УДК 004

ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ В ВЫЧИСЛИТЕЛЬНОМ ЦЕНТРЕ РОССИЙСКОЙ АКАДЕМИИ НАУК – ОСНОВНЫЕ РАЗРАБОТКИ

В. А. Серебряков[1]

Вычислительный центр им. А.А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, г. Москва, 119333, ул. Вавилова, 40

[1] serebr@ultimeta.ru

Аннотация

Рассмотрены основные проекты, которые были реализованы в Вычислительном центре им. А.А. Дородницына Российской академии наук (ВЦ РАН) за последние 20 лет, т. е. с 1998 года. Одним из первых был реализован пилотный проект «Интегрированной системы информационных ресурсов (ИСИР) РАН». Успешное завершение этого проекта позволило развернуть работы по интеграции разнородных научных информационных ресурсов в общеакадемическую научную информационную систему. Важным этапом был проект создания Единого Научного Информационного Пространства (ЕНИП) РАН. Этот проект основывался на подсистеме «Научный институт РАН», созданной в ВЦ РАН и Центре научных телекоммуникаций (ЦНТК) РАН. Учитывая важность формирования цифровых библиотек, Российская академия наук приняла в 2006 году целевую научную программу «Создание ЦБ «Научное наследие России»», в соответствии с которой была реализована цифровая библиотека. Созданный портал «ГеоМета» – это стандартизированная и децентрализованная среда управления пространственной информацией, разработанная для доступа к базам геоданных, картографическим продуктам и связанным с ними метаданным из различных источников, облегчающая обмен пространственной информацией между организациями и ее совместное использование посредством интернета.

В настоящее время основное направление работ – цифровая персональная семантическая библиотека LibMena. Основная задача этой системы заключается в предоставлении пользователю унифицированного представления для

возможности автоматизированного извлечения интересующей его информации по определенной предметной области.

***Ключевые слова**:* предметная область, научная предметная область, научная информация, научные знания, обобщенное представление научной предметной области, таксономии, тезаурусы, глобальные онтологии, поисковые системы, организация научных знаний, цифровые библиотеки

ВВЕДЕНИЕ

Работы в направлении создания систем, интегрирующих информационные ресурсы Российской академии наук (РАН), были начаты в 1998 году. Благодаря поддержке Межведомственной программы «Национальная сеть компьютерных телекоммуникаций для науки и высшей школы» был реализован пилотный проект «Интегрированная система информационных ресурсов (ИСИР) РАН». Успешное завершение этого проекта позволило развернуть работы по интеграции разнородных научных информационных ресурсов в общеакадемическую научную информационную систему. В 2001 году по инициативе Отделения математики РАН была принята новая программа целевых расходов Президиума РАН «Информатизация научных учреждений и Президиума РАН». Главной задачей этой программы стала поэтапная интеграция информационных ресурсов организаций РАН в объединенное информационное пространство – Единую информационную систему (ЕИС) РАН. Координация этих работ осуществлялась Советом РАН «Научные телекоммуникации и информационная инфраструктура». Основная часть работ по собственно разработке системы была выполнена в Отделе систем математического обеспечения Вычислительного центра (ВЦ) РАН и Отделе информационных технологий Центра научных телекоммуникаций (ЦНТК) РАН. Первоочередной задачей проекта ЕИС РАН стала разработка концептуальной основы и инфраструктуры для интеграции разнородных информационных и вычислительных ресурсов организаций РАН в единое информационное пространство. Единое информационное пространство (информационную инфраструктуру фундаментальных и прикладных исследований РАН) должны составлять всевозможные цифровые библиотеки, информационные и вычислительные системы организаций РАН, использующие как собственные принципы организации, так и, по возможности, технологию открытой архитектуры проекта ЕИС или непосредственно ее релизы. В

результате был подготовлен системный проект, который определил структуру системы, как таковой, типы информационных ресурсов, участвующих в системе, общую функциональность компонентов системы. В проекте также были отражены принципы организации распределенной системы и интеграции в систему уже существующих ресурсов.

Важным этапом был проект создания Единого Научного Информационного Пространства (ЕНИП) РАН. Этот проект основывался на подсистеме «Научный институт РАН», созданной в ВЦ РАН и ЦНТК РАН. Эта подсистема обеспечивает возможность интеграции информационных ресурсов отдельных организаций в ЕИС. На базе этой системы были реализованы веб-информационные системы ряда институтов и отделений РАН, а также такие информационные системы, как Научное наследие России, портал интеграции пространственных данных Геомета и ряд других. Системы была запущена в опытную эксплуатацию.

1. ИНТЕГРИРОВАННАЯ СИСТЕМА ИНФОРМАЦИОННЫХ РЕСУРСОВ (ИСИР) РАН

С 2001 года выполнялась целевая программа Президиума РАН «Информатизация научных учреждений и Президиума РАН» (с 2004 года – «Информатизация»). С самого начала официальной деятельности по программе значительные усилия были приложены к выработке согласованного системного взгляда на стоящие проблемы и пути их решения, к формированию целей и задач, подходов к решению, базовых требований к используемым методам (технологиям, стандартам и т. п.). В связи с этим с участием всех заинтересованных сторон был разработан ряд документов, положенных в основу большинства проектов, выполняющихся в рамках Программы. В целом надо сказать, что до определенного момента все процессы, связанные с применением информационных технологий в РАН, двигались полностью бессистемно, не управлялись и не контролировались. Тем более не было никакого анализа полученных результатов, эффективности вложений и т. п. С момента деятельности рабочих групп по программе информатизации эта бессистемность постепенно начала исчезать. Кроме бессистемности, проблемы были еще и такими:

• отсутствие полного понимания, согласованного со всеми заинтересованными сторонами в РАН, необходимости развития работ в направлении интеграции;

• как следствие, задержки при окончательной формулировке и принятии общей концепции и программы работ по информатизации РАН;

• отсутствие юридической базы, которая могла бы создать условия для защиты авторских прав и прав интеллектуальной собственности на разработки, выполняемые в РАН;

• различные уровни подготовленности организаций РАН к внедрению и использованию современных ИТ;

• отсутствие или недостаточная подготовленность к интеграции базовых информационных блоков, которыми должны быть информационные системы Институтов, Центральных библиотек, Отделений и Президиума РАН;

• отсутствие адаптированных к требованиям РАН разработок в области стандартизации объектов и механизмов единой системы;

• как следствие, отсутствие возможности полноценного обмена информацией в электронном виде.

Общая задача ИСИР РАН [1, 2] состоит в организации единого информационного пространства. Это требует решения задач по извлечению и структуризации метаданных, по обеспечению их ввода в структурированном в виде. Второй класс задач состоит в предоставлении средств интеграции информации разнообразных информационных систем (репозиториев), тем или иным способом накопивших структурированную информацию. Модель данных представлена на рис. 1. С точки зрения пользователя ИСИР представлена как Портал РАН. Портал реализован (совместно с ИПИ РАН) как Информационно-поисковый справочник РАН, ориентированный на накопление и предоставление оперативной научно-административной информации. В настоящий момент основными типами ресурсов справочника (Портала РАН) являются следующие:

• Организации РАН в соответствии со структурным делением РАН (президиум, отделения, секции, научные центры, филиалы РАН) и сведения о них;

• Сотрудники РАН (аппарат РАН, аппарат отделений, руководство организаций и учреждений, научные сотрудники) и сведения о них (адреса, телефоны и т.д.);

• Публикации;

• Проекты.

