Платформа интеграции электронных архивов

© Марчук А.Г., Марчук П.А.

Институт систем информатики им. А.П.Ершова СО РАН
mag@iis.nsk.su, peter@iis.nsk.su

Аннотация

В статье рассмотрена задача интеграции информационных систем фактографической направленности. Интеграция рассматривается и как информационное объединение ресурсов и как переход отдельных информационных систем на унифицированное решение при сохранении функциональности и интерфейсов каждой в отдельности. Важным элементом предложенного подхода является ориентация на распределенную систему с сохранением контроля владельца информационного ресурса над «своими» данными.

Обсуждаемый подход реализован в системе «Электронный фотоархив СО РАН».

Введение

Институт систем информатики СО РАН, уже в течение более 10 лет ведет работу по созданию разнообразных информационных систем архивной направленности. В частности, созданы: электронный архив академика А.П. Ершова [1], исторический портал ММФ НГУ [2], электронная версия хроники Сибирского отделения [3] и ряд других. В настоящее время, институту поручено создание фотоархива Сибирского отделен-ия РАН. Указанные системы создавались как автономные технические и информационные решения, содержащие традиционные составляющие: базу данных, логику работы с архивом, пользовательские интерфейсы. Причем, как правило, имеется пользовательский или публичный, интерфейс и интерфейс информационной поддержки и редактирования. Каждая из созданных информационных систем продолжает жить и поддерживаться специалистами института и, в некоторых случаях, заказчика.

Проблема заключается в том, что эксплуатация и обеспечение жизненного цикла столь широкой гаммы электронных архивных систем ложится тяжелым бременем на институтские службы и стоимость такой эксплуатации практически линейно зависит от количества архивов. Большие проблемы ожидаются при изменении технологической базы, что ведет к необходимости модернизации и переходу к новым техническим решениям. Такая ситуация, вот уже около 30 лет характерна для наиболее ранней из информационных систем, созданной и эксплуатируемой в ИСИ. Речь идет о библиотечной системе института, обслуживающей мемориальную библиотеку А.П. Ершова [4]. Первый вариант системы был создан еще на БЭСМ-6, в качестве носителя информации сначала использовались перфокарты. В дальнейшем, система подвергалась неоднократным модернизациям, программы, с тех пор не раз были переписаны, интерфейсы перепроектированы, преемственность осуществлялась по данным: библиотечные карточки и другие элементы данных, конвертировались в новое представление и именно они, составляют основную ценность информационной системы.

Современные методологические и технологические решения группы Semantic Web [5] дают основу для интеграции информационных систем, что позволит расширять количество и тематическую направленность таких систем без существенного увеличения штата и стоимости эксплуатации, позволит на регулярной основе производить техническое обслуживание и модернизацию. Важным и концептуальным является еще одно преимущество, появляющееся при интеграции данных информационных систем: расширение информационного поля для каждой из них. Речь идет об обобществлении информации о традиционных сущностях, фиксируемых практически в любой информационной системе, особенно архивной направленности: людях, организациях, датах, событиях.

1. Методология построения электронного архива

Основу архива, как правило, составляет однородная коллекция предметов, например, коллекция документов. Соответствующий электронный образ архива начинается с коллекции (электронных) карточек, описывающих нужные свойства конкретных экземпляров. Эта часть структуризации данных – достаточно обычна и соответствует традиционным метаинформационным построениям. Соответственно, применяемая структуризация, имеет общие черты с многочисленными библиографическими системами, системами описания информационных ресурсов, а набор используемых полей, и у нас тоже, соответствует построениям типа Dublin Core [6].

Было замечено, что линейная коллекция (электронных) карточек не достаточна для скольнибудь объемного архива, содержащего десятки тысяч и более описаний. Попытки внести иерархию

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

в данное однородное множество, например, группированием по темам, позволяет достичь существенных результатов в академической (профессиональной) работе с архивом, но для широкого круга пользователей не знакомых с используемой иерархией тем, является слабым подспорьем для решения их задач поиска информации.

