частности, можно привести следующие строго специфицированные схемы обмена персональной информацией:

- Формат **vCard** (Internet Mail Consortium) применяется для автоматизации процесса обмена контактной информацией, как альтернативы обычным визитным карточкам [12]. Она может использоваться в различных интернет-приложениях, включая электронную почту, веб-браузеры, IP-телефонию, видео конференции, факсы и др. С одной стороны, этот формат имеет достаточно простую структуру, что минимизирует время его внедрения в разнообразных приложениях. С другой стороны, в спецификации формата зафиксирован ограниченный набор полей; он не подлежит дальнейшему расширению и это сильно сужает сферу его применения.
- **LDAP** (Lightweight Directory Access Protocol) определяет протокол взаимодействия клиента и сервера для извлечения данных. Данный протокол активно используется для хранения данных о пользователях [13, 14] и организациях [15]. Стоит отметить, что LDAP хорошо зарекомендовал себя в качестве информационно-справочной системы с небольшим объемом редко обновляемой информации. В остальных случаях он заметно уступает классическим СУБД. Кроме того, реализация взаимодействия с сервером LDAP на основе XML-документов сложной спецификацией соответствующей схемы данных.
- Набор стандартов **CIQ OASIS** [16] описывает детализированную XML-схему для представления и обмена сведениями о персонах и организациях. Основной целью разработки этой схемы является создание обменного формата для CRM-систем. Следует отметить, что, несмотря на большой объем и гибкость схемы, получаемые документы достаточно наглядны и читабельны. Однако на сегодняшний день перечисленные стандарты распространены в гораздо меньшей степени.

Несомненно, для различных классов задач оптимальным может быть применение любого из перечисленных стандартов, а также стандартов, не упоминающихся в данной работе. Именно этим обуславливается необходимость разработки и применения технологий интеграции разнородных каталогов.

3.2 Интеграция: цели и задачи

Прежде всего, необходимо декларировать *цели* интеграции – варианты использования, описывающие практически полезное решение задач научно-организационного характера:

1. Аккумуляция данных из различных источников о каком-либо объекте (чаще персоне или организации) в одной точке виртуального информационного пространства. Причем аккумуляция может быть как ссылочной (как в портальных технологиях, где в одном месте собраны ссылки на разнородные источники), так и содержательной (когда в одном месте логически собраны все данные о заданном объекте, как копии сведений из источников).
2. Установление соответствия между ресурсами различных источников. Например, по каждому сотруднику из кадровой БД организации требуется сопоставить список его публикаций или получить сведения об образовательной деятельности.
3. Осуществление обмена предметными данными между различными системами. В частности, существует потребность в автоматизированном сервисе подачи заявки на участие в конференции, в котором персональные сведения вводятся не вручную, а запрашиваются из кадровой БД организации, LDAP-каталога или единого справочника сотрудников Сибирского отделения РАН.

Важный фактор для разработки и применения интеграционных технологий состоит в том, что большинство электронных систем, предоставляющих различные информационные услуги и сервисы, на текущий момент реализованы и успешно функционируют на протяжении длительного времени. В связи с этим, модификация схем данных, программного кода, алгоритмов и регламентов их работы в сколько-нибудь значительных объемах крайне нежелательна. Поэтому применение интеграционных технологий должно осуществляться в соответствии с принципом «неразрушающего расширения», когда модификация системы должна иметь целью

реализацию новой функциональности без изменения в худшую сторону существующей.

В целом, основываясь на практическом опыте, можно выделить следующие задачи, выполнение которых необходимо для достижения перечисленных целей:

1. Разнородность протоколов доступа к интегрируемым ресурсам (например, LDAP, Z39.50, CORBA, ODBC, Web Services).
2. Необходимость сопоставления несогласованных форматов, схем данных и, в конечном итоге, различных онтологий смежных предметных областей.
3. Определение «релевантных» ресурсов – ресурсов, содержащих сведения об одном и том же объекте реального мира.
4. Необходимость сохранения источников данных – возможность определить, откуда была получена информация и какова степень ее достоверности.
5. Разработка пользовательского интерфейса, предоставляющего доступ ко всем данным, включая ссылки на внешние источники и связанные ресурсы.
6. Необходимость оперировать актуальной информацией вследствие естественной изменчивости предметных данных. Для этого интеграционные технологии должны обладать адекватными механизмами мониторинга и актуализации устаревших данных, функционирующих в автоматическом режиме (в отдельных случаях может возникать необходимость в хранении истории изменений и предоставления информации в задаваемых временных срезах).

