ISBN 978-5-19601-103-6

Федеральное государственное бюджетное учреждение науки
ИНСТИТУТ КОСМИЧЕСКИХ ИССЛЕДОВАНИЙ
РОССИЙСКОЙ АКАДЕМИИ НАУК
Федеральное государственное бюджетное учреждение науки
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР им. А.А. ДОРОДНИЦЫНА
РОССИЙСКОЙ АКАДЕМИИ НАУК

ИНФРАСТРУКТУРА НАУЧНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ И СИСТЕМ

Сборник избранных научных статей

Под редакцией
доктора техн. наук Е.Б. КУДАШЕВА,
доктора физ.-матем. наук В.А. СЕРЕБРЯКОВА

Том I

Москва
2014

УДК [002:004.9] (063)
ББК [73+32.973.233]я43

Инфраструктура научных информационных ресурсов и систем. Сборник избранных научных статей. Труды Четвертого Всероссийского симпозиума (С.-Петербург, 6–8 октября 2014 г.). Под ред. Е.В. Кудашева, В.А. Серебрякова. В 2-х тт. Т. 1. М.: ВЦ РАН, 2014.

Симпозиум проводится ежегодно с 2011 г. по Плану научных конференций Отделения математических наук РАН. В 2011 г. и 2012 г. Симпозиум проводился в С.-Петербурге при поддержке РФФИ, в 2013 г. – в Абхазии, г. Сухум – при поддержке Академии наук Абхазии. В 2014 г. Симпозиум проводился в С.-Петербурге при поддержке РФФИ на базе Петербургского Отделения Математического института им. В.А. Стеклова РАН – ПОМИ РАН.

Научная программа Симпозиума «**Инфраструктура научных информационных ресурсов и систем»** ориентирована на рассмотрение проблем и перспектив развития информационно-телекоммуникационных систем; методов, технологий и средств применительно к доступу, хранению и интеллектуальному анализу данных в различных областях фундаментальной науки, разработки информационных систем для научных исследований.

Основные цели Четвертого Симпозиума: методы и технологии интеграции электронных коллекций; взаимодействия информационных ресурсов и формирования электронного документного пространства научных исследований и инноваций, развитие электронных библиотек.

The symposium is held annually since 2011 on the Plan of scientific conferences Department of Mathematical Sciences of RAS. In 2011 and 2012 Symposium was held in St. Petersburg and was supported by RFBR, in 2013 – in Abkhazia, Sukhum – with the support of the Academy of Sciences of Abkhazia. In 2014, the Symposium is held in St. Petersburg on the basis of the St. Petersburg Branch of the Mathematical Institute of the Academy of Sciences – PDMI RAS and is supported by RFBR.

The scientific program of the Symposium is oriented to the infrastructure of scientific information resources and systems geared to the problems and prospects of development of information and telecommunication systems; methods, tools and technology with respect to access, storage, and data mining in various fields of basic science, development of information systems for research.

The main objectives of the Fourth Symposium: methods and techniques of integration of digital collections; interaction of information resources and the generation of the electronic document space research and innovation, the development of digital libraries.

Рецензенты: Г.Н. Заварза, К.Б. Теймуразов

Научное издание

Программный комитет
Четвертого Всероссийского Симпозиума:

Гордов Е.П., профессор (Институт мониторинга климатических и экологических систем СО РАН, Томск);

Кудашев Е.Б., акад. РИА (Институт космических исследований РАН / Мех.-матем. факультет МГУ им. М.В. Ломоносова);

Жижимов О.Л., профессор (Институт Вычислительных технологий СО РАН, Новосибирск);

Федотов А.М., член-корреспондент РАН (Институт вычислительных технологий СО РАН / Новосибирский государственный университет);

Серебряков В.А., профессор (ВЦ РАН, Москва);

Малков О.Ю. , профессор (Институт астрономии РАН, Москва);

Попов М.А., профессор (Научный центр аэрокосмических исследований Земли Институт геологических наук Национальной академии наук Украины),

Прохоров М.Е., профессор (ГАИШ МГУ им. М.В. Ломоносова);

Вязилов Е.Д., профессор (ВНИИ гидрометеорологической информации – Мировой Центр данных), Обнинск.

Председатели Программного комитета Симпозиума:
академик Российской инженерной академии ***Е.Б. Кудашев***
и профессор ***В.А. Серебряков***.

Оргкомитет Симпозиума

Председатель Оргкомитета
Серебряков В.А., профессор (ВЦ РАН, Москва);

Е.Б. Кудашев (ИКИ РАН), академик РИА;

Теймуразов К.Б. (ВЦ РАН, Москва);

Пестунов И.А. (Институт вычислительных технологий СО РАН, Новосибирский университет);

Гордов Е.П., проф. (Институт мониторинга климатических и экологических систем СО РАН, Томск);

Малков О.Ю., проф. (ИНАСАН, Москва);

Фазлиев А.З. (Институт оптики атмосферы СО РАН, Томск);

Желенкова О.П. (Специальная астрофизическая обсерватория РАН, Архиз).

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

В настоящем Сборнике представлены доклады Четвертого Всероссийского симпозиума «ИНФРАСТРУКТУРА НАУЧНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ И СИСТЕМ», организованного Вычислительным центром им. А.А. Дородницына РАН и Институтом космических исследований РАН в рамках Плана научных конференций РАН в 2014 году и поддержанного РФФИ (проект № 14-07-20321). Заседания Симпозиума проходили 6–8 октября 2014 г. в Санкт-Петербурге в Санкт-Петербургском отделении Математического института им. В.А. Стеклова РАН.

Симпозиум проводится ежегодно в С.-Петербурге при поддержке РФФИ, начиная с 2011 г., что отражает возросшее значение и особую актуальность научных проблем, связанных с созданием и исследованием современных информационных систем для научных исследований. С учетом того, что в последние годы созданы принципиально новые технологии обработки данных, которые позволяют оперативно получать информацию о различных объектах и явлениях на очень больших территориях, необходимо отметить широкий интерес к фундаментальным и прикладным научным исследованиям в области развития инфраструктуры научных информационных ресурсов, а также к использованию геопространственных данных.

В связи с лавинообразным ростом объема цифровых научных данных, подходы к длительному хранению и непрерывному доступу к научным информационным ресурсам вызывают особый интерес. На Симпозиуме традиционно обсуждаются исследова-

ния цифровых e-Infrastructures с целью формирования распределенных научных информационных ресурсов, развития взаимосвязанных каталогов и создания сети интегрированных интероперабельных баз данных.

Важным научным направлением состоявшегося Симпозиума является обсуждение результатов исследований в области создания e-Science инфраструктуры научных данных и информационных систем, разработки программных средств по созданию такой инфраструктуры, включающих в себя средства сбора, моделирования и представления данных на основе международных стандартов, исследований информационных технологий электронного правительства и управления государственными информационными системами, развития общества знаний на основе применения информационных технологий в социально-гуманитарной сфере.

В докладах на Симпозиуме отмечалось, что научные данные рассредоточены, их использование ограничено зачастую рамками того проекта, где они созданы, затруднен или невозможен поиск существующих данных и доступ к ним, не налажен обмен данными. Причина этого – отсутствие эффективной системы управления пространственными данными. Ее создание позволило бы 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.

Развитие e-Science Infrastructures должно стать основой формирующихся систем коллективной работы исследователей на ос-

нове виртуального объединения информационных и вычислительных ресурсов. В нашей стране накоплен большой опыт использования геоинформационных технологий, реализованы многочисленные геоинформационные проекты, созданы информационные системы научных ресурсов.

На Симпозиуме были заслушаны доклады из различных регионов России: это Вычислительный центр им. А.А. Дородницына РАН, Институт космических исследований РАН, Объединенный Институт высоких температур РАН, Институт вычислительных технологий СО РАН, Мировой Центр данных – Всероссийский НИИ гидрометеорологической информации, Институт оптики атмосферы СО РАН, Институт мониторинга климатических и экологических систем СО РАН, Институт астрономии РАН, Специальная Астрофизическая обсерватория РАН (Архыз), ГАИШ МГУ, Тихоокеанский Институт географии ДВО РАН, Географический факультет МГУ и другие ведущие научные центры России. Были также рассмотрены результаты работ, ведущиеся в институтах РАН и Национальной Академии наук Украины по разработке научных и прикладных аспектов создания общеакадемической инфраструктуры пространственных данных и формирования распределенной среды интегрированных информационных ресурсов.

Основным направлением работы Четвертого Симпозимума были вопросы применения современных подходов в технологии развития информационных систем к задачам информационной поддержки научных исследований. Важным и интенсивно развивающимся направлением информационной поддержки научных

исследований являются электронные библиотеки, технологии которых обеспечивают интеграцию разнородных данных. Исходя из целей ЭБ и анализа существующих систем, направленных на поддержку научных исследований, сформулированы следующие функциональные требования к модели ЭБ по научному наследию: надежное долговременное и защищенное от исчезновения хранение информации; актуальность, полнота, достоверность происхождения документов; историчность информации; географическая привязка информации; наличие большого числа словарей-классификаторов (справочников), для обеспечения идентификации и классификации ресурсов; поддержка неоднородных и слабо структурированных информационных ресурсов; поддержка взаимосвязей информационных ресурсов; предоставление информации пользователю в виде, выбранном пользователем; наличие интеллектуальных служб обслуживания запросов пользователя; наличие программных интерфейсов для поддержки аналитической работы пользователя с помощью программных приложений; поддержка требований интероперабельности как на программном, так и на семантическом уровне; поддержка работы с внешними источниками.

Редакторы и авторы Сборника выражают свою благодарность и признательность Российскому фонду фундаментальных исследований за поддержку исследований в области развития инфраструктуры научных информационных ресурсов и предоставленную возможность издания их результатов.

Е.Б. Кудашев, В.А. Серебряков

СИСТЕМА УПРАВЛЕНИЯ ЭЛЕКТРОННЫМИ БИБЛИОТЕКАМИ В ИРИС СО РАН[1]

Ю.И. Шокин, А.М. Федотов, О.Л. Жижимов, О. А. Федотова

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет,
Государственная публичная научная библиотека СО РАН

Начиная с 1998 года в ИВТ СО РАН, ведутся работы по созданию ИРИС СО РАН (Интегрированной Распределённой Информационной Системы), организованной в виде электронной библиотеки. Статья посвящена истории создания и описанию технологических решений, применяемых при создании системы. Описываются архитектура информационной системы и принципы интеграции с внешними источниками, правила представления и преобразования метаданных, а также описана работа со словарями, которые используются для систематизации и классификации информационных ресурсов, и моделированию связей между ними.