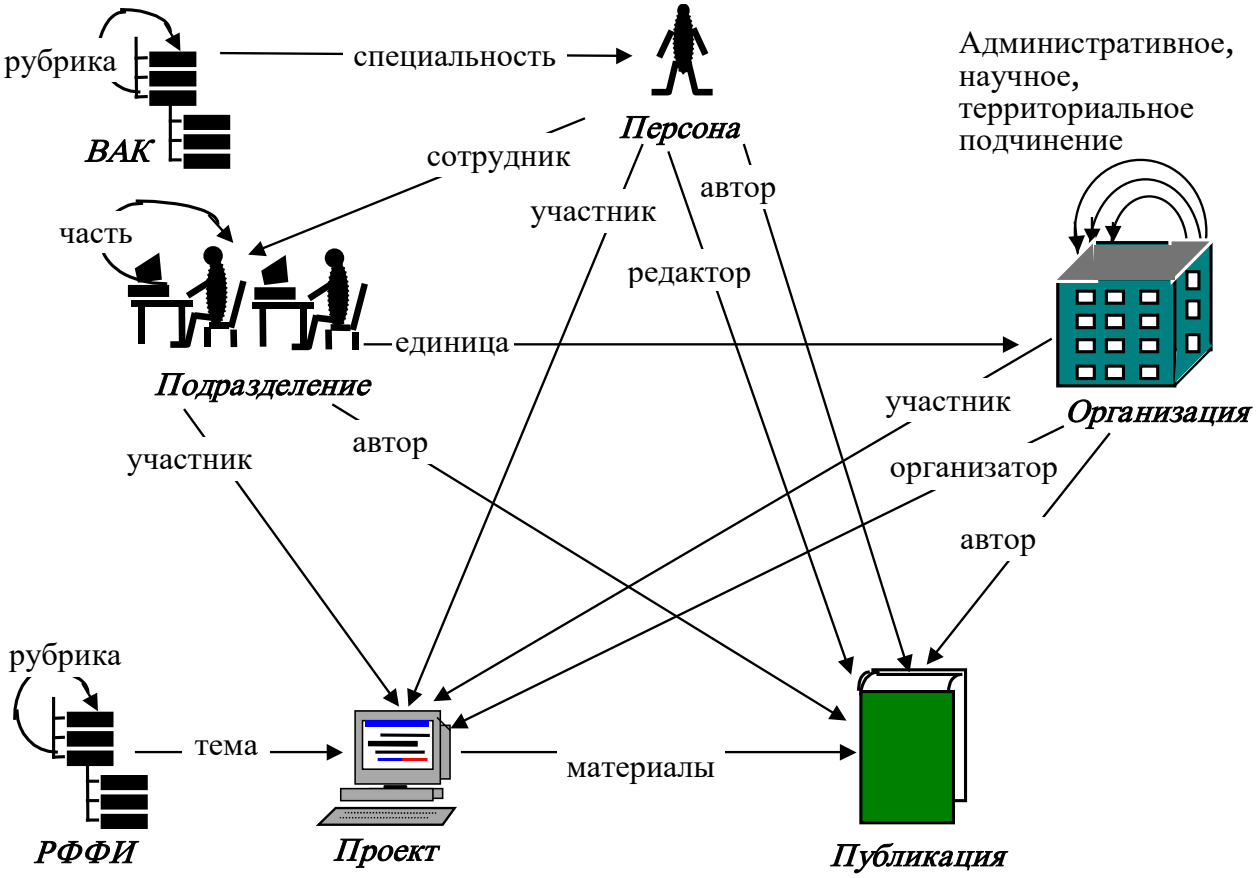

Рисунок 1. Модель данных ИСИР РАН

Справочник РАН отражает организационно-структурное деление РАН, позволяет получить информацию о структурных подразделениях РАН и обеспечивает доступ к информационным ресурсам этих подразделений, данным о сотрудниках учреждения, их научной деятельности. Исполнительная система справочника обеспечивает следующие возможности:

- Просмотр информации и средства навигации по структуре информации;
- Поиск информации по различным видам запросов и просмотр выданной по запросам информации;
- Средства ввода, редактирования и сопровождения информации;
- Средства администрирования непосредственно в подразделениях РАН.

При выборе платформы для был произведен анализ существующих инструментальных средств, пригодных для создания подобной системы. Выбор был сделан в пользу платформы ASP.NET, как обеспечивающей максимальную производительность, удобные средства разработки, компонентную ориентированность, открытость и расширяемость архитектуры, позволяющую вмешиваться практически во все этапы обработки поступающих веб-запросов. Текущее состояние портала представлено на рис. 2.

Рисунок 2. Портал РАН

2. ИНФОРМАЦИОННАЯ СИСТЕМА НАУЧНЫЙ ИНСТИТУТ (НИ) РАН

Рассмотрим типовой научный институт, входящий в состав РАН. Он представляет собой полноценную организацию со сложной административной структурой, основным направлением деятельности являются научные исследования.

Задачи, решаемые каждой такой структурной единицей РАН, можно разделить по своему типу – административные, научные, публичные и т. д.

Административные задачи. В любой организации для нормального функционирования требуется постоянное решение управленческих задач, влияющих прямым образом на деятельность организации в целом и выполнение конкретных задач на всех уровнях. Сюда входят такие задачи, как управление организационной структурой и кадрами, управление проектами, обеспечение документооборота и пр.

Научные задачи. Основным направлением деятельности любого научного института РАН являются научные исследования, а основной задачей организации становится в этой плоскости обеспечение научной деятельности сотрудников.

Публичные задачи. Взаимодействие с другими научными учреждениями, организация и проведение конференции и научных семинаров, публикация научных трудов сотрудников, предоставление доступа к результатам научных экспериментов, научным данным – все это составляет неотъемлемую часть деятельности научного института. Информационная система Института РАН должна, с одной стороны, стать центром научно-информационного сервиса сотрудников Института, а с другой, – обеспечивать полное представление информации о научной деятельности Института для мирового сообщества. Информационная система Института РАН должна представлять собой узел в распределенной архитектуре множества узлов – информационных систем Институтов РАН. На основе описанных выше задач научных организаций в составе РАН можно сформулировать набор требований к программному комплексу ИС «НИ РАН». Информационная система «Научный Институт РАН» должна:

• обеспечивать решение основных информационных задач научного института в составе Российской Академии Наук;

• позволять гибко изменять конфигурацию системы под нужды конкретной организации, реализацию новых модулей для решения специфических задач;

• предоставлять средства интеграции и структуризации существующих данных;

• обеспечивать поддержку распределенного взаимодействия, в том числе со сторонними системами (через специализированные адаптеры, создаваемые отдельно).

Система должна включать:

• средства интеграции существующих данных;

• автоматизированные интерактивные средства структуризации и пакетной загрузки данных;

• пользовательские и административные интерфейсы ввода новых данных и управления уже находящимися в системе данными;

• систему (возможно распределенную) хранения данных;

• систему безопасности, обеспечивающую аутентификацию пользователей и авторизацию доступа к ресурсам системы;

• спецификации по разработке дополнительных модулей, обеспечивающих решение специфических задач научного института.

ИС «НИ РАН» [3, 4] представляет типовой собой программный комплекс автоматизации информационной деятельности научного института в составе Российской академии наук, обеспечения научной деятельности его сотрудников, взаимодействующий с другими информационными системами в составе ЕНИП. Разработанная платформа ИС «НИ РАН» предоставляет широкие возможности по конфигурированию под нужды конкретного научного института. Ядро всей системы составляют инфраструктурные службы. Они обеспечивают хранение, индексирование и поиск ресурсов, обеспечивают безопасность и взаимодействие между другими модулями. Базовые компоненты ИС «НИ РАН» обеспечивают выполнение самых общих информационных задач научного института – управление содержанием портала, организационной структурой, ведение сведений о публикациях и проектах сотрудников. Все действия конечный пользователь производит через веб-интерфейс. ИС «НИ РАН» представляет собой модульную расширяемую систему, решающую типовые информационные задачи научного института в составе РАН. Но реальные потребности таких организаций и их сотрудников зачастую бывают очень специфичными и относятся к узкой предметной области. Для удовлетворения таких нужд разрабатываются прикладные подсистемы, расширяющие функциональность типового решения в конкретных экземплярах. Архитектура НИ РАН представлена на рис. 3.

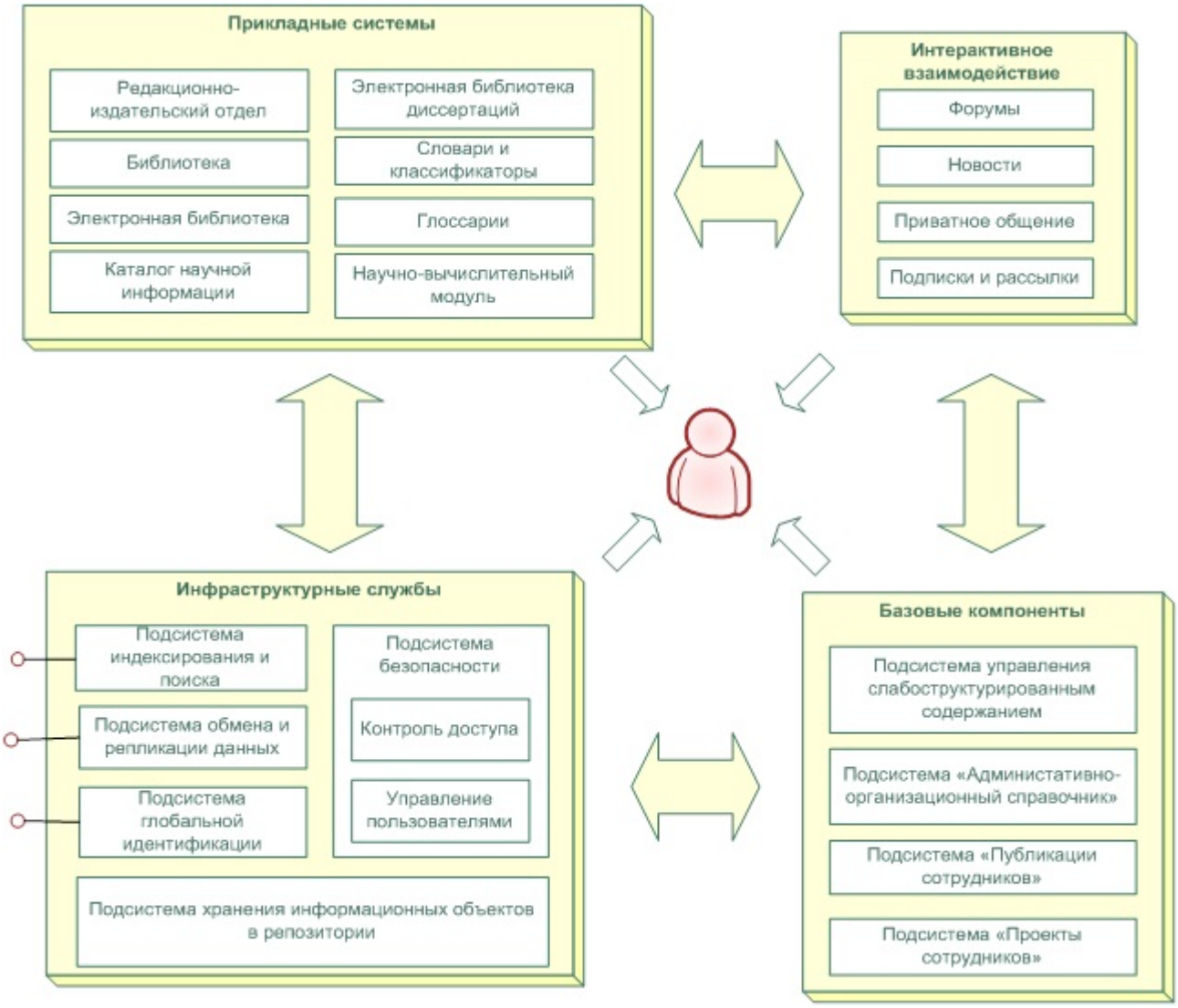

Рисунок 3. Архитектура НИ РАН

В основном профиле метаданных можно выделить общую поддержку следующих четырех основных групп информационных сущностей:

«*Участники научной деятельности*» – центральное звено, вся информация в РАН связана с научной деятельностью ее сотрудников, «Персон», образующих разнообразные организационные объединения от формальных («Организации» и «Подразделения») до неформальных («Коллективы», «Сообщества», «Рабочие группы»);

«*Научная деятельность*», в частности, «Проекты», отражающие процесс научной деятельности, информация о результатах проектов, патентах и т. п., а также «Научные мероприятия» – как разовые, так и повторяющиеся, такие, как «Конференции», «Семинары», «Симпозиумы»;

«*Результаты научной деятельности*», в которые могут входить «Интернет-системы» – вебсайты и пр., «Базы данных», предоставляющие автономные коллекции информации с той или иной степенью интеграции с ЕНИП и т. п., «Экспериментальные данные» и их «Математические модели», «Программные системы», в частности, «Научные вычислительные приложения», «Экспериментальные установки», «Изобретения», «Технологии» и т. п.

«Документы и публикации» – ресурсы этого типа представляют собой научные труды, статьи, отчеты сотрудников (научные «Публикации» и «Диссертации» сотрудников), возможно, административные «Постановления» и «Распоряжения». Примерами специализации публикации могут служить, например, «Тезисы конференций» и т. п. Профиль метаданных НИРАН представлен на рис. 4.

На базе информационной системы Научный институт РАН были созданы информационные системы ряда организаций (институтов и отделений) РАН:

- Отделение общественных наук (на базе системы Соционет);
- Библиотека по естественным наукам (БЕН РАН);
- Вычислительный центр (ВЦ РАН);
- Институт физики твердого тела;
- Палеонтологический институт;
- Пермский научный центр и Институт механики сплошных сред УРО;
- Институт проблем химической физики и Научный центр Черноголовка;
- Тихоокеанский океанологический институт им. В.И. Ильичева;
- Отделение математических наук (ОМН);
- Санкт-Петербургский научный центр;
- Дальневосточное отделение (ДВО РАН);
- Институт научной информации по общественным наукам (ИНИОН);
- Институт США и Канады (ИСКРАН);
- Институт проблем информатики (ИПИ РАН);

- Портал пространственных метаданных ГеоМета;
- Цифровая библиотека Научное наследие РАН;
- Северокавказский научный центр.

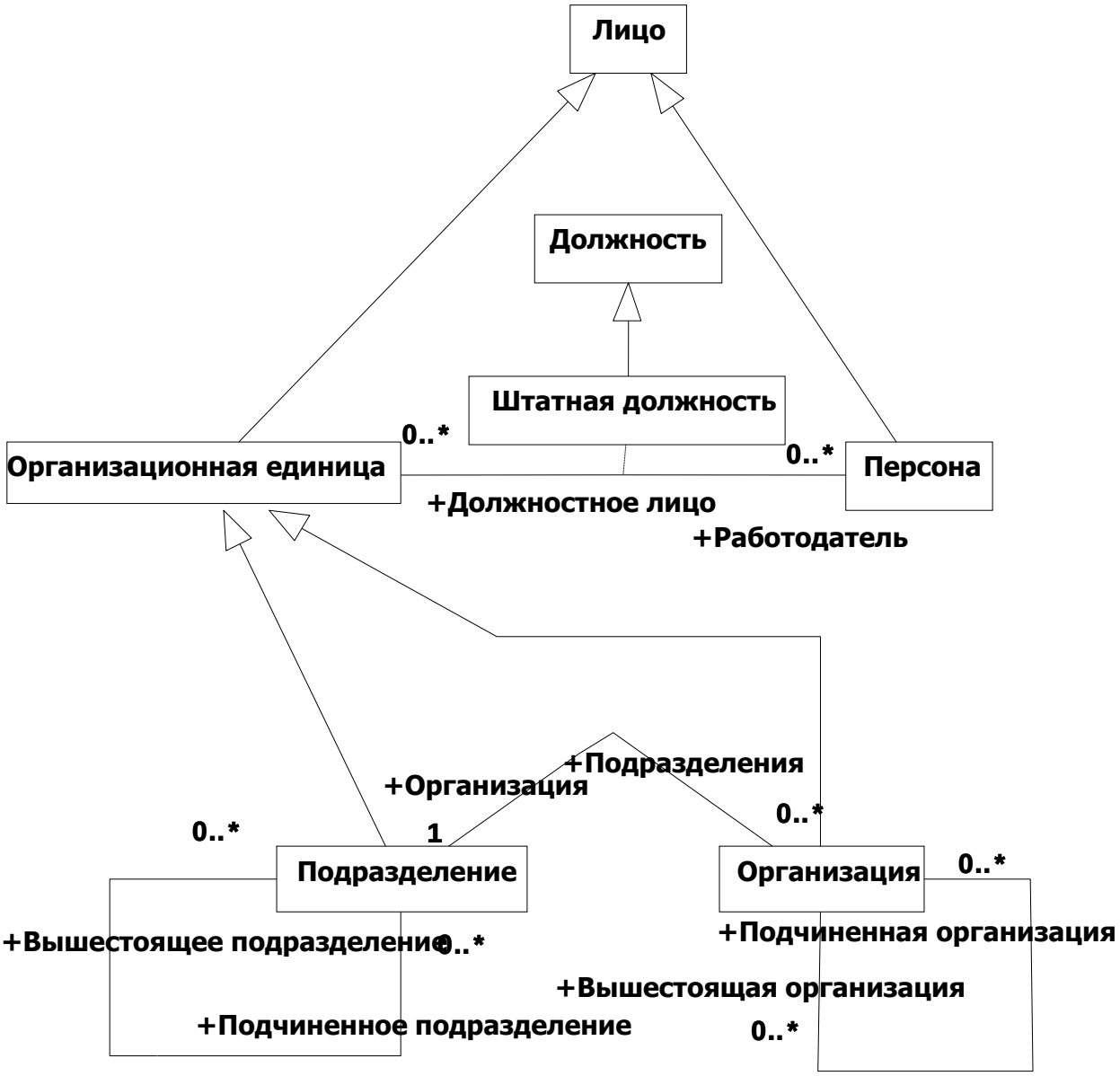

Рисунок 4. Основной профиль метаданных НИ РАН

3. ЕДИНОЕ НАУЧНОЕ ИНФОРМАЦИОННОЕ ПРОСТРАНСТВО РАН

Российская академия наук обладает уникальными научными информационными ресурсами. Среди них – опубликованные результаты научных исследований и экспериментов, библиографические и фактографические базы данных, сведения об ученых, их научной деятельности, публикациях, проектах и т. п. Эти ресурсы представляют значительный интерес для сотрудников РАН, членов мирового научного сообщества, для представителей промышленности и предпринимателей, которые заинтересованы во внедрении результатов научных исследований. Предполагалось, что Единое Научное Информационное Пространство (ЕНИП) РАН должно стать интегрированным источником научной информации.

Система предусматривала объединение сведений о разнородных научных информационных ресурсах РАН, обеспечение актуальности этих сведений и широких возможностей для достаточно точного поиска научных ресурсов на основе этих сведений, поддержку средств научной коммуникации, сервисов, связанных с возможностью оперативного информирования пользователей о необходимых им ресурсах, и т. п.

Такая система могла обеспечить пользователей актуальными данными о текущем состоянии и характеристиках информационно-научной базы институтов РАН и их подразделений, упростить анализ состояния и тенденций развития науки. Облегченный доступ к информации мог бы изменить способы ведения научной деятельности, способы обучения. Для обеспечения взаимодействия существующих разнородных научных систем в рамках ЕНИП предполагалось выработать корпоративные стандарты на интерфейсы взаимодействия, а также профили метаданных, что позволило бы реализовать инструментальные средства, обеспечивающие интеграцию данных в единую среду. Результатом решения этих первоочередных проблем должны были явиться предложения ЕНИП по:

• типовым интерфейсам взаимодействия (форматы данных, протоколы обмена) отдельных информационных источников (организаций РАН, поддерживающих собственные научные информационные ресурсы);

• профилям метаинформации, предоставляемой этими источниками; в частности, производится разработка набора элементов метаданных для научной информации общего характера, предложений по формированию элементов

метаданных для отдельных областей науки и согласование их с научным сообществом и международными открытыми стандартами;

• справочникам и классификаторам ресурсов;

• реализации политики информационной безопасности и требований по разграничению прав доступа к цифровым ресурсам.