Начиная с работ над электронным архивом академика А.П. Ершова, мы придерживаемся следующего подхода: одновременно с описанием однородной коллекции элементов архива, создается база данных традиционных сущностей (люди, организации, даты, города, события), хорошо знакомых непрофессиональному пользователю и элементы архива «привязываются» к базе данных. Привязка осуществляется по таким естественным отношениям, как авторство, временные и географические аспекты, связанные с элементом коллекции, принадлежность к организационным системам, информационное отражение в документе сущностей реального мира. Создается как бы модель внешнего мира, и экспонаты погружаются в эту модель. В итоге, пользователь отталкивается от модели, которую он знает гораздо лучше, а на ее элементах (людях, организациях), находит привязанные к ним элементы коллекции. Выполненный в таком стиле электронный архив, позволяет выполнять как профессиональную работу с его содержимым, так и производить быстрое ознакомление через нажатие гиперссылок, ориентируясь, главным образом, на знакомые пользователю понятия и образы. Перемещения, как правило, выполняются в последовательности: персона (или организация) – экспонат из множества экспонатов, связанных с персоной – другая персона из множества персон (организаций), связанных с экспонатом. Имеющаяся в базе данных информация позволяет сузить выборку элементов данных за счет указания пространственных, временных и тематических ограничений.

2. Методология интеграции данных разных архивов

Как было показано в предыдущем разделе, при примененном подходе, в архивах имеется существенная часть общей по смыслу информации. Это информационное поле и есть основа интеграции по данным разных по тематике архивов. Действительно, глядя на архивную информацию глазами пользователя, мы отталкиваемся от традиционных сущностей, таких как люди, организационные структуры, географические образования и обнаруживаем связи этих сущностей с элементами описаний различных архивов. С познавательной точки зрения, это не обедняет, а обогащает возможности пользователей получать информацию и более комплексно раскрывать различные темы. Например, интересуясь конкретным университетом, пользователь получит из одного архива информацию о документах, связанных с деятельностью университета, из другого архива – фотографии, иллюстрирующие наиболее интересные аспекты университетской жизни, из университетской базы данных ему представлена структура организации, известные ученые и преподаватели, знаменитые ученики. Дополнительную информацию о найденных персонах, пользователь сможет получить из общей базы данных и других архивных материалов.

Создавая общее информационное поле, мы можем использовать не только внутренние имеющиеся источники, но и внешние (например, какие-либо единые государственные реестры, другие электронные архивы). И если внутренние данные мы можем (хотя это необязательно) физически разместить в одном месте, предоставляя разные интерфейсы для работы с ними, то в другом случае у нас появляется концепция распределенного хранения данных, когда при объединении информационного поля у нас не происходит физического перемещения информации.

На этом этапе появляется ряд методологических проблем, как-то: разные реализации деления на сущности и отношения, разная идентификация объектов, наличие дублирования информации, согласование используемых онтологий.

Также в разных базах данных существует свой информационный контекст. Например, предположим, что имеется база данных персон в какой-то конкретной организации. Эти персоны являются сотрудниками этой организации. Но это не единственная информация, которую можно оттуда выжать. Это и расположение персон – конкретный город, и то, что персоны еще, скорее всего, живы и косвенно доступна некоторая коммуникационная информация (напр. телефонный код города).

Кроме того, нам требуется эволюционное согласование баз данных и отдельных ее частей. Со временем схема данных или онтология может поменяться в любой из частей распределенной системы. Поэтому нам необходимы механизмы, во-первых, указания версии используемой онтологии, а во-вторых, возможность трансформировать данные из старых онтологий в новые. Таким образом, речь идет не только о статическом представлении об онтологиях, но и о динамике онтологий.