Ключевые слова: ИРИС, информационная система, электронная библиотека, словарь-справочник, классификация информационных ресурсов, цифровой репозиторий, информационно-поисковый тезаурус, ключевые термины, протокол OAI-PMH, метаданные.

1. Введение

В начале 1998 г. в Сибирском отделении РАН была сформирована целевая программаразвития информационных ресурсов Отделения под общим названием “Электронная библиотека Сибирского отделения РАН”. Для решения проблемы информационной обеспеченности сотрудников Отделения было принято решение о создании собственной универсальной Интегрированной Распре-

[1]Работа выполнена при частичной поддержке РФФИ (проекты 12-07-00472, 13-07-00258), президентской программы «Ведущие научные школы РФ» (грант НШ–5006.2014.9).

деленной Информационной Системы (ИРИС) Отделения [1,2], содержащей полнофункциональную систему об интеллектуальном потенциале Отделения (информационную систему об институтах, сотрудниках, достижения и др. аспектах, связанных с работой Отделения) и систему электронной поддержки сбора и накопления научной информации (электронных атласов, электронных коллекций, баз данных и т.п.).

Основные направления программы связаны с формированием собственных электронных ресурсов по основным отраслям наук (математика, науки о земле, химия, биология, археология и др.), созданию и поддержке электронных коллекций и электронных публикаций, организации удобных систем доступа к библиотечным и библиографическим базам данных и базам данных Институтов Отделения [3]. Результаты работ должны обеспечить:

- Единую распределенную информационную среду Отделения.
- Информационную поддержку исследований по фундаментальным и прикладным направлениям.
- Поддержку профессионально-ориентированных систем подготовки и обмена научных документов.
- Поддержку профессионально-ориентированных систем доступа и интерфейсов с хранилищами данных.
- Коллективное использование приобретаемой электронной литературы, каталогов, баз данных и библиографических изданий.
- Поддержку электронных версий научных журналов, издаваемых институтами Отделения.
- Поддержку принятия и реализации организационных и управленческих решений в Отделении.

Основные ресурсы, созданные тогда не потеряли своей актуальности до сих пор. В первую очередь здесь стоит упомянуть

электронный атлас «Биоразнообразие животного и растительного мира Сибири» [4], систему поддержки проведения конференций [5], виртуальной музей СО РАН [6], база данных организаций и сотрудников СО РАН [7].В результате работ по программе сложилось понимание того, что информационная система для поддержки научных исследований ИРИС должна основываться на использовании концепции электронных (цифровых) библиотек [8-11]. В рамках нашего подхода цифровые библиотеки рассматриваются как отдельная конкретная технология работы с цифровой информацией, образующая новый класс информационных систем (ИС), предназначенных для управления информационными ресурсами [12, 13].

Под термином *электронная библиотека* (ЭБ) в данной работе будем понимать систему управления структурированными каталогизированными коллекциями разнородных электронных (цифровых) объектов (ресурсов). ЭБ не только обеспечивает многосторонний поиск и навигацию по каталогам (в отличие от печатных изданий, микрофильмов и других носителей), но и предоставляет пользователю непосредственно найденный ресурс (публикацию, документ, фотографию, описание факта и др.), а также дополнительные сведения о нем, например, географическую привязку, информацию об авторах, информацию о фактах, библиографию, перечень организаций и т. д.

Концепция ИРИС

На сегодняшний день наиболее эффективным способом решения проблем организации доступа к распределенным информационным ресурсам является организация информации о них в информационные системы, обличенных в форму электронных

библиотек[2]. В работах[14,15] были сформулированы основные принципы реализации ИРИС, основанные использовании идеи электронных библиотек.

Отметим, что ЭБ – явление относительно новое, но уже достаточно популярное [16]. Тем не менее ЭБ сегодня следует рассматривать как множество слабосвязанных сущностей, объединяемых на первый взгляд только общим названием [13, 17]. Современное название «электронная библиотека» – это не только и не столько дань моде, сколько попытка охарактеризовать новый феномен – возникновение принципиально нового класса систем, призванных аккумулировать и распространять информацию в электронной форме [18]. А большой интерес к самим системам данного класса объясняется потребностями общества и ростом возможностей по их удовлетворению [19].

В связи с этим можно сформулировать основные цели, стоящие перед ЭБ (системами управления информационными ресурсами) [20]:

- управление информационными ресурсами;
- обеспечение и управление доступом к информации;
- долговременное хранение информации;
- сохранение научного и культурного наследия;
- поддержка аналитической работы с информацией;
- повышение эффективности научных исследований и обучения.

В существующих разработках ЭБ, как правило, поиск и доступ к информации обеспечиваются только посредством визуальных

[2] По-простому, понятие электронной библиотеки заключается в том, что любой ресурс должен быть *стандартным* образом каталогизирован, снабжен *метаданными,* правилами доступа и уникальным идентификатором.

графических интерфейсов. Это хорошо для пользователя-человека, но не годится для пользователя-системы. Для реализации функций поиска вне графических интерфейсов требуется поддержка специальных сетевых сервисов и языков запросов. В идеальном случае все ИС должны поддерживать единый поисковый профиль и единый язык запросов [12].

Однако в общем случае под словосочетанием «электронная библиотека» могут фигурировать совершенно различные объекты, такие как архивы цифрового контента и наборы программного обеспечения для управления этим контентом. ЭБ может называться система сетевых сервисов, предоставляющих доступ к цифровому контенту и объединенных единой системой управления этим доступом [20]. Такое определение ЭБ полностью соответствует определению традиционной библиотеки как организации в системе, например, Министерства культуры [12].

Ввиду того, что информация в ИС отображает некоторые сущности реального мира (физические объекты: предметы, процессы, явления, персоны, публикации, документы, алгоритмы, программы, файлы, факты, ключевые термины и т. д.), следует рассматривать ИС как множество информационных объектов – наборов данных, представляющих (описывающих) эти сущности в ИС. Отметим, что разработка модели ЭБ должна использовать онтологические описания и концептуальные модели, обобщающие накопленный опыт в сфере создания и использования ЭБ [20]. Неплохой обзор существующих концептуальных моделей ЭБ приведен в [22].

Онтологическая модель ЭБ ИРИС основана на концептуальных моделях ЭБRMOAIS[3] [23] и DELOSDLRM[4] [24].

[3]Open Archival Information System
[4]Digital Library Reference Model

В соответствии с концептуальной моделью DELOS Информационный Ресурс (ИР) – это абстрактное понятие, выражаемое экземплярами одной из своих специализаций. В частности, экземплярами понятия ИР являются экземпляры информационного объекта любого типа (например, документы, базы данных, коллекции, функции и т.п.). Каждый ресурс в соответствии с моделью DELOS:

- имеет идентификатор;
- организован в соответствии с описанием ресурса. Ресурс может быть сложным и структурированным, поскольку он, в свою очередь, может состоять из меньших ресурсов и иметь связи с другими ресурсами;
- может регулироваться функциями, управляющими его жизненным циклом;
- выражается через информационный объект;
- должен быть описан метаданными, а также может быть описан или дополнен дополнительными метаданными и аннотациями.

В основе реализации Системы Управления Электронными Библиотеками (СУЭБ) в ИРИС лежит метамодель, исходящая из того, что каждый информационный ресурс характеризуется набором присущих ему атрибутов, и методов, характеризующих его свойства и связи с другими ресурсами. Эффективным средством описания информационных объектов являются метаданные – данные, являющиеся неотъемлемой частью информационного объекта и описывающие реальный объект или группу объектов.

Каждый информационный объект в ИС состоит из (см. рис. 1):

- Информационного содержания объекта (первичный информационный объект: например, изображение, полный текст и т.д.) – объект, который может использоваться самостоятельно;

- объект метаданные – объект, главная цель которого состоит в том, чтобы дать информацию об ИР (как правило о первичном информационном объекте);
- объект аннотация – объект, главная цель которого состоит в том, чтобы аннотировать ИР или его часть. Примеры таких аннотаций включают примечания, структурированные комментарии и связи. Объекты аннотации помогают интерпретировать ИР, содержат либо поддержку, либо детальные объяснения, либо информацию о том, как можно использовать ИР.

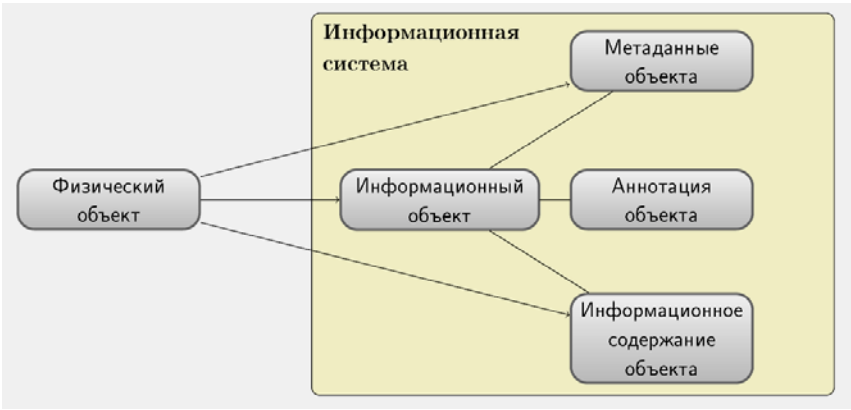


Рис.1. Структура сущностей ИС.

Документы

В информационном пространстве все информационные ресурсы: события, факты, программы и любые другие сущности реального или виртуального мира существуют только в форме некоторых информационных объектов[25, 26], концептуальная модель которых была описана выше.

Бельгийский ученый П. Отле – пионер и основоположник науки «информатика» – в своем трактате о Документации [27] опре-

деляет расширение понятия (термина) Документ[5]:«материальный объект, содержащий информацию, специально предназначенный для ее передачи в пространстве и времени» – который трактуется как основной «объект», с которым оперирует любая информационная система [28]. Таким образом, Документ – это информационный объект, представлявший структурированное описание реальной сущности (объекта, субъекта, факта или понятия), совокупность которых составляют информационное наполнение системы.

В ИС Документ – это целостный информационный объект, представленный в цифровом виде, имеющий некоторый стандартный набор атрибутов и функций и допускающий однозначную идентификацию. Документом может описывать статью из журнала, сам журнал, персону, оцифрованное изображение, экспериментальные данные, программа или вычислительный алгоритм, база данных, фрагмент базы данных и т.п. [25].Однако документ не обязан содержать полный текст статьи или персону, но в качестве атрибута он может содержать указатель на хранилище данных (репозиторий), где полный текст хранится. Кратко требования к документу можно охарактеризовать так:

- документ имеет уникальный идентификатор;
- документ имеет структурированное описание (метаданные);
- документ имеет набор атрибутов (*свойств*) и методов (функций);
- взаимодействие с документом (например, работа с атрибутами) происходит через набор *методов.*

Ключевым моментом в работе с документами является использование метаданных для формирования структуры, схемы

[5]отлат. Documentum –доказательство

данных и свойств документов – информационных элементов системы и ведения каталога системы. Любая ЭБ опирается на метаинформацию, содержащую онтологию, описывающую принципы организации информации. Онтология, описывающая конкретную предметную область, конкретизируется в схеме данных (атрибутов), описывающих информацию в метаданных[26].