Инициатива по организации Единого Научного Информационного Пространства (ЕНИП) РАН призвана была помочь научным коллективам сделать ряд шагов в направлении интеграции разнородных научных информационных и программных ресурсов отдельных научных учреждений, предоставлении пользователям более эффективных средств интеграции и поиска информации, научной коммуникации, сотрудничества и совместной работы. Под единым пространством понимается не формирование централизованной системы, не навязывание всем одних и тех же решений, а стремление последовательностью практических шагов, совместными усилиями научных коллективов:

• сформулировать взаимосогласованный набор соглашений, правил и открытых стандартов;

• приготовить совокупность макетов и типовых решений для реализации адаптеров прикладных систем, инфраструктурных служб, поддерживающих разные уровни интероперабельности распределенных гетерогенных данных и приложений;

• создать ряд информационных систем общего назначения, следующих этим соглашениям, использующих эти реализации, допускающих модульную организацию, наращивание функциональных возможностей;

• применить эти результаты для решения соответствующих задач научных учреждений.

Основу ЕНИП РАН составляют, прежде всего, стандарты на метаданные информации, циркулирующей в ЕНИП. Эти стандарты должны отвечать следующим требованиям:

• включать в себя основные типы информации, требующейся для поддержки работы научного сотрудника;

• быть открытыми, т. е. обеспечивать доступ к соответствующей информации по этим описаниям;

• быть расширяемыми, т. е. обеспечивать возможность детализации описаний;

• обеспечивать возможности интеграции информации;

• обеспечивать возможности уникальной идентификации информации;

• обеспечивать возможности размещения и поиска информации в распределенной среде;

• быть ориентированными на современные и перспективные технологии описания и использования информации (в нашем понимании – ориентироваться на семантический Веб (Semantic Web));

• обеспечивать возможности интероперабельности с внешней средой.

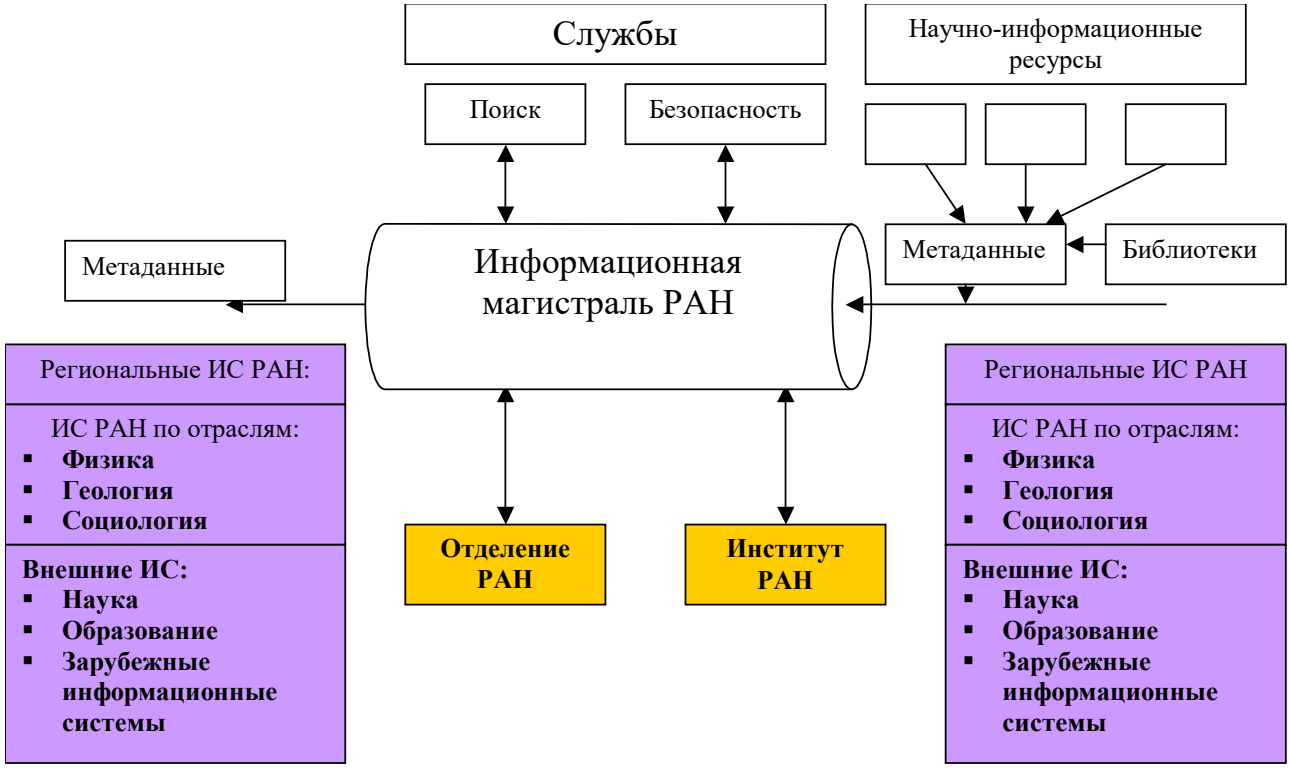

Рисунок 5. Информационная магистраль ЕНИП РАН

Основу единого информационного пространства РАН составляет Информационная магистраль ЕНИП РАН (рис. 5), представляющая собой комплекс аппаратных, программных и организационных мер, обеспечивающих:

• формирование состава цифровых ресурсов и служб ЕНИП РАН;

• предоставление доступа к цифровым ресурсам и службам ЕНИП РАН;

• обеспечение защиты цифровых ресурсов и служб ЕНИП РАН;

- ведение и поддержка в актуальном состоянии метаданных системы;
- поиск по хранимой метаинформации и идентификация ресурсов;
- интеграцию ресурсов различных областей и отраслей знаний.

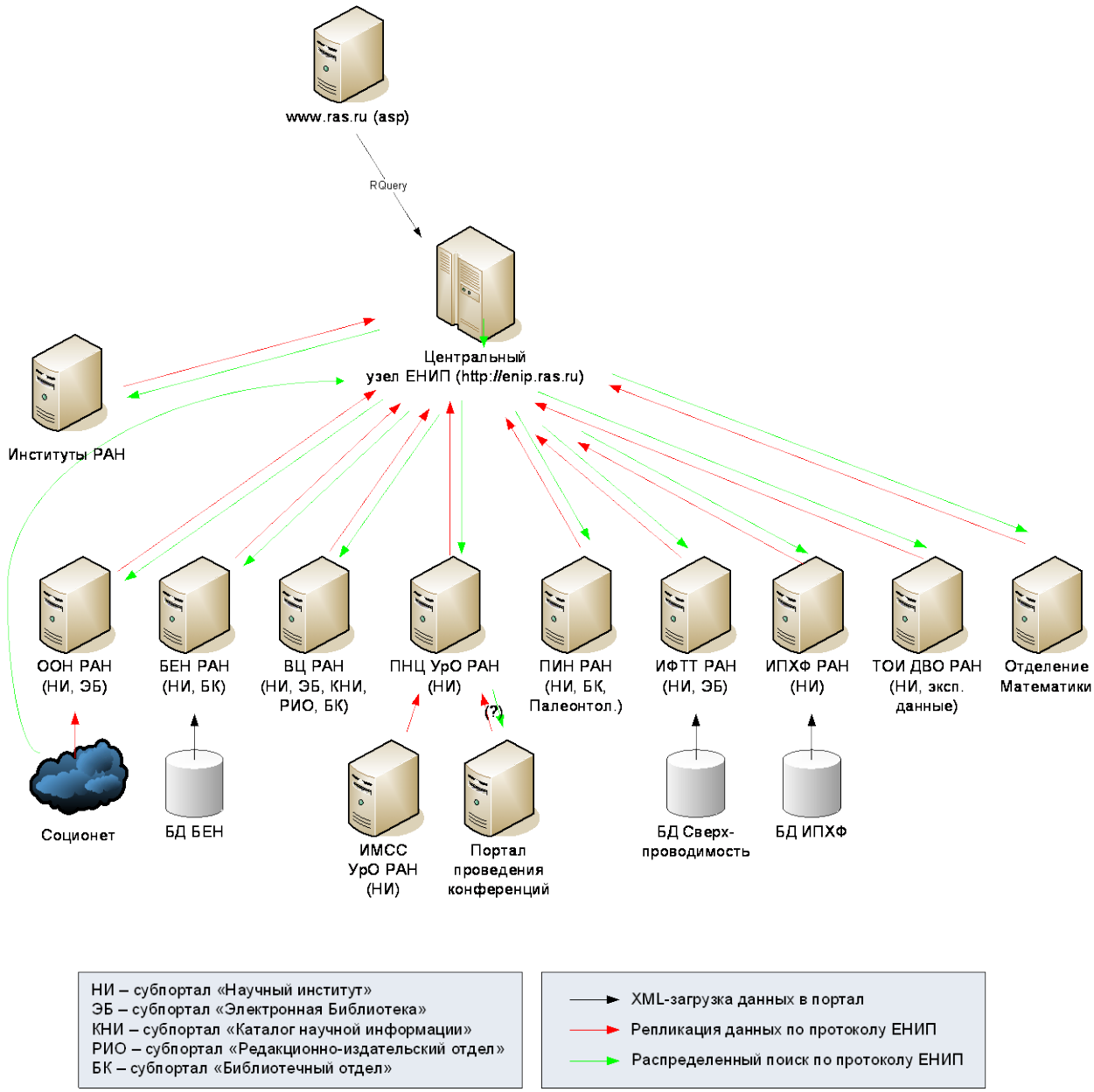

Рисунок 6. Схема взаимодействия узлов ЕНИП

На рис. 6 изображена схема взаимодействия узлов ЕНИП –четыре типа узлов ЕНИП: центральный узел; узлы организаций; независимые источники данных; независимые системы, включенные в ЕНИП.