3. Интеграция программ и интерфейсов

В целом проблему интеграции программного обеспечения объединяемых информационных систем решить на данный момент, не представлялось возможным. Однако удалось частично решить конкретные моменты. В частности, могут существовать единые средства для операторов данных (или информационных администраторов) для поиска информации или ее редактирования.

Программы могут быть реализованы на разных платформах. Для решения чего, соответственно, требуется в некоторых случаях создавать export/import модули. Либо интерфейсы сетевого взаимодействия.

Возможна система настраиваемых интерфейсов, причем настройка осуществляется по принципу ограничения взаимодействия оператора с данными и их элементами. В итоге, в целом ряде частных случаев удается с помощью базовых средств

породить специализированный интерфейс под конкретную потребность.

Кроме интерфейсов, нам требуются объединенные программы системного администрирования. Нам необходимо разграничить полномочия, предоставить возможность конфигурирования, а также сбора статистики.

Используя концепцию из Semantic Web «любой о любом может сказать любое», нам необходимо на системном уровне обеспечить неразрушаемость данных и защиты от злонамерения или некомпетентности. Одним из нами опробованных решений является решение структурного уровня, когда модификации кусков семантического графа могут располагаться в других файлах. Таким образом, имея защищенные файлы, в частности удаленные, можно производить их содержательное редактирование.

Существуют разные подходы реализации базовой RDF-машины. Типовыми являются три: 1. использовать хорошо зарекомендовавшие себя информационные СУБД; 2. использовать чье-то специализированное решение (Jena [7], Sesame [8] и др.); 3. создать свою систему управления базами данных.

В данном случае нами реализуется третий вариант, т.е. в проекте взят курс на создание простой системы работы с семантическими сетями RDF, оптимизированной на работу с сетью под спектр запросов, который характерен для архивной проблематики. Такое решение показало высокую эффективность при работе с данными, содержащими до миллиона фактов.

Также в плане интеграции, ситуация может осложняться тем, что во время этой интеграции, все еще происходит работа с данными, то есть данные либо изменяются, либо добавляются. В этом случае требуется использование версий интегрируемых данных.

4. Распределенное редактирование данных

В этом разделе речь идет прежде о мета объектах, которые содержатся в нашей информационной системе. Некоторые из них были проинтегрированы в нашу конкретную систему, некоторые могут находиться в удаленном положении при этом в нужном нам формате. Некоторые могут быть так же на другом удаленном сервере и к тому же в другом формате, но при этом у нас есть импорт-фильтр для их использования. В общем, исходя из первоначальных источников, мы получаем RDF-документы, в которых хранятся мета объекты. В совокупности эти RDF-документы дают нам единое информационное поле, с которым мы должны работать.

При работе с этими мета объектами в нашем информационном поле в режиме чтения, и если у нас нет проблем с дублированием идентификаторов, вся работа выстраивается достаточно простым способом. Например, мы должны найти объект, чтобы посмотреть его свойства, либо мы должны найти объекты, которые ссылаются на этот объект, чтобы отобразить их свойствами. Методология такого поиска объектов, части ассоциативного графа описана, например, в языке поиска SPARQL Query Language for RDF [9].

В этом разделе обозначена проблема, а так же предложен вариант решения этой проблемы в том случае, если нам требуется редактировать информацию в нашем информационном поле, при том, что у нас есть источники, которые находятся не в нашей собственности.