Электронные библиотеки

Как мы уже отмечали, что электронные библиотеки – это распределённые каталогизированные информационные системы, позволяющие хранить, обрабатывать, распространять, анализировать, а также организовывать поиск в разнообразных коллекциях электронных (цифровых) документов. Основная задача, решаемая электронными библиотеками – это управление информационными ресурсами и «интеграция информационных ресурсов (включая поддержку унифицированного доступа к ним), а также эффективная навигация в них» [25].

Под интеграцией информационных ресурсов понимается их объединение с целью использования (с помощью удобных и унифицированных пользовательских интерфейсов) разнородной информации с сохранением ее свойств, особенностей представления и пользовательских возможностей манипулирования с ней. При этом объединение ресурсов не обязательно должно осуществляться физически, оно может быть виртуальным, главное – оно должно обеспечивать пользователю восприятие доступной информации как единого информационного пространства. В частности, такие системы обеспечивают работу с гетерогенными наборами и базами данных или системами баз данных, обеспечивая пользователю эффективность информационных поисков независимо от особенностей конкретных систем хранения ресурсов, к которым осуществляется доступ [26].

Под эффективной навигацией в информационной системе понимается возможность для пользователя находить интересующую его информацию с наибольшей полнотой и точностью при наименьших затратах усилий во всем доступном информационном пространстве. При таком подходе хорошо известные информационно-поисковые системы, используемые в информационных системах и базах данных, являются частными случаями навигационных средств [25, 26].

В настоящее время существуют достаточно мощные ИС, удовлетворяющие в той или иной степени потребности научных работников в информации, однако основной недостаток большинства систем – ограниченность интеграции ресурсов как внутри каждой из них, так и с внешними системами. Основу разработки ЭБ составляют стандарты и международные рекомендации, формирующие профиль ЭБ, под которым понимается один или набор нескольких базовых нормативно-технических документов (стандартов и спецификаций), ориентированных на решение определенной задачи (реализацию заданной функции либо группы функций приложения или среды), с указанием, если нужно, выбранных классов, подмножеств, опций базовых стандартов, необходимых для выполнения конкретной функции [29]. Наиболее важным является профиль метаданных информации, циркулирующей в системе. Выбор профиля метаданных должен основываться на выполнении следующих требований [20, 25, 30,31]:

- включать описания основных типов информации, требующейся для поддержки научно-образовательной деятельности;
- быть открытым, т. е. обеспечивать доступ к информации в соответствии с ее описанием (метаданными);
- быть расширяемым, т. е. обеспечивать возможность детализации описаний;

- обеспечивать возможность интеграции информации;
- обеспечивать возможность уникальной идентификации информации;
- обеспечивать отбор, систематизацию и классификацию информации;
- обеспечивать возможности размещения и поиска информации в распределенной среде;
- быть ориентированным на современные технологии описания и использования информации;
- обеспечивать возможность интероперабельности с другими системами.

Отметим, что серьезной проблемой является идентификация информационных ресурсов [32], позволяющая получать библиографические сведения, а также устанавливать связи определенного ресурса с другими фактами и объектами.

При работе с цифровыми объектами уже выработан определенный набор стереотипов, отсутствие которых вызывает дискомфорт [12]. Одним из элементов этого набора является наличие взаимных ссылок между цифровыми объектами, например, в виде гиперсвязей в пользовательских графических интерфейсах просмотра информации. Реализация взаимных ссылок в цифровых документах не представляет большой сложности, однако здесь прослеживается своя специфика. Во-первых, электронный объект с реализованными связями уже не совсем соответствует своему печатному оригиналу. Во-вторых, внедренные в объект связи должны быть гарантированно актуальными. Так появляется требование обеспечения ссылочной целостности данных. Это очень жесткое требование, которое тяжело соблюсти даже в хорошо формализованных системах управления БД. Приемлемым решением здесь может быть замена жестких гиперссылок динамиче-

скими связями между документами на уровне системы управления. Результат –новый цифровой объект как самосогласованное хранилище цифрового контента или БД цифровых объектов.

Существует достаточно много технологических разработок ИС для ЭБ, так или иначе ориентированных на поддержку научных исследований, например, euroCRIS[6], eLibrary[7], Информика[8], athNET[9]. В большей степени удовлетворяет информационным потребностям научно-образовательного сообщества в информации система ИСИР (ЕНИП) РАН [30, 33].

Наиболее важным выводом из вышесказанного является следующий: информационная модель ЭБ должна быть многоуровневой и состоять как минимум из нескольких компонент [21, 25, 26]: хранилища данных (репозиторий), сервера метаданных, сервера приложений (диспетчера), словарей-справочников (см. рис.2).

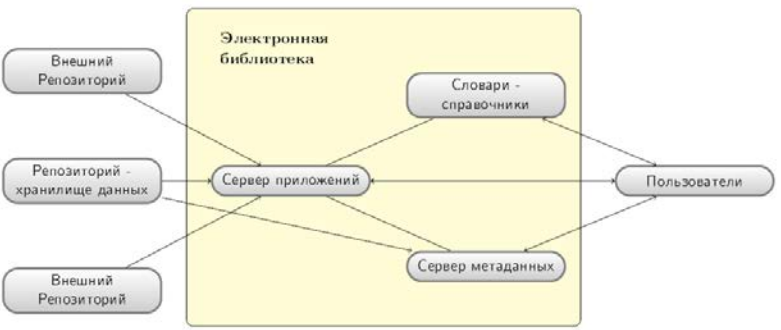


Рис.2. Архитектура ЭБ.

[6]http://www.eurocris.org/
[7]http://elibrary.ru/
[8]http://www.informika.ru/
[9]http://www.mathnet.ru/

Цифровые репозитории

Для организации системы долговременного хранения информационных ресурсов (репозиторие в цифровых объектов) международной организацией по стандартизации (ISO) предложен стандартISO-14721 (OpenArchiveInformationSystem– OAIS[10]) [23]. Эталонная модель для стандарта OAIS – это концептуальная модель, основанная на расширенной схеме данных DublinCore [34]. Эта модель была использована многими организациями для разработки наборов метаданных и создания крупных хранилищ цифровых объектов.

На основе данной модели создана концепция «институционального репозитория» как системы долговременного хранения, накопления информации и обеспечения надежного доступа к цифровым объектам, представляющим собой результат интеллектуальной деятельности научного или образовательного учреждения. К основным особенностям институционального репозитория относятся:

– обеспечение разграниченного доступа к разнородным цифровым объектам (публикациям, изображениям и т. д.);

– организация доступа к информационным ресурсам для мирового сообщества (в том числе с помощью полнотекстового индексирования мировыми поисковыми системами);

– унифицированный доступ к метаданным по стандартным протоколам (поддержка интероперабельности);

– возможность организации единой точки доступа к информационным ресурсам;

– сохранение других информационных ресурсов, в том числе неопубликованных, таких как диссертации, препринты и технические отчеты, программное обеспечение, мультимедиа и т. д.

[10]Открытая Архивная Информационная Система.

Согласно данным сайта OpenDOAR[11], большинство институциональных репозиториев основано на свободном программном обеспечении и построено в рамках модели OAIS на базетехнологий открытых систем. В настоящий момент в мире насчитывается более десятка систем поддержки институциональных репозиториев, наиболее популярные из них DSpace [35] (свыше 41% установок), E-Prints [36], Fedora [37] (сравнительную характеристику этих систем и описание используемых в них информационных моделей можно найти в [38]). Процесс интеграции репозитория в среду ЭБ для этих систем отличается лишьнесущественными деталями и основан на модели агрегирования и распространения метаданных. Применение этой модели закреплено в протоколе OAI Protocol for Metadata Harvesting (далее OAI или OAI-PMH) [39], который поддерживается большинством систем, предназначенных для хранения информационных ресурсов.

В качестве репозитория для ИС по научному наследию нами выбрана система DSpace. Выбор обусловлен тем, что она является самой популярной в мире и уже эксплуатируется в СО РАН[12] (а так же в ряде других институтов и университетов России) на протяжении десяти лет.

Кроме того, система DSpace имеет ряд привлекательных особенностей. Для базовой организации данных зафиксирована определенная модель данных (DIM – внутренний формат данных DSpace), основанная на схеме Dublin Core и ее расширениях. При некотором напряжении при помощи этой схемы можно отобразить основные элементы всех используемых в настоящий момент форматов. Система с помощью фильтров метаданных, которые используются для преобразования метаданных из внутренней схемы в схемы, пригодные для обмена метаданными с внешними

[11] The Directory of Open Access Repositories – http://www.opendoar.org/
[12]http://elib.nsc.ru:8080/jspui/

системами на основе XSLT[13] трансформаций, позволяет конвертировать и индексировать метаданные в разнообразных форматах (МЕКОФ[14], MODS[15], METS[16], QDC[17], MARC[18] и др.). Система хранит информацию о пользователях, поддерживает авторизацию и разграничивает доступ к содержимому репозитория по группам, сетевым адресам и на основе протокола LDAP [40], что при создании ИСдает возможность использовать уже существующую систем уаутентификации пользователей (а не разрабатывать свою собственную) и достаточно легко дифференцировать публичные и служебные ресурсы, оставляя при этом свободный доступ к метаданным. Провайдеры данных для протоколов OAI-PMH [39], Z39.50 [41, 42] и SRW/SRU [43], стандартных для библиотечного сообщества, позволяют разрабатывать программный интерфейс для взаимодействия ИС поддержки ЭБ с хранилищем данных, поостренном на основе DSpace.

Отметим, что для обеспечения работы ЭБ в рамках сформулированных выше функциональных требований, недостаточно метаданных, используемых для долговременного сохранения цифровых объектов (нужны дополнительные метаданные для организации связей между документами, способов представления документов и т. д., что подробнее изложено в следующем разделе), поэтому необходимо, чтобы метаданные хранились в системе

---

[13] http://www.w3.org/TR/xslt/

[14] Этот формат наиболее привлекателен в России, т. к. подержан ГОСТом [35].

[15] Metadata Object Description Schema (MODS) – http://www.loc.gov/standards/mods/

[16] Metadata Encoding and Transmission Standard (METS) – http://www.loc.gov/METS

[17] Qualified Dublin Core (QDC) [29].