Следует отметить, что академическая специфика позволяет сделать существенные допущения в требованиях к функционалу системы. Так, система не должна быть полноценным «аккумулятором» всей информации – такое хранилище на сегодняшний день не является реально востребованным, а дублирование информации в еще одной системе приведет, скорее, к новым противоречиям, чем к извлечению новых знаний.

Гораздо более эффективным в части отношения «полезность/трудозатраты» является разработка системы, для которой объектами интеграционной функции были бы не данные, содержащиеся в ресурсах, а сами ресурсы. При этом частичная обработка содержания ресурсов имеет место исключительно для извлечения связей с другими ресурсами, а не для целей сохранения в собственной БД. В результате может быть создана система, концептуально напоминающая современные web-порталы, – она не содержит никаких сведений о внешних ресурсах, за исключением ссылок на них. И именно эта целевая функция является основной, поскольку получение собственно информации пользователь осуществляет во внешнем источнике, который и был для этого предназначен.

3.3 Интеграция: концепция системы

Для реализации ссылочной аккумуляции ресурсов и установления соответствия между ними целесообразно реализовать интегрирующую систему, имеющую интерфейсы взаимодействия с источниками. В соответствии с перечисленными требованиями данную систему можно разделить на следующие логические компоненты [17]:

1. **База данных** для хранения предметных сведений. Для решения задач интеграции БД должна содержать:
   - Минимальный набор полей для каждого предметного объекта (персоны, организации, публикации, проекты, мероприятия).
   - Метаданные ресурсов, из которых извлекаются сведения о предметных объектах, для целей каталогизации.
   - Связи между одним и более предметными объектами – любые два предметных объекта могут быть связаны друг с другом типизированным отношением (например, персона *является сотрудником* организации, персона *является автором* публикации).
   - Параметры взаимодействия с источниками – сетевые адреса, протокол доступа, параметры авторизации.
   - Регламенты и правила обработки ресурсов для каждого источника – периодичность обновления данных, интерфейсный драйвер для доступа к ресурсу, последовательность запуска компонентов для их обработки.

   БД должна быть легко расширяема новыми типами объектов и отношений, предоставлять возможность автоматически настраиваться на работу с новыми источниками, производить автоматизированную обработку данных. В качестве платформы для реализации БД авторы предлагают использовать либо реляционные СУБД как надежное и функциональное средство управления данными, либо реализацию RDF-хранилища. Преимущество RDF состоит в том, что этот язык является формализованным инструмента для представления онтологических описаний, содержащего прозрачные правила идентификации ресурсов, возможность формулирования фактов наличия

именованных связей между ресурсами и возможность вывода новых фактов из уже существующих.

2. **Модуль взаимодействия с источниками** должен поддерживать интерфейсы для обмена предметными сведениями с внешними системами. Инициатором обмена может выступать как интегрирующая система (когда обновление сведений производится по заданному регламенту), так внешний источник (по факту обновления собственных данных). Протоколы взаимодействия должны соответствовать открытым стандартам. Основными компонентами модуля являются *интерфейсные драйверы*. Каждый драйвер обслуживает свой интерфейс (причем каждый интерфейс используется для доступа ко всем источникам с аналогичным протоколом обмена) и преобразует полученные данные из внешнего источника в собственное внутреннее представление, к которому применяются последующие процедуры обработки и каталогизации. Так, для существующих информационных систем, ориентированных на web, авторы предлагают использовать технологию web-сервисов [11]; в этом случае интерфейсный драйвер будет представлять собой клиента, взаимодействующего с web-сервисом. Для систем, реализованных на других платформах (LDAP, Z39.50) потребуется разработка соответствующих интерфейсных драйверов.
3. **Модуль обработки ресурсов** осуществляет анализ полученных данных и размещение их в БД. Также в его функции входит установление факта соответствия нескольких ресурсов одному предметному объекту, либо выявления отношения между двумя предметными объектами. Таким образом, можно выделить следующие основные компоненты данного модуля:
   - Компоненты типа *DataLoader*, которые извлекают базовый набор сведений о предметном объекте из внутреннего представления ресурса (сформированного интерфейсным драйвером) и сохраняют их в БД.
   - Компоненты типа *DataMiner*, которые извлекают из внутреннего представления ресурса связи между различными предметными объектами.

   В отличие от компонентов DataLoader, компоненты DataMiner анализируют имеющиеся предметные объекты, выявляют возможные связи между объектами и определяют степень их достоверности. В случае если достоверность связи приемлема, такая связь заносится в БД с указанием источника связи. Характерным примером выявления связи является сопоставление авторов публикации, полученной из каталога Z39.50 с уже имеющимся списком сотрудников института – когда прямой связи между публикацией и ее автором нет; такую связь можно восстановить путем сопоставления фамилии и инициалов персоналий.