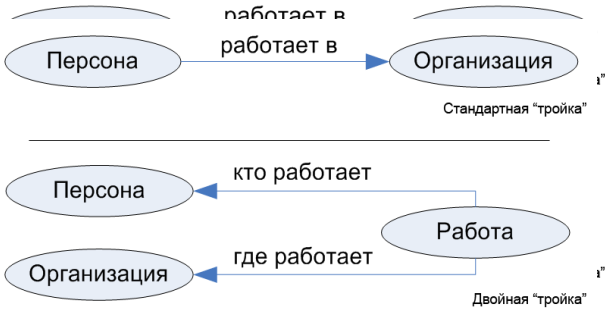
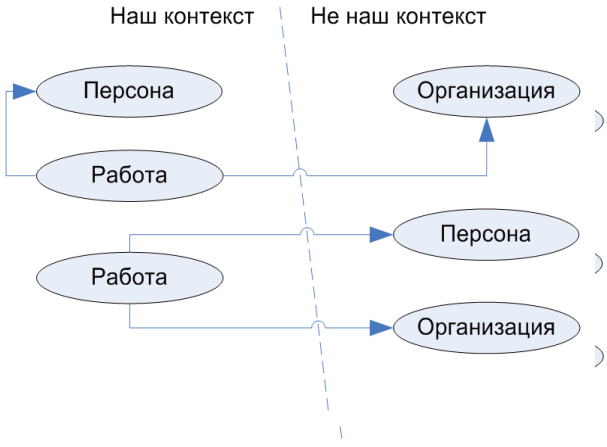


Рис. 1. Стандартная и двойная «тройка»

Сначала, обозначим некоторую принципиальную для данного раздела особенность нашей онтологии. В стандартной концепции графов RDF, ассоциативные связи между объектами обозначаются при помощи «троек» (Triples) RDF. В тройку входит субъект, предикат и объект (см. рисунок, первая часть). В нашей же онтологии в качестве базовой схемы установления ассоциативной связи между объектами мы берем следующую схему. Ассоциативную связь, которую мы устанавливаем, мы так же заводим в качестве мета объекта в базу данных. И эта связь ссылается на объекты, между которыми мы эту связь устанавливаем (см. рисунок, вторая часть). Фактически одна связь представлена в виде двух «троек».

Теперь определим контекст работы. Зададим простейшую модель, при которой в нашей единой

Рис. 2. Возможность построения связей с объектами в других контекстах



распределенной информационной системе существуют только 2 источника. Наш источник, где мы

можем редактировать данные, и чужой источник, данные которого мы не можем редактировать, но можем просматривать, а он наши данные ни просматривать, ни редактировать не может. При этом нашей областью просмотра является вся распределенная система, а контекстом редактирования только наш источник. Исходя из этой модели и нашей онтологии, мы можем добавлять в наш контекст новые мета объекты, редактировать их свойства и устанавливать связи между нашими мета объектами. Кроме того, мы можем устанавливать связи и между объектами разных источников (см. рисунок). При этом, имея областью просмотра все источники, мы получим полную информационную картину. Второй источник полной картины не получит, так как его область просмотра по нашему определению не включает нашего источника, но и противоречий наши добавленные объекты не вызовут.

Так как основу семантической сети представляют собой все-таки связи между объектами, нежели конкретные свойства каждого объекта, то, используя приведенную выше модель, мы достигаем заметного результата, однако полной универсальности данного решения пока недостаточно. Например, поскольку мы создаем много новых мета объектов, то из-за их количества свойства этих объектов из-за своей множественности так же начинают играть ощутимую роль. Мы можем захотеть изменить уже имеющееся свойство, например, неправильное название организации, либо добавить новое свойство, например, дату основания. Кроме того, мы можем посчитать, что какой-то мета объект не верен, например, человек никогда не работал в такой-то организации, а в базе данных есть соответствующее упоминание. Таким образом, мы должны удалить какой-то мета объект. Понятно, удалить его из своего контекста не составляет труда, однако это становится проблемой в случае с чужим контекстом.

В случае с главной централизованной базой данных, данная проблема решается на уровне полномочий. Пользователю дается право на редактирование определенной области данных, возможно после редактирования, это действие должно быть подтверждено модератором соответствующей области. В этом случае после изменений, мы даже можем сохранить предыдущую версию, которая была до изменений, чтобы возможно потом эти изменения отменить, если нужно. В принципе, мы можем перенести эту схему и на распределенную модель хранения данных. Тогда каждый источник может определить полномочия других пользователей информационной системы и контролировать свой «куст» данных. Недостатком этого подхода, например, является обязательное наличие серверной части, отслеживающей полномочия всех пользователей. Кроме этого, свойства этих объектов тогда становятся частью «куста» данных и мнением не того человека, кто это сказал, а того человека, где этот мета объект находится. В ряде случаев, обозначенные недостатки могут не являться проблемными местами информационной системы и подход может использоваться для распределенного редактирования данных.