[18] Machine-Readable Cataloging (MARC) – http://www.loc.gov/marc/

поддержки ЭБ, независимой от той, которая использовалась при их создании.

Для синхронизации метаданных информационных ресурсов между ИС ЭБ и хранилищем данных используется сервис, основанный на использовании протоколов OAI-PMH [39], SRW/SRU [43] и сервера приложений на основе ZooPARK-ZS [44] (в зависимости от типа хранилища данных). В задачи этого сервиса входит извлечение метаданных из репозитория, конвертирование (при необходимости) и передача их серверу метаданных. Под конвертированием здесь понимается как преобразование метаданных и схем метаданных (например, QDC [34] в ГОСТ 7.19 [45] или RUSMARC [46]), так и преобразование форматов (например, в XML [47] или ISO-2709 [48]).

Следует подчеркнуть, что при взаимодействии с удаленными системами, обмен метаданными должен происходить согласованно, с использованием форматов обмена.

Выбор метаданных для ЭБ

Для поддержки сложных функций поиска и классификации необходимо обеспечить возможность поиска по атрибутам, полнотекстового поиска, а также просмотра ресурсов по категориям и словарям-классификаторам.

В существующих ИС информационные ресурсы разрознены, недостаточно систематизированы и структурированы. В ходе создания их описаний недостаточное внимание уделяется вопросам интероперабельности: слабо применяются соглашения и рекомендации по стандартизации представления документов и

средства интеграции разнородных информационных ресурсов. Под интероперабельностью ИС понимается степень ее способности взаимодействовать с другими ИС, в том числе и с человеком. Но если при взаимодействии с человеком (как с ИС) основная нагрузка на достижениевзаимопонимания ложится на последнего, способного обработать даже плохо организованную информацию, то для обеспечения эффективного взаимодействия между собственно ИС требуются специальные технологические методы и общие соглашения. Это влечет за собой необходимость соблюдения соответствия всех схем данных, интерфейсов и протоколов международным стандартам и рекомендациям [12, 20].

В работах [20, 21] был определен профиль ЭБ как необходимый набор стандартов и компонентов информационной системы, ориентированной на научные исследования.

В настоящее время существует большое количество систем метаданных, предназначенных для описания различных классов информационных объектов. Использование систем метаданных (схем данных) пока еще недостаточно формализовано. ИС, ориентированные на одинаковые классы информационных объектов используют различные, зачастую оригинальные системы метаданных и форматы метаописаний, а также разные подходы к решению прикладных задач. Устранением подобных несоответствий занимаются многие организации по всему миру, например, такие как как W3C[19], DCMI[29], OCLC[20], IFLA[21], IETF[22], ISO[23].

[19] World Wide Web Consortium (W3C) – http://www.w3.org/
[20] Online Computer Library Center – http://oclc.org/
[21] International Federation of Library Associations – http://www.ifla.org/

Метаданные необходимы для решения следующих задач:

- предоставление сведений об объекте для получения представления о его содержании, структуре, способах использования и т. д.;
- сбор и систематизация информации об объектах описания;
- выбор из множества объектов определенного подмножества по формальным признакам и сопоставление объектов по формальным признакам;
- внутрисистемные технологические задачи, связанные с обеспечением подготовки объектов, размещением объектов в информационном фонде и т. д.;
- внешние технологические задачи, связанные, прежде всего, с обменом данными с внешними информационными системами.

Основу содержания ЭБ в ИРИС составляют документы (информационные объекты), представляющие основные типы сущностей:

субъекты: персоны, организации и т. д.;

объекты (единицы хранения): публикация, документ, факт, научный результат, мероприятие, фотография и др.;

отношения: понятие, ключевой термин, событие, время, место и т.п.

В отличие от общепринятых ЭБ указание на субъекты дается с помощью ссылки на экземпляр сущности «субъект», что позволяет корректно решать задачу идентификации объектов.

Используемый профиль определяет список элементов данных (полей), необходимых для создания записи соответствующего

[22] Internet Engineering Task Force (IETF) – http://www.ietf.org/
[23] International Organization for Standardization – http://www.iso.org/

типа, и раскрывает содержание элементов данных. Для эффективной работы сервера приложений используется набор словарей-классификаторов, содержащих как классификационные признаки, так и наборы ключевых терминов (с отношениями порядка), по которым производится систематизация и классификация материала.

Для формирования метаданных применяются несколько стандартов, являющихся расширениями рекомендаций Dublin Core[24] и Qualified Dublin Core (QDC). Для документов нами была расширена стандартная схема метаданных QDC полями, включающими основные требования государственного стандарта МЕКОФ [45].

Словари (ключевые признаки, ключевые термины) – это особый вид метаданных, отражающих наиболее существенные свойства объекта и имеющие наиболее важное значение с точки зрения ИС. Специфика словарей определятся терминологией конкретной предметной области, которой посвящена ЭБ. Необходимо рассматривать различные типы ключевых терминов (ключевые термины в стандартном понимании; ключевые термины, описывающие персону; ключевые термины, описывающие организацию; ключевые термины, описывающие временные периоды; ключевые термины, описывающие географические понятия), а также тематические словари-классификаторы, тезаурусы, описания предметной области данной научной школы, и классификаторы документов в соответствии с МЕКОФ.

Существует ряд российских (например, УДК[25], ГРНТИ[26]) и международных (например, MSC-2000[27], ORTELIUS[28]) словарей

[24] http://www.dublincore.org/

[25] УДК – Универсальная десятичная классификация.

для классификации научных данных. Однако в целом эти словари содержат только общенаучную информацию и не подходят (хотя использовать их все равно нужно) для систематизации материалов.

Метаданные существенным образом зависят от природы и структуры объектов реального мира, от способа представления их в виде информационных объектов и от специфики ИС. Учитывая это, необходимо классифицировать описываемые объекты. Совокупность правил, достаточная для формирования метаданных в определенном классе ИСи (или) для решения определенного рядазадач над информационными объектами, представляет собой систему метаданных.

Практическая реализация

Рассмотренная модель ЭБ реализована в виде СистемыУправления ЭБ (СУЭБ ИРИС), созданной и эксплуатируемой в ИВТСОРАН с 2004 года [25]. СУЭБ ИРИС оперирует электронными коллекциями документов. Электронная коллекция – это набор документов, объединённых по смысловому признаку и имеющих одинаковую структуру (схему данных) [26]. СУЭБ позволяет работать с двумя видами коллекций: каталогами и тезаурусами. Принципиальное отличие каталогов от тезаурусов состоит в том, что в первыхможно организовывать иерархические зависимости (родитель-потомок, часть-целое и т.п.) между записями. Коллекции-каталоги предназначены для хранения и обработки метадан-

[26] ГРНТИ – Государственный рубрикатор научно-технической информации.

[27] MSC-2000 – Математический классификатор – http://www.ams.org/msc/msc.html

[28] The «Ortelius Thesaurus on Higher Education» – http://cordis.europa.eu/cerif/src/sum_concl.htm

ных о документах различной природы: публикации, ключевые термины, персоны, организации, фотографии и т. д. Коллекции-тезаурусы предназначены для работы со словарями классификаторами.

Сервер метаданных СУЭБ содержит служебную коллекцию Основной каталог метаданных, которая содержит документы, описывающие все метаданные, которые можно использовать в системе. Документы Основного каталога содержат описания схемы метаданныхQDC, расширенной метаданным и для соответствия МЕКОФ и описания служебных метаданных, описывающими структуру объектов, пользовательские интерфейсы, ассоциативные связи между документами, права доступа к документам и т. д.(при желании он может быть расширен новыми метаданными).

Априори каждая коллекция (в зависимости от вида) имеет минимальный обязательный набор метаданных. Администратор коллекции имеет возможность доопределить схему метаданных коллекции, исходя из имеющихся метаданных из основного каталога.

В СУЭБ представленодва вида ассоциативных связей между документами (записями): жесткие и мягкие. Жесткие связи реализованы средствами СУБД путем ссылок на первичные ключи записи. К сожалению, такой тип связи не защищен от нарушения целостности (в случае неправильного изменения или удаления записи). Мягкие связи реализуются через процедуру поиска соответствий. Такой способ установления связей защищенот любых нарушений целостности БД и достаточно удобен пользователям, поскольку для указания на необходимость связи используются наглядные мнемонические определения. Соответствия устанавливаются следующими способами:

- Ссылка на идентификатор записи – уникальный, в пределах одной коллекции, текстовый код, формируемый в рамках конкретной коллекции по определенным правилам. Например, для коллекции, содержащей описания персон, идентификатор формируется (на русском языке) последовательно из фамилии – инициалов – года рождения. Отметим, что за десять лет эксплуатации СУЭБ не было зафиксировано ни одного конфликта при формировании идентификаторов.
- Ссылка на ключевой термин – особый вид метаданных, выбираемый из словаря ключевых терминов, по существу представляющий собой тезаурус предметной области коллекции. Ссылка определяет запись, в которой данный ключевой термин присутствует в метаданных.

Для внутреннего долговременного хранения цифровых объектов используется репозиторий DSpace [35]. Стандартная схема метаданных DSpace, основанная на схеме DCMI [34], расширена полями, отвечающими основным требованиям МЕКОФ[45]. Для поддержки процесса наполнения полнотекстовых БД созданные профили метаданных зарегистрированы в системе DSpace, и в соответствии с ними настроены рабочие процессы, а также пользовательский интерфейс системы.

С целью организации обмена метаданными между DSpace и сервером метаданных (а также с другими системами с расширенным профилем) создан специальный сервис, выполняющий преобразование метаданных из внутренней схемы DSpace в другие схемы метаданных, в том числе и в схему DCMI, с использованием квалификаторов (QDC[29]), а также в схему МЕКОФ (представление ISO2709 или XML). Реализован OAI-PMH сервис, который

[29] Qualified Dublin Core (QDC) – http://www.dublincore.org/documents/dcmi-terms/

в пакетном режиме периодически, в соответствии с расписанием, проводит синхронизацию метаданных репозитория и сервера метаданных. Для заполнения основного каталога метаданных в соответствии с созданными схемами метаданных используются контролируемые словари из справочного блока сопровождения. Для обеспечения интероперабельности данных DSpace также задействован сервер приложений на основе ZooPARK-ZS [44], реализующий доступ к метаданным системы по протоколам Z39.50 [41, 42] и SRW/SRU [43].

Разработанная модель может быть использована как типовая модель системы для работы с документами, связанными с научно-образовательной деятельностью, поскольку решает основные задачи, предъявляемые к этим системам: обеспечение системы надежного долговременного хранения цифровых (электронных) документов с сохранением всех смысловых и функциональных характеристик исходных документов; обеспечение «прозрачного» поиска и доступа пользователей к документам, как для ознакомления, так и для анализа содержащихся в них фактов; организация сбора информации по удаленным цифровым репозиториям, поддерживающим протоколы OAI-PMH, SRW/SRU, Z39.50.