Центральный узел осуществляет интеграцию данных с других узлов с помощью механизма репликации (копирования) метаинформации. На основе

реплицируемой на центральный узел метаинформации строятся поисковые индексы и на их основе осуществляется единый поиск по этим узлам. Загрузка данных в узлы системы может осуществляться из других источников, например, из сайтов организаций. Независимые информационные системы могут быть включены в ЕНИП самостоятельно, если обеспечены протоколы взаимодействия. Система ЕНИП оперирует такими ресурсами, как персоны, публикации, организации, подразделения и проекты. Данные по этим ресурсам обновляются каждую неделю. Центральный узел предоставляет пользователям две возможности поиска: поиск по локальной базе данных и полнотекстовый поиск. Поиск по локальной базе осуществляется по стандартным ресурсам: персона, организация, публикация, проекты. Актуальность информации может составлять разницу в 6 дней от информации на сервере-источнике данных. Полнотекстовый поиск позволяет получить полную и актуальную информацию, но выполняется дольше.

4. ЦБ «НАУЧНОЕ НАСЛЕДИЕ РОССИИ»

Учитывая важность формирования цифровых библиотек, Российская академия наук приняла в 2006 году целевую научную программу «Создание ЦБ «Научное наследие России»». ЦБ призвана аккумулировать цифровые копии книг, статей, документов, хранящихся в библиотеках, архивах и музеях РАН. В первую очередь акцент сделан на перевод в цифровую форму редких и уникальных изданий, важнейших документов по истории РАН, материалов экспозиционного характера, включая аудио-видеоматериалы [5]. Основной целью создания ЦБ является предоставление через интернет всем желающим информации о выдающихся российских ученых, внесших вклад в развитие фундаментальных естественных и гуманитарных наук, с возможностью ознакомления с полными текстами опубликованных ими наиболее значительных работ. Исходя из этой цели, в ЦБ было решено включать не только электронные версии книг, но и развернутые сведения о российских ученых – биографические данные, основные этапы их научной деятельности, разнородную архивную и музейную информацию, отсканированные фотографии, аудио- и видеозаписи, относящиеся к теме научного наследия.

Другой целью создания ЦБ является обеспечение сохранности оригиналов изданий, являющихся исторической ценностью, поскольку возможность работы с цифровыми копиями существенно снижает потребности в работе с печатными

материалами, а каждая «книговыдача» на руки раритетных изданий сокращает срок их «жизни».

Третьей целью создания ЦБ является включение сведений об ученых и основных результатах их научной деятельности в Единое научное информационное пространство (ЕНИП) РАН. На начальном этапе реализации основными задачами Программы были разработка основных принципов формирования ЦБ, технологии сканирования, обработки и предоставления пользователям материалов, включаемых в Библиотеку, а также создание программного обеспечения, сопровождающего все этапы создания ЦБ.

В основу технологии формирования ЦБ положен принцип распределенного наполнения и централизованной поддержки. Руководство Программой осуществляет Межведомственный суперкомпьютерный центр РАН, осуществляющий вместе с ВЦ РАН и БЕН РАН разработку технологии и программного обеспечения наполнения и поддержки ЦБ. Основными поставщиками информации для загрузки в ЦБ в настоящее время являются центральные академические библиотеки (БАН и БЕН РАН с их отделами в институтах и научных центрах РАН), ИНИОН, Центральный архив РАН с его Санкт-петербургским филиалом, Геологический музей РАН им. В.И. Вернадского, Институт русской литературы РАН (Пушкинский дом). В настоящее время наполнение ЦБ осуществляется копиями изданий, которые не подпадают под действие закона о защите авторских прав (в основном это издания, вышедшие из печати до 1920-го года). Основные элементы функциональности распределенной цифровой библиотеки следующие (рис. 7):

• доступ к ресурсам – запрос, определение местоположения, извлечение, трансформация и сохранение ресурса; поиск может осуществляться как по атрибутам ресурса, так и по полным текстам;

• управление ресурсом – создание нового ресурса, внесение его в ЦБ, удаление старого ресурса и изменение существующего;

• управление метаданными – их создание, обработка и преобразование; состав метаданных определяется соглашениями;

• управление словарями – их создание, обработка и преобразование; состав словарей определяется соглашениями;

• управление участниками – их регистрация, подписка, права доступа и персональная информация;

• управление цифровой библиотекой – управление коллекциями, группами пользователей, членством, так же, как общее управление политикой, качеством или функциональностью;

• системное администрирование – установка, конфигурирование, необходимые периодические мероприятия, восстановление после сбоев и мониторинг ЦБ.

Рисунок 7. Функциональная схема ЦБ «Научное Наследие»

Цифровая библиотека строится как распределенная информационная система с выделенным центральным узлом. Узлы системы, с одной стороны, являются точками входа в цифровые библиотеки организаций – участников проекта, с другой, – поставщиками информации для всей распределенной системы. Таким образом, ключевой принцип архитектуры – независимое развитие цифровых

библиотек организаций – участников с одновременной интеграций данных в единое информационное пространство. Это достигается стандартизаций предоставления метаданных, форматов предоставления данных, интерфейсов поиска и словарей. Таким образом, каждая из цифровых библиотек организаций – участников может хранить данные в собственных форматах и предоставлять собственные сервисы, но в то же время должна обеспечить единые для всех интерфейсы, упомянутые выше.

Центральный узел системы должен обеспечить навигацию, поиск и предоставление данных по всем цифровым библиотекам в соответствии с унифицированными форматами и сервисами. Сервера хранения оцифрованных данных обеспечивают надёжное хранение и резервирование оцифрованных данных библиотеки, а также подмножества метаданных, отражающих структуру информации (например, оглавление книг). Кроме того, на них возлагается задача по предоставлению доступа к данным конечных пользователей, перенаправленных с центрального портала цифровой библиотеки. Серверами хранения данных для центров оцифровки предоставляются также средства автоматизации размещения и поддержания актуальности данных. Центральный веб-портал цифровой библиотеки «Научное Наследие РАН» осуществляет консолидацию метаданных, полученных из центров оцифровки, в рамках централизованного хранилища, обеспечивая, таким образом, централизованный доступ к ним пользователей. Взаимодействуя с серверами хранения оцифрованных данных, он является также единой точкой доступа к электронным версиям научных трудов. Второй задачей, решаемой центральным порталом, является обеспечение интеграции библиотеки в ЕНИП РАН путём предоставления на центральный сервер ЕНИП метаданных, по которым возможен распределённый поиск.

Функциональная схема ЦБ «Научное Наследие» приведена на рис. 7, главная страница – на рис. 8.