Рассмотрим случай с отсутствием серверной части, определяющей какие объекты можно редактировать. Например, чужие «кусты» мы просто получаем при доступе через http-протокол, и там не предусмотрено редактирование.

В этом случае, нам все равно требуется внести изменения, однако нам надо сделать это исключительно в своем контексте редактирования. При этом, при визуализации, изменения должны корректно показываться, то есть, по крайней мере, в нашей части системы они должны показываться именно так, как мы хотим их увидеть.

Итак, чтобы внести изменения в свойства мета объекта, который находится в удаленном кусте, предлагается следующий вариант решения. Мы создаем новую сущность у себя в контексте редактирования. Наделяем ее при этом всеми теми же свойствами, что были в удаленном объекте. Теперь этот мета объект находится в нашем кусте, и мы можем произвести все необходимые нам изменения. Однако в семантической сети ссылаются все еще на старый объект. В своем контексте редактирования мы, в принципе, можем исправить все ссылки на старый объект ссылками на новый добавленный объект, однако другие ссылки на старый объект остаются в нередактируемых источниках. Поэтому мы добавляем дополнительное указание для просматривающей системы, что произошла смена идентификатора, и новый объект «заменил» старый. Тогда информационная система должна при получении старого идентификатора автоматически его заменять на новый.

При некотором изменении данной схемы мы аналогично можем удалять мета объекты из чужих «кустов», обозначая в нашем контексте, что этого мета объекта не существует (для нас).

5. Опыт выполненных разработок

В результате выполнения упомянутых в начале статьи проектов, в Институте систем информатики был сформирован подход, основа которого изложена в данной работе. Платформой интеграции предыдущих разработок и апробации новых решений явился выполняемый в настоящее время проект создания фотоархива СО РАН.

В нем реализованы принципы, изложенные далее.

Наиболее важным моментом для интеграции данных является наличие онтологии для описания информации. В проекте онтологии описываются при помощи OWL (Web Ontology Language). В частности, в [10] выделяются пять базовых классов сущностей: персоны, организационные системы, гео-системы, документы, коллекции. Те, в свою очередь, разделяются на подклассы. Установлен также набор сложных (атрибутированных) отношений между сущностями и необходимыми полями значений.