Рассмотренная модель ЭБ реализована на примере научной школы Алексея Андреевича Ляпунова – основателя теоретического программирования и российской кибернетики, в виде ЭБ по моделям динамики биосферы, а также в виде ЭБ учебных пособий[30] по курсам «Современные проблемы информатики и вычислительной техники», «Вычислительные системы», «Информатика» и «Экология» и др.

[30] http://fedotov.nsu.ru/lecture.php

Литература

1. Шокин Ю.И., Федотов А.М. Распределенные информационные системы // Вычислительные технологии. 1998. Т.3. № 5. С.79-93.
2. Шокин Ю.И., Федотов А.М. Электронная библиотека Сибирского отделения РАН // Электронные библиотеки: рос. науч. электронный журн. – 1999. Т. 2. – Выпуск 4. [Электронный ресурс]. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/1999/part4/fedotov (дата обращения: 04.09.2014).
3. Шокин Ю.И., Федотов А.М., Жижимов О.Л., Мазов Н.А. Интегрированная распределенная информационная система (ИРИС) Сибирского отделения РАН // В сб.: Материалы выездного заседание научно-координационного совета по целевой программе "Информационно-телекоммуникационные ресурсы СО РАН", Иркутск, 29-30 августа 2002: Ин-т географии СО РАН, 2003, с. 139-149.
4. Федотов А.М., Артемов И.А., Ермаков Н.Б., Красников А.А., Потемкин О.Н., Рябко Б.Я., Федотов А.А., Хорев А.Г. Электронный атлас «Биоразнообразие растительного мира Сибири» // Вычислительные технологии. - 1998. - Т.3. - № 5. - С.68-78.
5. Федотов А.М., Гуськов А.Е., Молородов Ю.И.Информационная система поддержки проведения конференций СО РАН // В сб.: Материалы выездного заседание научно-координационного совета по целевой программе "Информационно-телекоммуникационные ресурсы СО РАН", Иркутск, 29–30 августа 2002: Ин-т географии СО РАН, 2003, с. 91110.
6. Шокин Ю.И., Ламин В.А., Федотов А.М. и др. Виртуальный музей Науки и Техники СО РАН// В сб.: Материалы выездного заседание научно-координационного совета по целевой программе "Информационно-телекоммуникационные ресурсы СО РАН", Иркутск, 29-30 августа 2002: Ин-т географии СО РАН, 2003, с. 118-125.
7. Леонова Ю.В., Клименко О.А., Федотов А.М. Информационная система «База данных организаций и сотрудников СО РАН». – Новосибирск: РИЦ Прайс-Курьер. - 2005. – 55 с.

8. Федотов А.М., Шокин Ю.И. Электронная библиотека Сибирского отделения РАН // Информационное общество. – 2000. – № 2. – С.22-31.
9. Шокин Ю.И., Федотов А.М. Библиотека, работающая круглосуточно // ЭКО. - 2000. - № 6. - С.163-172.
10. Шокин Ю.И., Федотов А.М. Информационная система Сибирского отделения РАН // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Сб. докладов. Протвино, 2000. - С.6-15.
11. Шокин Ю.И., Федотов А.М. Электронная библиотека Сибирского отделения РАН // Информационно-библиотечное обеспечение науки. Проблемы интеграции информационных ресурсов: Сб. науч. тр. – 2000. – М. – С.118-128.
12. Жижимов О. Л., Мазов Н.А., Федотов А.М. Некоторые заметки об эволюции цифровых репозиториев традиционных библиотек к полнофункциональным электронным библиотекам // Вестник Владивостокского государственного университета экономики и сервиса. Территория новых возможностей. 2010. Т. 7. № 3. С. 55-63.
13. Антопольский А.Б., Вигурский К.В. Концепция электронных библиотек [Электронный ресурс]. Электронные библиотеки: рос. науч. электронный журн. Т. 2. Вып. 2. 1999. URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/1999/part2/antopol (дата обращения: 04.05.2013).
14. Федотов А.М. Концептуальные подходы к построению распределенных систем // Труды Международной конференции по вычислительной математике (МКВМ-2004): Рабочие совещания. Новосибирск: ИВМиМГ СО РАН, 2004. - С.132-143.
15. Шокин Ю.И., Федотов А.М. Поддержка и развитие распределенных информационно-вычислительных ресурсов в СО РАН // Вычислительные технологии. (Совместный выпуск). Вестн. КазНУ им. аль-Фараби. Серия: Математика, механика, информатика. Ч. 4. - 2004. - Т.42. - Ч.4. - № 3. - С.324-334.

16. Земсков А.И., Шрайберг Я.Л. Электронные библиотеки: учеб. пособие для студентов ун-тов и вузов культуры и искусств и др. учеб. заведений. 3-е изд. М.: ГПНТБ России. 2004.
17. Воройский Ф.С. Электронные и традиционные библиотеки – суть не одно и то же [Электронный ресурс]. Электронные библиотеки: рос. науч. электронный журн. Т.6. Вып.5. 2003. URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part5/voroisky (дата обращения: 04.05.2010).
18. Акимов С.И., Елизаров А.М., Ершова Т.В., Когаловский М.Р., Федоров А.О., Хохлов Ю.Е. Научно-методическая поддержка разработки научных электронных библиотек [Электронный ресурс]. Электронные библиотеки: рос. науч. электронный журн. 2005. Т. 8. № 1. URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2005/part1/AEEKFH.
19. Вигурский К. В. Что такое электронная библиотека? Доклад на конференции "Информационные технологии в образовании - 2005". [Электронный ресурс]. http://rd.feb-web.ru/library.htm.
20. Федотов А.М., Барахнин В.Б., Жижимов О.Л., Федотова О.А. Технология создания корпоративных информационных систем учета трудов научных работников // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2011. Т. 9. № 2. С. 31-41.
21. Жижимов О.Л., Федотов А.М., Федотова О.А. Построение типовой модели информационной системы для работы с документами по научному наследию // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2012. Т. 10. № 2. С. 5-14.
22. Резниченко В.А, Проскудина Г.Ю., Кудим К.А. Концептуальная модель электронной библиотеки [Электронный текст] / В. А. Резниченко, Г. Ю. Проскудина, К. А. Кудим// Труды XI Всероссий-

ской научной конференции RCDL'2009», Россия, г. Петрозаводск (Карелия), 17-21 сентября 2009 г. С. 23-31
23. ISO-14721 Reference Model for an Open Archival Information System (OAIS), Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1 August 2009.
24. Candela L., Castelli D., Dobreva M., Ferro N., Ioanni-dis Y., Katifori H., Koutrika G., Meghini C., Pagano P., Ross S., Agosti M., Schuldt H., Soergel D. The DELOS Digital Library Reference Model Foundations for Digital Libraries. IST-2002-2.3.1.12. Technology-enhanced Learning and Access to Cultural Heritage. Version 0.98, December 2007.
25. Шокин Ю.И., Федотов А.М., Гуськов А.Е., Жижимов О.Л., Столяров С.В. Электронные библиотеки - путь интеграции информационных ресурсов Сибирского отделения РАН // Вестник КазНУ. Серия: Математика, механика, информатика. 2005. – Алматы: КазНУ. № 2. С.115-127.- ISSN 1563-0285.
26. Федотов А.М. Методологии построения распределенных систем // Избранные доклады X Российской конференции «Распределенные информационно-вычислительные ресурсы» (DICR-2005), Новосибирск 6-8 октября 2005 г. / Вычислительные технологии. 2006. Т.11.- С.3-16. – Новосибирск: ИВТ СО РАН. ISSN 1560-7534.
27. Otlet P. Traite de documentation. // Bruxelles: Ed. Mundaneum, 1934.
28. Отле П. Библиотека, библиография, документация: Избранные труды пионера информатики / Поль Отле. М.: ФАИР-ПРЕСС: Пашков Дом, 2004. 348, [1] с. (Специальный издательский проект для библиотек). Библиогр.: с. 312-327. Имен. указ.: с. 340-342.- ISBN 5-8183-0624-0
29. ГОСТ Р ИСО / МЭК ТО 10000-2-99. Информационная технология. Основы и таксономия функциональных стандартов. Часть 2. Принципы и таксономия профилей ВОС.

30. Бездушный А.Н., Бездушный А.А., Серебряков В.А., Филиппов В.И. Интеграция метаданных Единого Научного Информационного Пространства РАН. М.: ВЦ РАН, 2006.
31. Федотов А.М., Барахнин В.Б., Жижимов О.Л., Федотова О.А. Модель информационной системы для поддержки научно-педагогической деятельности // Вестник НГУ. Сер.: Информационные технологии. 2014. Т.12, № 1. С.89-101.- ISSN 1818-7900.
32. Федотов А.М., Жижимов О.Л., Князева А.А., Колобов О.С., Мазов Н.А., Турчановский И.Ю., Федотова О.А. Проблемы авторитетного контроля для распределенных электронных библиотек и библиографических баз данных. Вестник НГУ. Серия: Информационные технологии. 2011. Т. 9. № 1. С.89-101.
33. Захаров А.А., Серебряков В.А. Система управления электронными библиотеками LibMeta // Труды 12-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010. с. 28-37.
34. DCMI – Dublin Core Metadata Initiative (http://www.dublincore.org/).
35. DSpace [Электронный ресурс] : an open source solution for accessing, managing and preserving scholarly works // dspace.org [web-сайт] / MIT Libraries; HP Labs. 2007. <http://www.dspace.org/>
36. EPrints Free Software [Электронный ресурс] // EPrints for Digital Repositories [web-сайт] / School of Electronics and Computer Science, University of Southampton, UK. – 2008. <http://www.eprints.org/>
37. Fedora [Электронный ресурс] : Fedora Repository System // Fedora Commons [web-сайт] / Gordon and Betty Moor Foundation; Cornell University Information Science; University of Virginia Library; The Andrew W. Mellon Foundation. 2007. <http://www.fedora-commons.org/>
38. Кудим К.А,. Проскудина Г.Ю, Резниченко В.А. Сравнение систем электронных библиотек EPrints 3.0 и DSpace 1.4.1 / Девятая всероссийская научная конференция «Электронные библиотеки: пер-