   Очевидно, для каждого типа объекта необходима реализация собственного алгоритма обнаружения соответствующих ресурсов. Аналогично, для каждого типа отношений требуется собственный критерий установления отношения.
4. **Диспетчерский модуль** должен выполнять функции мониторинга работы системы, управлять выполнением заданных действий по расписаниям (например, автоматически производить обновление ресурсов из внешних каталогов), не допускать возникновения конфликтов при одновременном доступе к сервисам интегрирующей системы.

   Типичная процедура управления обработки ресурсов внешнего источника состоит в последовательном прохождении следующих шагов:
   1. Определение конфигурации источника;
   2. Запуск интерфейсного драйвера и получение ресурсов;
   3. Запуск компонентов DataLoader;
   4. Запуск компонентов DataMiner;
   5. Подготовка отчета о выполнении.

5. **Модуль пользовательских интерфейсов** должен предоставлять пользователям доступ к сведениям из БД. Дополнительно необходима реализация администраторского интерфейса для управления работой системы. Отметим, что современная технологическая база диктует необходимость создания пользовательских интерфейсов на базе web-технологий.

4. Заключение

Разработанная технология способна обеспечить децентрализованное хранение ресурсов, которые регистрируются в едином каталоге, служащим ядром распределенной системы. Автоматизация процесса актуализации информации, хранящейся в каталоге, резко снижает текущие затраты на поддержку системы, являясь залогом перспективности данной технологии.

Литература

[1] European Research Gateways Online [http://www.cordis.lu/ergo]

[2] Библиотека Конгресса США [http://www.loc.gov/]
[3] Единое Научное Информационое Пространство РАН [http://www.ras.ru/]
[4] База данных организаций и сотрудников СО РАН [http://www.sbras.ru/sbras/db/]
[5] Государственный НИИ информационных технологий и телекоммуникаций "Информика" [http://www.informika.ru]
[6] Университетская информационная система РОССИЯ [http://www.cir.ru/index.jsp]
[7] Научная электронная библиотека eLIBRARY.RU [http://elibrary.ru/defaultx.asp]
[8] Соционет [http://socionet.ru/]
[9] *Фейгин Д.* Концепция SOA // Открытые системы. – 2004. – N 6. [http://www.osp.ru/os/2004/06/184447/_p1.html]
[10] *Бездушный А.Н., Кулагин М.В., Серебряков А.А., Бездушный А.А., Нестеренко А.К., Сысоев Т.М.* Предложения по наборам метаданных для научных информационных ресурсов // Вычислительные технологии. – 2005. – Т. 10. – Специальный выпуск. – С. 29-48.
[11] *Барахнин В.Б., Ведерников В.В.* Автоматизированная каталогизация электронных журнальных публикаций // Труды международной конференции "Вычислительные и информационные технологии в науке, технике и образовании". Казахстан, Павлодар, 20-22 сентября 2006 г. – Т. I. – С. 209-214.
[12] vCard MIME Directory Profile [http://www.ietf.org/rfc/rfc2426.txt]
[13] A Summary of the X.500(96) User Schema for use with LDAPv3. [http://www.ietf.org/rfc/rfc2256.txt]
[14] Definition of the inetOrgPerson LDAP Object Class. [http://www.ietf.org/rfc/rfc2798.txt]
[15] The COSINE and Internet X.500 Schema [http://www.ietf.org/rfc/rfc1274.txt]
[16] OASIS CIQ v2.0 committee specifications / standards page // [http://www.oasis-open.org/committees/ciq/ciq.html]
[17] Гуськов А.Е. Модель виртуальной среды для обмена результатами научных исследований // Труды международной конференции «Вычислительные и информационные технологии в науке, технике и образовании». I том. – Павлодар: ТОО НПФ «ЭКО», 2006. С. 372-380.

Developing information system for scientific society based on the integration of heterogeneous heterogenius heterogenius collections of resources

A.M. Fedotov, V.B. Barakhnin, A.E.Guskov, J.V. Leonova

In this paper the technology for information system “Organizations and employees of SB RAS directory” creation is described. This technology based on the principle of decentralized information storage with an integrated resource catalogue. Such approach provides the possibility of automatic actualization of information and system’s interoperability – the possibility of heterogeneous resource integration within the system and between external systems.

♣ Работа выполнена при частичной поддержке РФФИ: проекты 06-07-89060, 06-07-89038, 07-07-00271, президентской программы ”Ведущие научные школы РФ” (грант № НШ-9886.2006.9) и интеграционных проектов СО РАН.