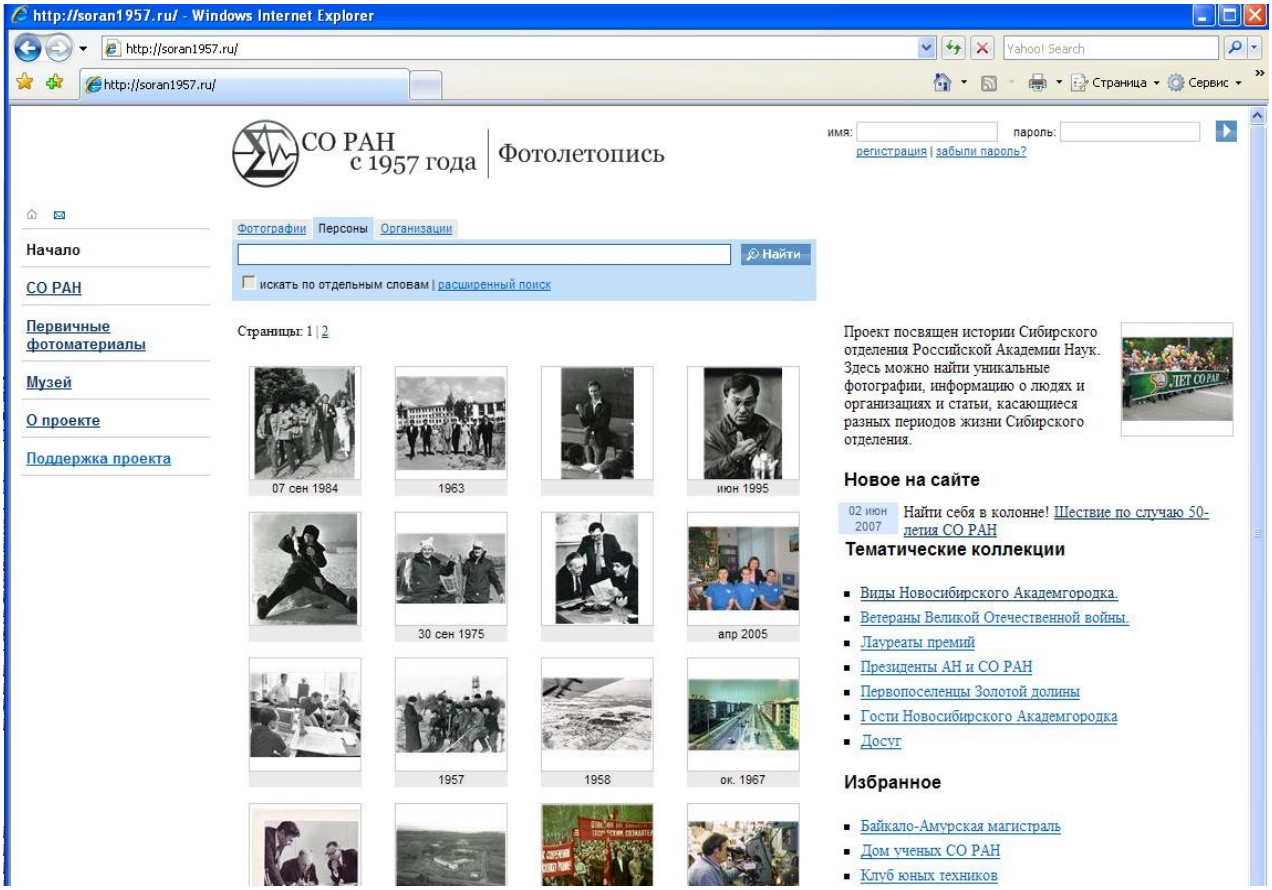
В проекте принято решение поддерживать

распределенное хранение информации, в разных файлах, на разных компьютерах. В качестве информационных источников используются RDF (XML)-документы, однако возможно использование также и реляционных таблиц, погруженных в реляционные СУБД и некоторых других хранилищ.

В программном обеспечении предусмотрена возможность использования разных информационных источников как в единой информационной сети – Интернете, так и в разных информационных сетях, при помощи подключения дополнительных информационных ресурсов, конкретные данные которых могут находиться в оффлайновом состоянии, обновляясь периодически.

Предоставлена возможность корректировки данных не только хранящихся в данной информационной системе, но и в других источниках. Однако изменение делается не непосредственно в данных первоисточника, что зачастую невозможно или нецелесообразно, а в области данных интегрирующей информационной системы.

Рис. 3. Фрагмент публичного интерфейса «Электронного фотоархива СО РАН»



5.1 Система хранения информационных документов

Документ отправляется в хранилище традиционным и понятным пользователям способом – при помощи помещения его в некоторую папку с фото, видео, аудио и текстовыми документами. После чего документу присваивается уникальный идентификатор, который в дальнейшем используется системой. Наличие директорий (папок) не означает, что документ будет храниться именно в этой статической директории, однако, так как эта директория может помогать пользователю ориентироваться в документах, то она остается, правда, в качестве коллекции (виртуальной директории, вообще говоря, одной из многих).