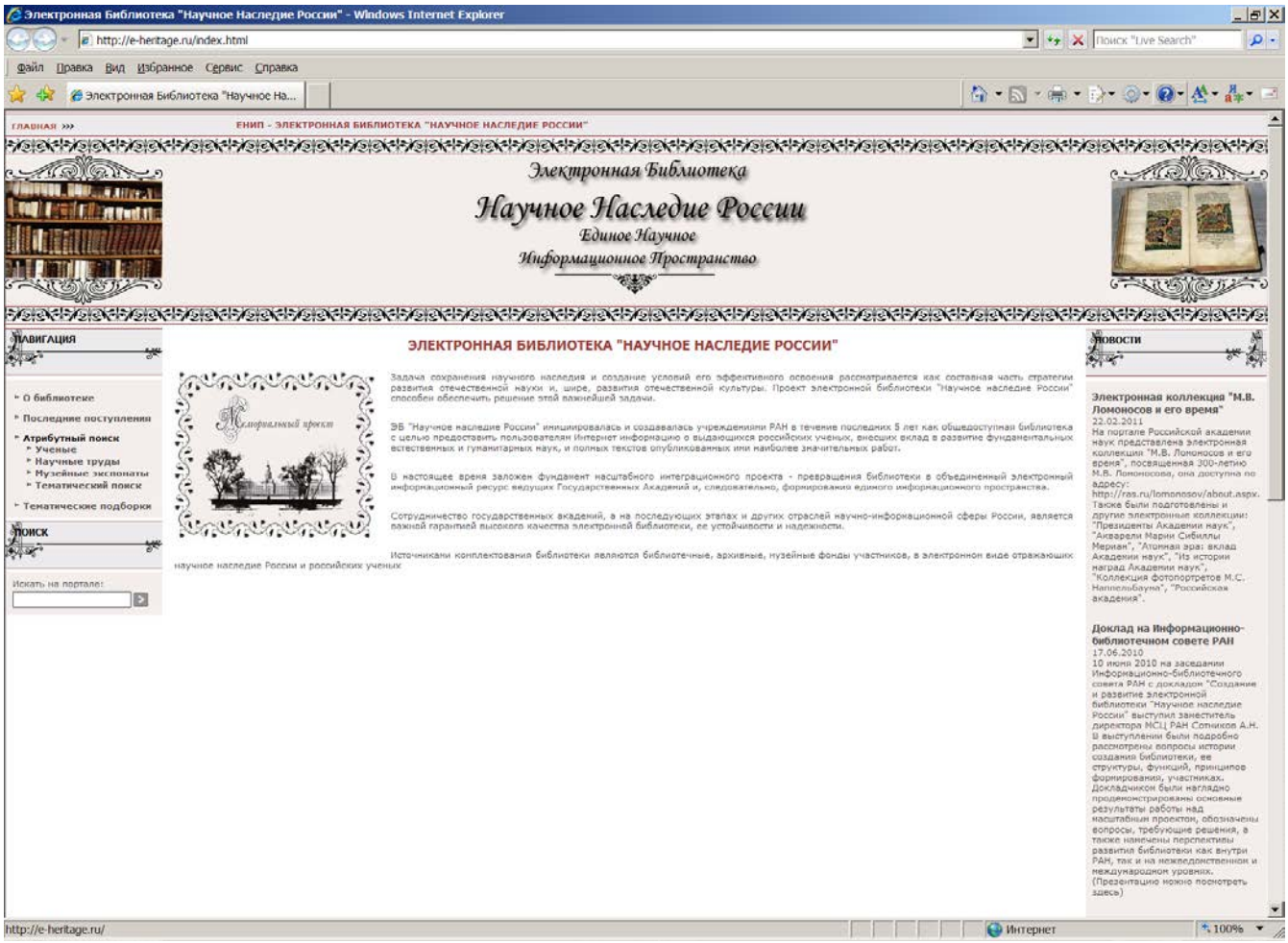

Рисунок 8. Главная страница ЦБ «Научное Наследие»

5. ПОРТАЛ ГЕОМЕТА

К настоящему времени в учреждениях РАН накоплен большой опыт использования геоинформационных технологий, реализованы многочисленные геоинформационные проекты, созданы базы и банки пространственных данных. Академические ресурсы пространственных данных составляют значительную часть национальных информационных ресурсов. Основным производителем пространственных данных являются учреждения геологического, геофизического, географического и экологического (природоохранного) профилей. В то же время данные рассредоточены, их использование ограничено зачастую рамками того проекта, где они созданы, затруднены или невозможны поиск существующих данных и доступ к ним, не налажен обмен ими. Причина этого – отсутствие эффективной системы управления пространственными данными. Ее создание позволит интегрировать данные и знания о территории, строить и использовать модели

природных и социально-экономических явлений и процессов, их взаимодействия в системе «общество – природная среда», использовать методы пространственного анализа, обеспечивать территориальное планирование и управление. В целом в учреждениях РАН имеется опыт выполнения разнообразных геоинформационных проектов для различных приложений, сформированы подразделения, отделы и лаборатории геоинформатики, укомплектованные высокопрофессиональными научными кадрами, располагающими необходимой технической базой, современными программными средствами геоинформационных систем (ГИС) и данными, то есть созданы необходимые условия для разработки ГИС и их интеграции.

Основным инструментом интеграции и предоставления пространственных данных в настоящее время являются геопорталы. Понятие «геопортал» означает точку входа в интернет с инструментами просмотра метаданных, поиска географической информации, ее визуализации, загрузки, распространения и, возможно, поиска геосервисов. Современное требование к системам поддержки геопорталов – независимость, расширяемость и гибкость компонентов, являющаяся важной особенностью современной программной системной архитектуры. Существует потребность в объединении этих данных, имеющих распределенный характер, в концептуально одну информационную систему, в обеспечении централизованного доступа к ним, в создании на основе интернета технологий единого информационного пространства геоданных.

Портал «ГеоМета» [6] – это стандартизированная и децентрализованная среда управления пространственной информацией, разработанная для доступа к базам геоданных, картографическим продуктам и связанным с ними метаданным из различных источников, облегчающая обмен пространственной информацией между организациями и ее совместное использование посредством интернета. Этот подход к управлению географической информацией имеет целью предоставить широкому сообществу пользователей средства для простого и своевременного доступа к имеющимся пространственным данным и существующим тематическим картам, которые могут оказаться полезными для поддержки информированного принятия решений. Главная цель портала – увеличить доступность разнообразных междисциплинарных данных различного масштаба вместе с

сопутствующей информацией, организованных и документированных стандартным и непротиворечивым способом, улучшить кооперацию и координацию усилий при сборе данных, сохраняющих ресурсы и, в то же самое время, ограждающих данные и информацию от нежелательного доступа.

Портал «ГеоМета» представляет собой платформу для создания распределенной среды интеграции неоднородных источников геоинформационных данных и предоставления к этой среде единой точки входа (веб-портала), которая позволит ученым в сфере наук о Земле легко находить специализированные данные и приложения, производить вычислительные эксперименты, визуализировать результаты деятельности.

Благодаря тому, что портал «ГеоМета» построен на базе ИС «НИ РАН» [1], являющейся базовым инфраструктурным компонентом ЕНИП [2], он может интегрироваться в ЕНИП с предоставлением расширения схемы геопространственными метаданными и геоданными. К функциональностям ГИС-части системы относятся:

• каталогизация, сбор, поиск геопространственных метаданных;

• размещение геоданных в собственном хранилище и предоставление к ним доступа;

• предоставление доступа к распределенным геопространственным данным по стандартизованным протоколам;

• визуализация карт, редактирование элементов.

Интерфейс системы представлен веб-порталом, поэтому для ГИС-части основным методом доступа пользователя к информации является обычный доступ к веб-страницам портала через любой распространенный браузер. Ядро системы предоставляет следующие возможности: управление статическим содержанием; хранение объектов системы (представленных RDF-тройками) в реляционных СУБД; индексирование и полнотекстовый поиск; обеспечение безопасности системы. Система поддерживает следующие основные типы ресурсов: *Пространственные данные* (картографические данные и их метаданные) и дополнительные типы ресурсов, такие, как *Организация, Персона, Публикация, Проект,* различные рубрикаторы и классификаторы.

Ресурс *Пространственные данные* содержит наборы пространственных данных и метаданные распределенных пространственных данных. Ресурс

Организация включает организации РАН, научные центры и другие организации. Данные об их сотрудниках сопоставлены ресурсу *Персона*. Ресурс *Проект* поддерживает сведения о проектах, выполненных или ведущихся в РАН и других ведомствах. Ресурс *Публикация* представляет данные о публикациях и научной деятельности.

Доступ к порталу осуществляется интерактивно через интернет посредством веб-браузера (например, Netscape Navigator или Microsoft Internet Explorer) по ссылке http://www.geometa.ru.

Главная страница портала представлена на рис. 9.