спективные методы и технологии, электронные коллекции». Переяславль-Залесский, 15–18 октября 2007 года.
39. The Open Archives Initiative Protocol for Metadata Harvesting [Электронныйресурс]: Protocol Version 2.0 of 2002-06-14 // Open Archives Initiative: [web-сайт] / The OAI Executive; OAI Techical Commitee. 2004 (http://www.openarchives.org/).
40. RFC 4510 [Электронныйресурс]: Lightweight Directory Access Protocol (LDAP): Technical Specification Road Map / OpenLDAP Foundation. – 2006. <http://www.apps.ietf.org/rfc/rfc4510.html>
41. ANSI/NISO Z39.50-2003. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. NISO Press, Bethesda, Maryland, U.S.A. November 2002.
42. Жижимов О.Л., Мазов Н.А. Принципы построения распределенных информационных систем на основе протокола Z39.50. ОИГГМ СО РАН, Новосибирск: ИВТ СО РАН. 2004. – 361 с.
43. SRU (Search/Retrieve via URL) [Электронныйресурс]. Режим доступа: http://www.loc.gov/standards/sru/ (дата обращения: 23.08.2013).
44. Жижимов О.Л., Федотов А.М., Шокин Ю.И. Технологическая платформа массовой интеграции гетерогенных данных // Вестник НГУ. Серия: Информационные технологии. 2013. Т. 11. № 1. С. 24-41.
45. ГОСТ 7.19-2001. Система стандартов по информации, библиотечному и издательскому делу. Формат для обмена данными. Содержание записи
46. RUSMARC [Электронный ресурс]: Российский коммуникативный формат // Российская Библиотечная Ассоциация [web-сайт] / Российская Библиотечная Ассоциация. http://www.rba.ru:8101/rusmarc/
47. Extensible Markup Language (XML) 1.0 (Fourth edition) [Электронныйресурс] // World Wide Web Consortium [web-сайт]. – 2006. <http://www.w3.org/XML/>
48. ISO 2709:2008: Information and documentation – Format for information exchange // ISO - International Organization for Standardization. – 2008.

ZOOSPACE В ПРОЕКТАХ ИНТЕГРАЦИИ РАЗНОРОДНЫХ РАСПРЕДЕЛЕННЫХ РЕСУРСОВ: СОСТОЯНИЕ И ПЕРСПЕКТИВЫ[1]

О.Л. Жижимов[a], *А.М.Федотов*[a], *Ю.И Шокин*[a], *А.Е.Гуськов* [a,b]

[a] Институт вычислительных технологий СО РАН, Новосибирск

[b] Государственная публичная научно-техническая библиотека СО РАН

zhizhim@mail.ru, fedotov@sbras.ru, dir@ict.nsc,ru, guskov@ict.sbras.ru

Рассматриваются вопросы, связанные с использованием технологической платформы ZooSPACE, созданной в результате выполнения Государственного контракта № 07.514.11.4130 от 6.06.2012 Министерства образования и науки РФ в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы» по теме «Разработка принципов и программных средств виртуальной интеграции распределенных источников данных на основе международных стандартов для создания масштабных информационных инфраструктур», в проектах интеграции информационных ресурсов ведомственного, муниципального и федерального уровней. Обсуждаются направления дальнейшего развития ZooSPACE.

Ключевые слова: распределенные информационные системы, интеграция гетерогенных данных, управление доступом к информационным ресурсам, Z39.50, LDAP, SRW/SRU, ZooSPACE.

The problems associated with the use of a technological platform ZooSPACE, created as a result of the State contract № 07.514.11.4130 from 6.06.2012 Ministry of Education and Science of the Russian Federation within the framework of the Federal Program "Research and development on

[1]Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации (ГК № 07.514.11.4130, ГК № 14.521.11.0004), при частичной поддержке РФФИ (проекты 12-07-00472, 13-07-00859), интеграционных проектов СО РАН, проектов ФНИ и президентской программы «Ведущие научные школы РФ» (грант НШ–5006.2014.9).

priority directions of scientific-technological complex of Russia for 2007-2013 "on the" Development of principles and software virtual integration of distributed data sources on the basis of international standards for the development of large-scale information infrastructures", in projects of integration of information resources, departmental, municipal and federal levels. The directions further development ZooSPACE are discussed.

Keywords: distributed information systems, the integration of heterogeneous data, access control to information resources, Z39.50, LDAP, SRW/SRU, ZooSPACE.

Результатом выполнения Государственного контракта № 07.514.11.4130 от 6.06.2012 Министерства образования и науки РФ в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы» по теме «Разработка принципов и программных средств виртуальной интеграции распределенных источников данных на основе международных стандартов для создания масштабных информационных инфраструктур» явилось создание технологической платформы интеграции разнородных распределенных данных ZooSPACE [1-3].

Платформа ZooSPACE ориентирована на создание распределенных информационных систем (РИС), интегрирующих разнородные информационные ресурсы, управляемые различными СУБД, на основе единых политик организации доступа к этим ресурсам. Программные компоненты ZooSPACE функционируют на различных программно-аппаратных платформах серверов, сосредоточенных в узлах распределенной системы. Взаимодействие узлов между собой осуществляется посредством сетевых протоколов прикладного уровня на основе транспортного протокола TCP/IP. Количество узлов в ZooSPACE не нормируется и может быть любым. Система ZooSPACE может состоять из одного единственного узла. Такой выбор инфраструктуры узлов позволяет обеспечить достаточно гибкую распределенную информационную систему и реализовать всю необходимую функциональность, которая обеспечивается подсистемами ZooSPACE.

В качестве подсистем ZooSPACE выступают следующие [1]:

ZooSPACE-L обеспечивает функционирование справочной и административной подсистемы ZooSPACE. Она интегрирует совокупность LDAP серверов узлов, функционирующих в соответствии с единой для всех политикой и хранящих в виде единой иерархической базы данных (системный каталог ZooSPACE, далее каталог ZooSPACE) всю конфигурационную и административную информацию ZooSPACE. Все LDAP серверы подсистемы ZooSPACE-L связаны правилом двусторонней репликации каталога ZooSPACE по сетевому протоколу LDAP(S). Количество LDAP серверов в ZooSPACE-L не нормировано. Общую функциональность может обеспечить один единственный сервер LDAP.

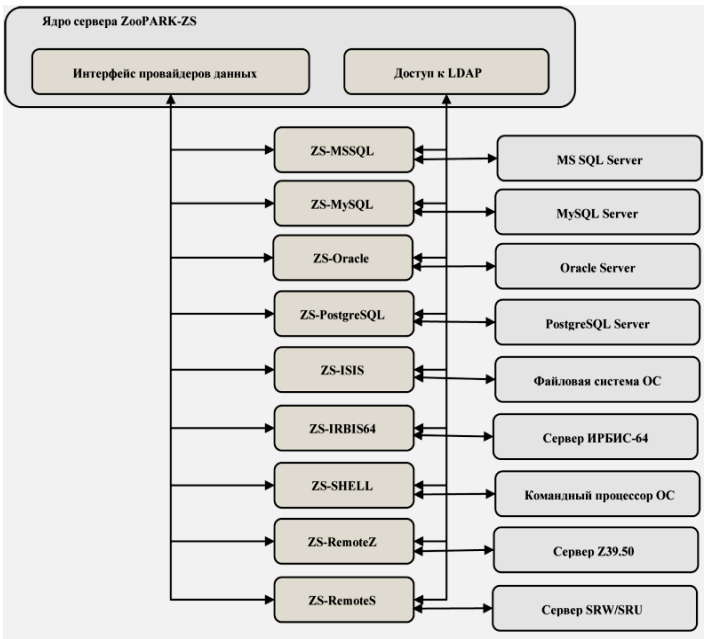


Рис. 1. Организация доступа к данным для сервера ZooPARK-ZS.

ZooSPACE-Z обеспечивает функционирование подсистемы доступа к базам данных системы ZooSPACE. Она интегрирует совокупность Z39.50 и SRW/SRU серверов узлов, функционирующих в соответствии с единой для всех политикой. В качестве серверов Z39.50 и SRW/SRU используется модифицированный сервер ZooPARK – ZooPARK-ZS. Количество серверов ZooPARK-ZS в ZooSPACE-Z не нормировано. Общую функциональность может обеспечить один единственный сервер ZooPARK-ZS. Каждый сервер ZooPARK-ZS в ZooSPACE-Z взаимодействует с подсистемой ZooSPACE-L по протоколу LDAP/LDAPS для получения конфигурационной и административной информации из каталога ZooSPACE. Аутентификация и авторизация всех пользователей ZooSPACE-Z также происходит в подсистеме ZooSPACE-L. Каждый сервер ZooPARK-ZS в ZooSPACE-Z предоставляет интерфейсы доступа к данным по протоколам Z39.50 и SRW/SRU в соответствии со спецификациями этих протоколов и обеспечивает взаимодействие с серверами СУБД, которые по отношение к подсистеме ZooSPACE-Z являются внешними, но могут использовать политику аутентификации и авторизации своих пользователей в подсистеме ZooSPACE-L. Одной из обязательных функцией серверов ZooPARK-ZS является возможность переадресовывать запросы на доступ к данным на другие серверы ZooPARK-ZS подсистемы ZooSPACE-Z, а также на серверы Z39.50 и SRW/SRU, не входящие в ZooSPACE-Z, по соответствующим протоколам.

ZooSPACE-M обеспечивает функционирование системы мониторинга всех компонент ZooSPACE.

ZooSPACE-S обеспечивает функционирование подсистемы сбора статистики работы всех компонент ZooSPACE.

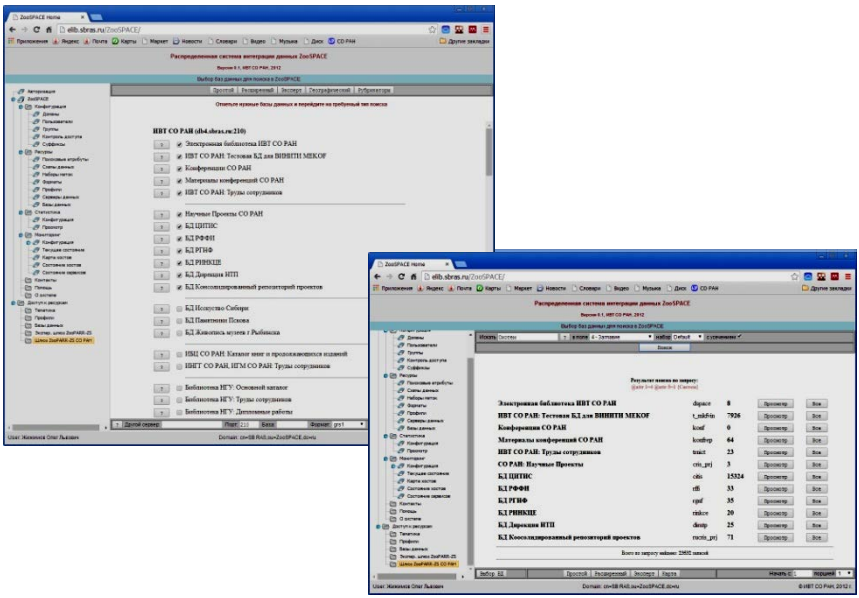


Рис. 2. WEB интерфейсы ZooSPACE-W.