Одной из частей информационной системы является система преобразования (копирования) информационных документов в более удобный для повседневного использования вид. Например, в случае фотографий, это фотографии с меньшим разрешением, достаточным для большинства пользовательских целей – просмотр, выставление в Интернет, thumbnails и т.п. Так как для преобразования документов в более удобный вид необходимо заметное количество времени, то для этого разработана специальная программная система, которая в фоновом режиме следит за добавленными документами и преобразует их, где надо.

Кроме автоматического преобразования (копирования) часто существует потребность и в неавтоматическом преобразовании. Например, для фотографий это может быть убирание полей (crop), поворот/зеркальное отражение (rotating, mirror), подчистка фотографий от физических повреждений Возмо-

жен вариант, когда архив получил документ (тот же самый) более высокого качества. В этом случае реализованным общим подходом является помещение нового документа в хранилище с установлением ассоциаций старого документа на новый, либо в частном случае (если качество ни для каких целей не могло быть ухудшено), непосредственная замена документа.

Также требуется система резервного копирования. В данном случае это отслеживание нового полученного содержимого по данным последнего изменения. Хранение требуется осуществлять не только на разных винчестерах, но и на разных компьютерах, и, если возможно, в разных сетях. В случае обнаружения каких-то несоответствий, требуется их выявить (подсказать пользователю о них), и, по подтверждению, выполнить корректирующие действия.

5.2 Служебный интерфейс и утилиты

Для описания сущностей и установления ассоциаций между ними, создан интерфейс взаимодействия операторов с системой (Backend). Интерфейс автоматически настраивается на предметную область через заданную онтологию.

В рамках проекта Фотоархив СО РАН помимо более традиционных формальных подходов к интеграции, требовалось также иметь возможность быстро превращать неформально описанные данные в данные с нужной онтологией. Для этого было создано программное обеспечение для ускоренного ввода специфических данных.

Ряд действий пока не доступен, в силу их критичности, даже информационным специалистам редактирующим фотоархив, а выполняется администратором системы. Для таких действий создан интерфейс административных действий и ряд утилит. К таким, критическим, задачам, в первую очередь относится обнаружение и устранение дублирования информации. По нескольким признакам, система получает кандидатов на возможное дублирование. В качестве признаков могут браться наименования, начальные и конечные даты, похожие ассоциации с другими объектами и т.п. Далее администратору предоставляется возможность утвердить предположение системы об имеющихся дублях и инициировать автоматическую процедуру слияния данных.

Создается система статистики, которая следит за объемами и балансировкой информационных доменов, добавлением информации конкретными операторами, за наиболее используемой и наиболее интересной для конечных пользователей информацией (для возможной оптимизации в дальнейшем информационной системы).

Литература

1] Архив А.П. Ершова. http://ershov.iis.nsk.su/
2] Исторический портал ММФ НГУ. http://www.globalmmf.ru/
3] Хроники Сибирского отделения. http://chronicle.iis.nsk.su/
4] Мемориальная библиотека А.П. Ершова. http://www.iis.nsk.su/
5] Berners-Lee Tim, Hendler James, Lassila Ora, The Semantic Web. In Scientific American, volume 284(5), pages 34-43, 2001.
6] The Dublin Core Metadata Initiative. http://dublincore.org/
7] Jena – A Semantic Web Framework for Java. http://jena.sourceforge.net/
8] Sesame. http://openRDF.org/
9] SPARQL Query Language For RDF // http://www.w3.org/TR/rdf-sparql-query/
10] Марчук А.Г. Распределенные электронные архивы, библиотеки и базы данных. Препринт 122, Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, 25с., 2004.

Digital archives integration platform

Alexander G. Marchuk, Peter A. Marchuk

The article examine the problem of integration of factographic information systems. Integration means both union of information resources and migration of information systems to uniform solution, preserving functionality and interfaces of specific system.

Proposed approach was implemented in the project “Digital photo-archive of SB RAS”.