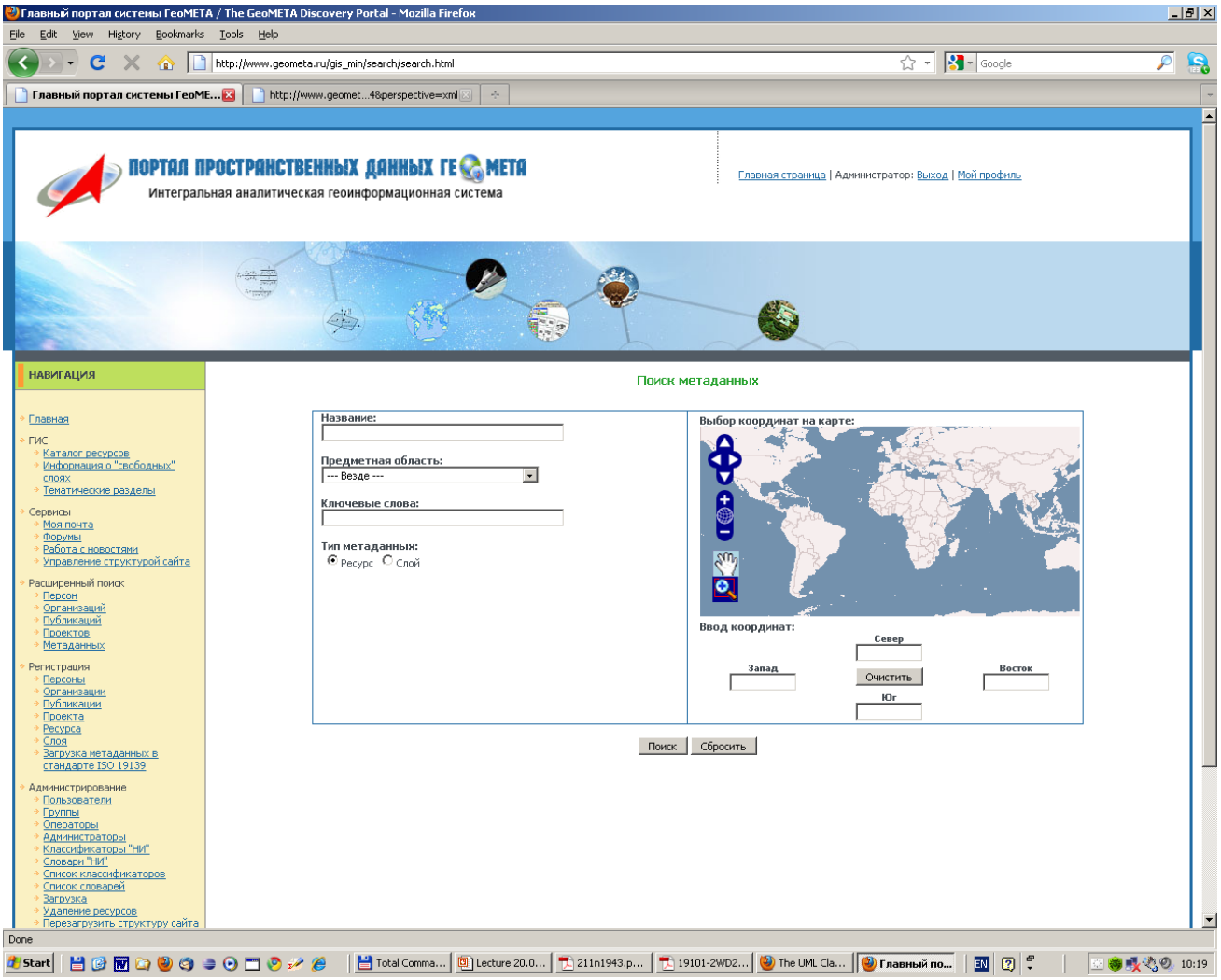

Рисунок 9. Главная страница портала ГеоМета

6. ПЕРСОНАЛЬНЫЕ СЕМАНТИЧЕСКИЕ ЦИФРОВЫЕ БИБЛИОТЕКИ

Под персональными семантическими цифровыми библиотеками подразумеваются такие цифровые библиотеки, наполнение которых индивидуально для каждого пользователя системы и выполняется в полуавтоматическом режиме из разнородных источников данных, интегрированных в облако LOD. Будем далее для краткости называть их персональными открытыми цифровыми библиотеками или ПОЦБ. Типы информационных ресурсов и их структура определяются пользователем, исходя из своих интересов, то есть пользователь описывает интересующую его предметную область, определяя тематическое наполнение библиотеки.

Основная задача системы заключается в предоставлении пользователю унифицированного представления для возможности автоматизированного извлечения интересующей его информации по определенной предметной области. Представление ресурсов библиотеки в виде связанных данных расширяет функциональность семантических цифровых библиотек, давая возможность:

- включения дополнительных элементов описания данных информационных ресурсов;
- полного или частичного обновления данных из источников;
- использования интерфейсов для создания запросов к интегрированным в LOD источникам данных на основе SPARQL;
- включения в описания ресурсов других типов информации.

Одна из задач, которая решается в ПОЦБ, – это реализация интеграции набора данных в пространство LOD с использованием онтологии предметной области информационных ресурсов, т. е. автоматизированное обнаружение новых наборов данных и, по возможности, установка и поддержка связей с элементами данных из этих наборов данных с уже имеющимися ресурсами в репозитории библиотеки, обеспечивая одновременно рекомендуемую проектом LOD функциональность в рамках одной системы.

Источники данных подразделяются на два типа: внешние и внутренние. Внешними мы называем те источники, которые интегрированы в LOD, и данные которых представлены в RDF и доступны нам с использованием SPARQL. Для своих практических целей мы использовали такие известные источники в LOD, как DBpedia, Europeana. Внутренние источники могут представлять собой любой

другой тип источника данных, который не интегрирован в LOD. На практике в качестве внутренних источников мы использовали другие библиотеки, которые предоставляли доступ к своим данным по протоколу OAI-PMH.

К основной функциональности системы, реализующей ПОЦБ относятся:

• функции атрибутного поиска;

• функция выделения неявных связей между ресурсами по их описаниям;

• функция работы с коллекциями;

• создание/просмотр/редактирование/объединение/вложенные коллекции;

• функция отображения онтологии ИД;

• функция детализации, которая обеспечивает преобразование в подзапросы, соответствующих различным ИД;

• функция для выполнения запросов и обработки результатов и предоставления окончательного результата пользователю;

• функция автоматического мониторинга ИД на наличие новых/измененных данных;

• создание словарей, классификаторов, тезаурусов;

• редактирование элементов;

• поддержка («гибкой») классификации ресурсов;

• поддержка настройки уровней доступа к различным ветвям тезауруса.

Исходя из определения источников данных ПОЦБ и перечня функций системы, можно выделить «внутренние» функции, т. е. те, которые оперируют данными в рамках системы и интегрируют данные из «внутренних» источников и фактически определяют обычную семантическую библиотеку. «Внешние» функции обеспечивают подключение и извлечение данных из LOD и позволяют задать тематическое наполнение библиотеки и установить связи, таким образом задавая фактически определение ПОЦБ.

Онтология ПОЦБ разработана в общем виде без привязки к конкретным методам и способам реализации семантических цифровых библиотек. Фактически общая онтология ПОЦБ состоит из двух онтологий:

1) онтология СЦБ, построенная на основе онтологии информационных систем, включающая в себя основные понятия, необходимые для обеспечения

основной функциональности библиотеки, такие, как ресурс, пользователь, коллекция, словарь, классификатор, запрос, источник и т. д.

2) онтология и тезаурус предметной области, для которой пользователь определяет ее понятия, их тип, структуру, совокупность словарей и классификаторов, которые представляют тезаурус предметной области, который обеспечивает доступ неквалифицированных пользователей, решающих задачи поиска информации, к знаниям предметной области в разных источниках. Эта онтология позволяет:

• выработать и зафиксировать общее понимание области знания;

• представить знания в удобном для обработки автоматизированными подсистемами виде, обеспечить возможность получения и накопления новых знаний, а также представить возможность многократного использования знаний.

Тезаурус же обеспечивает терминологическую поддержку и помогает пользователям сформулировать запрос к системе, в том числе, подобрать правильные ключевые слова для описания искомого результата, имеющихся данных и контекстной информации. Задача автоматизированного поиска релевантных источников данных осложняется тем, что чаще всего информация о связях между ними проставляется в основном на уровне данных с помощью связей sameAs, seeAlso. Даже простой анализ связей sameAs, seeAlso на уровне найденных данных позволит выявить эквивалентные классы, ранее не определенные связи между разными источниками или новые источники. Описание связей на уровне схем затем можно использовать при формировании запросов к источникам данных.

До недавнего времени связи между источниками на уровне схем описывались гораздо реже. В последние несколько лет эта задача решается с введением и активным распространением спецификации VOID. Для описания источников RDF-данных, в которой предоставляется информация о связанных источниках данных. VOID-описание содержит информацию об используемых словарях, статистическую информацию о том, сколько ресурсов того или иного типа или значений определенных свойств используются во множестве. При создании словаря VoID была сведена к минимуму необходимость создания новых свойств и классов путем использования существующих словарей. Например, для описания статистической информации используется словарь SCOVO. На основе этой информации можно делать вывод о релевантности источника тому или иному запросу или

предметной области. В рассматриваемой системе VoID-описание набора данных в хранилище генерируется с помощью D2R Server. В сгенерированное описание не попадают информация о подключенных источниках данных и статистика по имеющимся с ними связям. Для включения этой информации были использованы правила, по которым осуществляется поиск связанных данных. Полученное описание в рамках используемой системы позволяет формировать распределенные запросы к подключенным источникам данных в терминах онтологии, используемой в этой системе. С использованием VoID-описаний запросы из системы транслируются в термины уже источников данных. Также это описание применяется для отображения обобщенного результата поиска.

На рис. 10 представлена общая схема подключения различных источников данных с использованием технологий из стека проекта LOD.

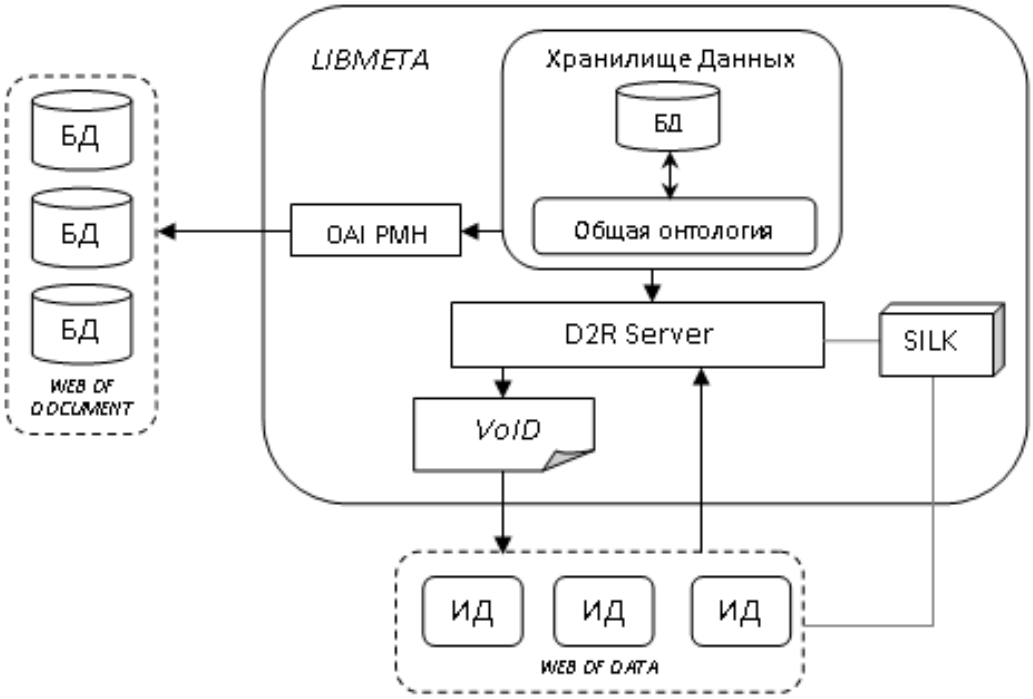

Рисунок 10. Схема подключения различных источников данных

Доступ к данным Libmeta осуществляется через ее общую онтологию, которая, как было сказано, состоит из: а) онтологии семантической библиотеки, б) онтологии предметной области, которая задает тематическое направление информационных ресурсов. При этом D2R Server использует онтологию Libmeta для

создания SPARQL-точки доступа к ее данным. Используются правила, которые задаются для каждого подключаемого источника (правил может быть несколько), с помощью которых осуществляются поиск и сохранение связей между данными Libmeta и источником из LOD. Для задания правил связывания используется фреймворк SILK. Правила описываются в соответствии с требованиями SILK и хранятся в определенном для каждого источника месте. После описания правила и указания его расположения все действия по запуску и анализу результатов работы SILK выполяются программно, для этого используется соответствующая задаче версия фреймворка.