ZooSPACE-W предназначена для предоставления административных и пользовательских WEB-интерфейсов для доступа к ZooSPACE. Подсистема ZooSPACE-W включает как минимум один WEB сервер. Каждый сервер подсистемы ZooSPACE-W хранит одинаковый набор программного обеспечения, реализующий необходимую функциональность для формирования WEB интерфейсов и внутренней обработки данных. Клиент может обращаться к любому из серверов без потери функциональности. Наличие нескольких серверов в подсистеме ZooSPACE-W повышает уровень доступности серверов и минимизирует трафик между разными узлами. Программное обеспечение сервера подсистемы ZooSPACE-W состоит из блоков:

- **Блок Z** реализует интерфейсы доступа к подсистеме ZooSPACE-Z. Этот блок обеспечивает поиск и представление данных из различных СУБД в соответствии с выбранным профилем.
- **Блок L** реализует интерфейсы доступа к подсистеме ZooSPACE-L как набор административных интерфейсов доступа к каталогу ZooSPACE. Доступ только для администраторов. Блок содержит функции авторизации пользователей. Фактически этот блок реализует интерфейсы для просмотра и модернизации каталога ZooSPACE.
- **Блок S** реализует интерфейсы доступа к подсистеме ZooSPACE-S. Возможны разные уровни доступа. Интерфейсы предназначены для просмотра статистической информации о работе системы ZooSPACE.
- **Блок M** реализует интерфейсы доступа к подсистеме ZooSPACE-M. Возможны разные уровни доступа. Интерфейсы предназначены для просмотра результатов мониторинга различных компонент ZooSPACE.

1. Проекты, направленные на развитие информационных систем СО РАН

В процессе выполнения различных проектов, в том числе и интеграционных проектов СО РАН, экспериментальная РИС на основе ZooSPACE, созданная как экспериментальный стенд для выполнения Государственного контракта № 07.514.11.4130, была существенно расширена как в части количества узлов, так и в части номенклатуры интегрируемых информационных ресурсов.

На сегодняшний день РИС ZooSPACE включает пять узлов (ИВТ СО РАН, ГПНТБ СО РАН, ТФ ИВТ СО РАН, ИВМ СО РАН, ИДСТУ СО РАН) в Новосибирске, Томске, Красноярске и

Иркутске. Каждый узел содержит серверы LDAP, Z39.50/SRW/SRU (ZooPARK-ZS), WEB-сервер, что позволяет организовать доступ к различным СУБД и предоставить унифицированные пользовательские и административные интерфейсы для доступа к информационным ресурсам. На текущий момент система обеспечивает доступ к более 70 базам данных различной тематической направленности [3]:

- каталоги книг, периодических изданий и авторефератов и диссертаций
- БД трудов сотрудников организаций СО РАН
- цифровые коллекции и репозитории
- ресурсы по культурному наследию
- реферативные библиографические БД
- и др.

Номенклатура доступных в ZooSPACE информационных ресурсов постоянно расширяется.

2. Проекты, выполняемые при поддержки Министерства образования и науки РФ в рамках ФЦП

Платформа ZooSPACE и РИС СО РАН на ее основе получили дальнейшее развитие в процессе выполнения гранта Министерства образования и науки РФ в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы» по теме «Разработка системы агрегирования данных по научным проектам из различных источников для обеспечения мониторинга реализации мероприятий и программ», шифр 2013-2.1-14-521-0017 (контракт № 14.521.11.0004).

В рамках этого проекта РИС ZooSPACE использовалась как инфраструктурная основа для сбора информации из распределенных разнородных источников данных по научным проектам.

Были разработаны и исследованы методы, процедуры и регламенты сбора информации о финансировании и результатах научно-технической деятельности в РФ, основанные на существующих стандартах взаимодействия распределенных систем и развивающемся формате описания данных о научных исследованиях CERIF [4].

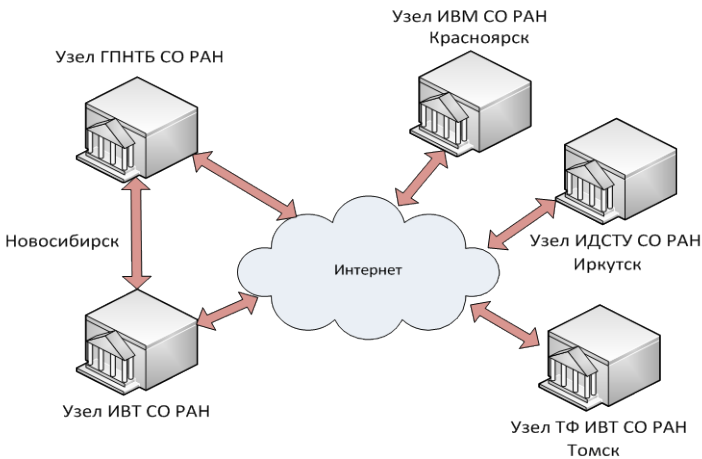


Рис. 3, Инфраструктура РИС СО РАН на основе ZooSPACE.

Разработан экспериментальный макет системы агрегирования данных по научным проектам из различных источников. В состав макета входит репозиторий информации о научных проектах и РНТД, подсистема сбора и обработки информации, пользовательский интерфейс для просмотра содержимого реестров репозитория, подсистема управления функционированием репозитория. Технология сбора данных, реализованная в макете, основана на

использовании стандарта Z39.50, протокола SRU/SRW и программного обеспечения ZooSPACE. Структура репозитория данных о научных проектах основана на схеме данных CERIF 1.5 [5].

Изначально техническим заданием предусматривалась демонстрация возможности интегрирования данных о научных проектах из пяти различных источников (ЦИТИС, РФФИ, РГНФ, РИНКЦЕ и Дирекция НТП). Однако в виду отсутствия физического доступа к реальным источникам, данные, полученные от поставщиков были загружены в созданные в ZooSPACE эмуляторы источников с сохранением первоначальной структуры данных. Эти эмуляторы реальных источников данных и участвовали во всех дальнейших испытаниях.

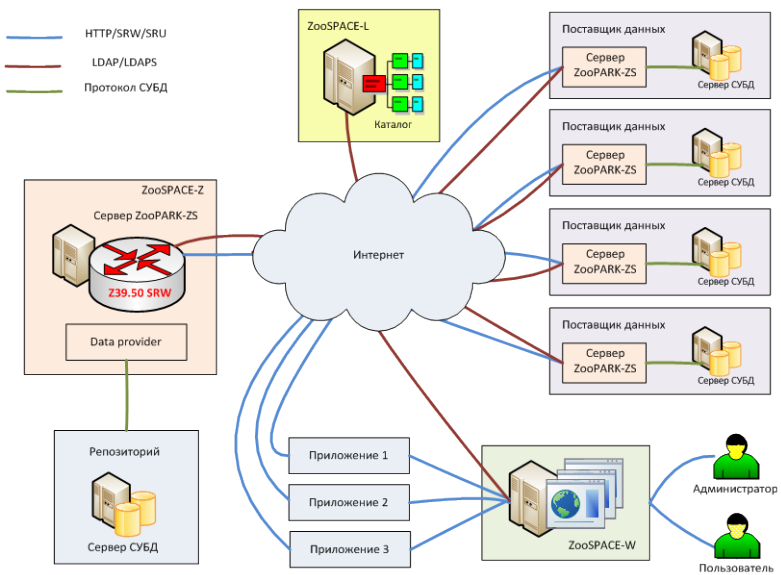


Рис. 4. Инфраструктура экспериментального стенда на основе ZooSPACE для интеграции данных по научным проектам.

В рамках экспериментальных испытаний были загружены данные о более чем 20 000 научных проектов, выполнявшихся в РФФИ, РГНФ, РИНКЦЭ, Сибирском отделении РАН, а также в рамках Федеральных целевых программ. Для апробации макета системы была разработана программа и методики испытаний, на основании которых были проведены испытания и подтверждено соответствие макета исходным требованиям [5].

3. Другие проекты

В 2014 году сотрудником ИВТ СО РАН был получен грант мэрии г. Новосибирска на выполнение проекта "Разработка модели и прототипа открытой краеведческой цифровой библиотеки Новосибирска"[2]. Работы по этому проекту предусматривают интеграцию созданных программных компонент и информационных ресурсов с РИС СО РАН на основе ZooSPACE.

4. Перспективы развития

На сегодняшний день просматриваются следующие направления развития ZooSPACE:

Усовершенствование и оптимизация программных компонент подсистем ZooSPACE

Реализованные в настоящее время программные компоненты ZooSPACE постоянно модернизируются в части как расширения их функциональных свойств, так и в части повышения их надежности для обеспечения устойчивой работы в штатных и в нештатных условиях.

[2] Список победителей конкурса на предоставление субсидий молодым ученым и специалистам в сфере инновационной деятельности в 2014 году// Официальный сайт города Новосибирска - http://www.novosibirsk.ru/articles/city_adm/departments/dpiip/u_nip/konkursy/young-scientists-2014/

В частности, в последнее время ведется большая работа по более тесной интеграции ZooSPACE с системой поддержки цифровых репозиториев, например, на основе DSpace [6]. Эта интеграция будет обеспечиваться специальным провайдером данных для сервера ZooPARK-ZS, реализующим интерфейсы доступа к данным Apache/Solr [7]. Провайдер будет функционировать в соответствии с общими спецификациями провайдеров данных ZooPARK-ZS, взаимодействовать с подсистемой ZooSPACE-L в части конфигурирования источников данных, конвертирования поисковых запросов, конфигурирования конвертеров данных для различных схем на основе преобразований XSLT. Этот провайдер также будет обеспечивать доступ к данным статистики, индексируемым в Apache/Solr.

По мере возникновения потребности в организации доступа к СУБД, неподдерживаемых в текущей реализации ZooSPACE, возможно расширение номенклатуры поддерживаемых СУБД при помощи создания новых провайдеров данных для ZooPARK-ZS в соответствии с общими спецификациями.

Оптимизация алгоритмов и протоколов взаимодействия подсистем ZooSPACE

Список используемых в ZooSPACE протоколов LDAP, Z39.50, HTTP, HTTP/SRU планируется расширить протоколами HTTP/SOLR и HTTP/OAI-PMH для обеспечения доступа к соответствующим источникам данных и расширения функциональных возможностей по частичной репликации данных.

Добавление новых узлов и новых ресурсов к существующей инфраструктуре ZooSPACE

Функционирующая сегодня распределенная информационная система, базирующаяся на пяти узлах, открыта для расширения по количеству узлов и номенклатуре доступных информационных ресурсов.