При каждом подключении нового источника или обновлении набора связей уже подключенных нужно обновлять VoID-описание множества данных Libmeta, анализируя полученный набор ссылок и правила, по которым они выполнялись. Это позволит обновить статистическую и структурную части VoID, необходимых для использования при формировании запросов в терминах общей онтологии и их преобразования в запросы к релевантным источникам в соответствующим им терминах.

Libmeta также исторически поддерживает обмен данными по протоколу OAI-PMH с библиотеками, неинтегрированными в LOD, выступая агрегатором, который интегрирует их данные в LOD.

В рамках создания первой версии ПОЦБ был реализован проект по созданию стандартизированной и децентрализованной среды управления информацией электронных фондов Libmeta. В проекте реализованы средства интеграции приложений с разными источниками/каталогами метаданных/данных, сервис директорий метаданных, унифицированный интерфейс поиска данных.

Существенное различие во внутренних моделях данных, используемых в различных музеях, библиотеках и архивах, является главной проблемой на пути решения задачи интеграции данных. Для преодоления этой проблемы в решаемой задаче интеграции данных было предложено участникам экспортировать метаданные из своего внутреннего формата в формат на базе Dublin Core с использованием синтаксиса XML, так как во внутренних используемых форматах удается выделить общую часть, которая ложится в рамки предложенного формата. В системе используется универсальный модуль загрузки метаданных в произвольном XML-формате в соответствии с протоколом OAI-PMH.

ЗАКЛЮЧЕНИЕ

За прошедшие годы была проделана значительная работа по созданию информационных систем для обеспечения доступа к информационным ресурсам РАН и интеграции этих ресурсов. К сожалению, так и не была решена задача создания единого информационного пространства РАН, основанного на современном подходе к интеграции данных и приложений, и эта проблема остается в том же положении, что и была в конце 1990-х и начале 2000 гг.

Благодарности

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований, проекты № 17-07-00214 и № 18-07-00297.

СПИСОК ЛИТЕРАТУРЫ

1. *Бездушный А.Н., Жижченко А.Б., Кулагин М.В., Серебряков В.А.* Интегрированная система информационных ресурсов РАН и технология разработки цифровых библиотек. // Программирование. 2000. №4. С. 12–20.

2. Интегрированная система информационных ресурсов: архитектура, реализация, приложения. Коллектив авторов под редакцией В.А. Серебрякова. М.: ВЦ АН, 2004. 240 с.

3. *Бездушный А.А., Бездушный А.Н., Серебряков В.А., Филиппов В.И.* Интеграция метаданных Единого Научного Информационного Пространства РАН. Вычислительный центр РАН. Москва, 2006, 238 с.

4. *Бездушный А.А., Бездушный А.Н., Нестеренко А.К., Серебряков В.А., Сысоев Т.М., Теймуразов К.Б., Филиппов В.И.* Информационная Web-система «Научный институт» на платформе ЕНИП. М.: Вычислительный центр им. А.А. Дородницына РАН, 2007. 257 с.

5. *Каленов Н.Е., Савин Г.И., Серебряков В.А., Сотников А.Н.* Принципы построения и формирования электронной библиотеки «Научное наследие России»// Программные продукты и системы. 2012. №4. С. 28–31.

6. *Атаева О.М., Кузнецов К.А., Серебряков В.А., Филиппов В.И.* Портал интеграции пространственных данных «ГеоМета». Препринт ВЦ РАН, 2010. 106 с.

7. *Атаева О.М., Серебряков В.А*. Персональная цифровая библиотека Libmeta как среда интеграции связанных открытых данных// Электронные библиотеки: перспективные методы и технологии, электронные коллекции:

Всероссийская научная конференция RCDL-2014 (Дубна, 13–16 октября 2014 г.): труды конференции / сост. Л.А. Калмыкова, М.Р. Когаловский. Дубна: ОИЯИ, 2014, С. 66–71.

ELECTRONIC LIBRARIES IN THE COMPUTING CENTER OF RUSSIAN ACADEMY OF SCIENCES – MAIN DEVELOPMENTS

V. A. Serebryakov

Dorodnicyn Computing Centre of Federal Research Centre "Computer Science and Control" of Russian Academy of Sciences, Moscow 119333. Vavilov str., 40

serebr@ultimeta.ru

Abstract

The main projects that have been implemented in the Computing Center named A.A. Dorodnitsyna of the Russian Academy of Sciences (CC RAS) for the last 20 years, that is, since 1998, are analyzed. One of the first was the implementation of the pilot project “Integrated Information Resource System (ISIR) RAS”. The successful completion of this project allowed the development of work on the integration of heterogeneous scientific information resources into the general academic scientific information system. An important stage was the project of creating the Unified Scientific Information Space (ENIP) of the RAS. This project was based on the subsystem “Scientific Institute of the Russian Academy of Sciences”, created at the CC of the Russian Academy of Sciences and the Center for Scientific Telecommunications (CNTK) of the Russian Academy of Sciences. Considering the importance of building digital libraries, in 2006 the Russian Academy of Sciences adopted the target scientific program “Creating the Central Bank “Scientific Heritage of Russia”, in accordance with which the digital library was implemented. The created GeoMeta Portal is a standardized and decentralized spatial information management environment designed to access geodatabases, map products and associated metadata from various sources, facilitating the exchange of spatial information between organizations and its sharing via the Internet. Currently, the main line of work is the LibMeta digital personal semantic library. The main task of this system is to provide the user with a unified view for the possibility of automated extraction of information of interest to him on a particular subject area.

Keywords: *subject area, scientific subject area, scientific information, scientific knowledge, generalized representation of scientific subject area, taxonomy, thesaurus, global ontology, search engines, organization of scientific knowledge, digital libraries*

REFERENCES

1. *Bezdushnyj A.N., Zizhchenko A.B., Kulagin M.V., Serebryakov V.A.* Integrirovannaya sistema informacionnyh resursov RAN i tekhnologiya razrabotki cifrovyh bibliotek // Programmirovanie. 2000. No 4. S. 12–20.

2. Integrirovannaya sistema informacionnyh resursov: arhitektura, realizaciya, prilozheniya. Kollektiv avtorov pod redakciej V.A. Serebryakova. M.: VCz AN, 2004. 240 s.

3. *Bezdushnyj A.A., Bezdushnyj A.N., Serebryakov V.A., Filippov V.I.* Integraciya metadannyh Edinogo Nauchnogo Informacionnogo Prostranstva RAN. Vychislitel'nyj centr RAN. Moskva, 2006, 238 s.

4. *Bezdushnyj A.A., Bezdushnyj A.N., Nesterenko A.K., Serebryakov V.A., Sysoev T.M., Tejmurazov K.B., Filippov V.I.* Informacionnaya Web-sistema «Nauchnyj institut» na platforme ENIP. M.: Vychislitel'nyj centr im. A.A. Dorodnicyna RAN, 2007. 257 s.

5. *Kalenov N.E., Savin G.I., Serebryakov V.A., Sotnikov A.N.* Principy postroeniya i formirovaniya ehlektronnoj biblioteki «Nauchnoe nasledie Rossii»// Programmnye produkty i sistemy. 2012. №4. S. 28–31.

6. *Ataeva O.M., Kuznecov K.A., Serebryakov V.A., Filippov V.I.* Portal integracii prostranstvennyh dannyh «GeoMeta». Preprint VC RAN, 2010. 106 s.

7. *Ataeva O.M., Serebryakov V.A.* Personal'naya cifrovaya biblioteka Libmeta kak sreda integracii svyazannyh otkrytyh dannyh// EHlektronnye biblioteki: perspektivnye metody i tekhnologii, ehlektronnye kollekcii: Vserossijskaya nauchnaya konferenciya RCDL-2014 (Dubna, 13–16 oktyabrya 2014 g.): trudy konferencii / sost. L.A. Kalmykova, M.R. Kogalovskij. Dubna: OIYAI, 2014. S. 66–71.

СВЕДЕНИЯ ОБ АВТОРЕ

СЕРЕБРЯКОВ Владимир Алексеевич, гл. н. с. ВЦ РАН, д. ф.-м. н., профессор, окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области алгоритмических языков и информационных технологий.

Vladimir Alekseevich SEREBRYAKOV, Chief Researcher of the Computing Center of the Russian Academy of Sciences, D. Sc., professor, graduated from M.V. Lomonosov Moscow State University. Specialist in the field of algorithmic languages and information technology.

e-mail: serebr@ultimeta.ru

Материал поступил в редакцию 10 декабря 2018 года