В частности, к концу года ожидается подключение к ZooSPACE краеведческих ресурсов Новосибирска в рамках выполнения гранта мэрии по проекту "Разработка модели и прототипа открытой краеведческой цифровой библиотеки Новосибирска" (см. выше).

Литература

1. *Жижимов О.Л., Федотов А.М., Шокин Ю.И.* Технологическая платформа массовой интеграции гетерогенных данных // Вестник НГУ. Сер.: Информационные технологии. - 2013. - Т.11. - № 1. - С.24-41. - ISSN 1818-7900.
2. *Жижимов О. Л., Лобыкин А. А., Турчановский И. Ю., Паньшин А. А., Чудинов С. А.* Автоматизированная система сбора статистической информации о событиях в распределенной информационной системе // Вестник НГУ. Сер.: Информационные технологии. - 2013. - Т.11. - № 1. - С.42-52. - ISSN 1818-7900.
3. *Жижимов О.Л., Федотов А.М., Шокин Ю.И.* Платформа ZooSPACE - организация доступа к разнородным распределенным ресурсам // Электронные библиотеки: российский научный электронный журнал. - 2014. - Т.17. - № 2. - ISSN 1562-5419.
4. CERIF 1.5 XML. Data Exchange Format Specification. euroCRIS. - 13 Feb 2013. [Электронный ресурс]. Режим доступа: http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.5/CERIF1.5_XML.pdf
5. *Guskov A.E., Zhizhimov O.L., Kikhtenko V., Skachkov D.M., Kosyakov D.* RuCRIS: A Pilot CERIF based System to Aggregate Heterogeneous Data of Russian Research Projects // Procedia Computer Science. - 2014. - Vol.33. - P.163-167. - ISSN 1877-0509.
6. DSpace Documentation [Электронный ресурс] // Режим доступа: https://wiki.duraspace.org/display/DSDOC/All+Documentation
7. Apache Solr [Электронный ресурс] // Web Site. The Apache Software Foundation. Режим доступа: http://lucene.apache.org/solr/

МЕТАДАННЫЕ – ОСНОВА АВТОМАТИЗАЦИИ ПО СОЗДАНИЮ ИНФОРМАЦИОННОЙ ПРОДУКЦИИ

Е.Д.Вязилов, Д.А.Мельников, Н.В.Чуняев, А.Е.Кобелев

ФГБУ «Всероссийский научно – исследовательский институт *гидрометеорологической* информации – Мировой центр данных»

vjaz@meteo.ru, melnikov@meteo.ru, chunyaev@meteo.ru, kobelev@meteo.ru

Представлены краткие сведения о Единой государственной системе информации об обстановке в Мировом океане (ЕСИМО). Рассмотрено повышение эффективности работы системы на этапе эксплуатации, связанное с объединением информационных ресурсов, маппированием классификаторов, слиянием накопленных данных с оперативными потоками данных и прогнозами, вычислением новых параметров, улучшением поиска, использованием атрибутов метаданных при визуализации данных, типизацией шаблонов визуализации информационных ресурсов, получением интерактивной продукции, мониторингом гидрометеорологической обстановки.

Short information on Unified system information on the World Ocean is providing. It is presenting the questions of effective exploitation of system. They connected with union of information resources; mapping qualifiers; join of historical data; operational data flows and forecasts; calculation of new parameters; search improvement; use metadata attributes at data visualization; standardization of visualization templates of information resources; receiving of interactive production; monitoring of a hydrometeorological situation.

1. Введение

С 1 января 2014 г. Единая государственная система информации об обстановке в Мировом океане (ЕСИМО, http://esimo.ru) находится в постоянной эксплуатации. Главный принцип реализации системы заключается в переходе от взаимодействия отдель-

ных серверов и систем хранения к сетевому взаимодействию узлов, состоящих из многих программных компонент. Каждый узел состоит из виртуальных машин, на которых установлены компоненты (портал, аналитический комплекс, картографический визуализатор, др.). И все это дополняется программным обеспечением, обеспечивающим автоматизацию доставки данных, сервисов, учета использования информационных ресурсов (ИР), мониторинга работы аппаратно-программных средств. При создании программного обеспечения использован портлетная технология.

Специфика интеграции данных в ЕСИМО заключается в том, что поставщик данных через сервер интеграции представляет данные в базу интегрированных данных (БИД) путем маппирования локальных имен параметров в общесистемные коды и используемых локальных классификаторов в общепринятые. В БИД возможны, как выборка, так и объединение однородных ИР, а также слияние данных и метаданных. За счет этого каждый пользователь может получить только ту информацию, и в том формате, который необходим ему. При этом возможно доставить данные от одного поставщика – множеству пользователей и из множества поставщиков – одному пользователю. Подключение новых поставщиков данных не приводит к изменениям ни в сервере интеграции, ни в БИД. Если у поставщика данных изменился состав ИР, то это никак не влияет на работу сервера интеграции и БИД, нужно только удалить или добавить новый системный элемент. Если возникли проблемы при маппировании параметров (например, несоответствие формата хранения параметра), то требуется либо ввести новый системный элемент, либо уточнить формат существующего элемента. ЕСИМО принимает практически любые форматы хранения данных от поставщиков данных. При этом

широко используются стандарты интероперабельности ИСО, OGC, W3C, др. То есть созданы масштабные типовые решения, способные хорошо тиражироваться и быстро настраиваться на нужды пользователей. Таким образом, начал осуществляться переход от разрозненных, фрагментарных баз (наборов) данных к единому пространству данных [4].

Целью эксплуатации системы является повышение качества и доступности данных о морской среде и морской деятельности. Повышение эффективности работы ЕСИМО на этапе эксплуатации связано с объединением ИР, маппированием классификаторов, слиянием накопленных данных с оперативными потоками данных и прогнозами, вычислением новых параметров, улучшением поиска, использованием атрибутов метаданных при визуализации данных, типизацией шаблонов визуализации ИР, получением интерактивной продукции, мониторингом гидрометеорологической обстановки.

2. Объединение информационных ресурсов

При передаче данных по глобальной сети телесвязи, хранении климатических обобщений для уменьшения дублирования значений атрибутов в исходных ИР применяются коды судов, буев, портов, гидрологических постов, прибрежных, метеорологических и аэрологических станций. А при выдаче информации пользователю необходима полная информация об этих объектах, которая находится в других ИР – будем называть их базовыми. Поэтому возникает задача объединения таких ИР. Примеры сведений о необходимости объединения ИР от различных поставщиков данных даны в табл. 1.

Таблица 1 - Сведения о необходимости объединения информационных ресурсов

ИД исходного ИР	Атрибут связи	ИД базового ИР	Атрибут связи	Название
RU_RIHMI-WDC_106	M4200	RU_RIHMI-WDC_2668	M4200	Прибрежные станции
RU_RIHMI-WDC_177	M4105	RU_RIHMI-WDC_1278	M4033	Организации
RU_FERHRI_26	M4202	RU_RIHMI-WDC_1283	M4215	Попутные суда
RU_PUGS_13	M4200	RU_RIHMI-WDC_2666	M4201	Гидрологические посты
RU_NFR_38	M4202	RU_NFR_01	M4200	Суда Росрыболовства
RU_MORSVJAZSPUTNIK_35	M4202	RU_MORSVJAZSPUTNIK_05	M4202	Суда Минтранс
RU_CNIIMF_66	M4032	RU_CNIIMF_27	M4032	Порты
RU_AARI_1401	M4201	RU_RIHMI-WDC_1288	M4200	Буй

Таблица 2 - Выдача названий вместо кодовых значений при визуализации информационных ресурсов на портале, в ГИС и других компонентах

ИД ресурса	Атрибут, код	Идентификатор классификатора	Название параметра
RU_RIHMI-WDC_105	M4200	140	Синоптический номер станции
RU_RIHMI-WDC_1165	P0399_00	233	Облачность общая: количество

Таблица 3 - Слияние накопленных данных с оперативными потоками данных и прогнозами для одной станции одного или нескольких параметров

ИД ресурса с текущими данными	Атрибут связи	ИД ресурса с историческими данными	Атрибут связи	Врем. масштаб	Параметр	Источник
RU_RIHMI-WDC_1198	M4200	RU_RIHMI-WDC_706	M4000	Сутки	P0229_03	КН-02. Т воды
RU_RIHMI-WDC_1198	M4200	RU_RIHMI-WDC_706	M4000	Сутки	P0229_01	КН-02 Т воды
RU_RIHMI-WDC_1198	M4200	RU_RIHMI-WDC_706	M4000	Сутки	P0229_02	КН-02 Т воды
RU_RIHMI-WDC_1220	M4200	RU_RIHMI-WDC_912	M4000	Сутки	P0074_00	КН-02 Напр. ветра

3. Маппирование классификаторов

Иногда в исходных ИР хранятся в виде кодов не только идентификационные сведения, но и сами данные. Кроме того, часто в источниках данных используются различные системы кодирования. Например, в сообщениях, передаваемых по глобальной сети телесвязи, используется синоптический номер станции, а пользователю необходимо название станции или вместо кода «количества облаков» выдавать полное название. Возникает задача приведения их к единым стандартам, т.е. требуется маппирование локальных кодов в общепринятые. Классификаторы также представлены в системе в виде ресурса. Поэтому перед визуализацией исходного ИР необходимо объединить его с ресурсом, содержащем классификаторы (табл.2). Фактически это частный случай объединения ИР.

4. Слияние накопленных данных с оперативными потоками данных и прогнозами

Принципиально новым видом информационной продукции являются результаты объединения исторических данных с оперативными потоками данных и прогнозами для одной станции, одного или нескольких параметров, на одном графике. Чтобы получить такую продукцию, необходимо на уровне БИД создать производный ИР, в котором исходные данные дополняются текущими оперативными данными или результатами прогнозов на ближайшие 120 часов (табл.3).

Таблица 4 - Вычисление новых параметров

ИД исходного ресурса	Исходные элементы, название	Исходные элементы, код	Вычисляемый параметр, название	Вычисленные элементы, код
RU_Hydrometcentre_42	U – зональная составляющая	P0696_00	Направление ветра	P0074_00
RU_Hydrometcentre_42	V – меридиональная составляющая	P0779_00	Скорость ветра	P0075_00

Таблица 5 - Сведения о возможности вывода представлений на карте, в виде временного ряда, графика

Код параметра	Полное название	Гео	Время	График	Горизонт	ИД классификатора
M4203	Платформа: тип	1		2		288
M4224	Судно: номер ИМО					335
M4302	Геообъект: код	2		2		287
P0001_00	Температура воздуха: измеренная	5	1	1	1	
P0001_01	Температура воздуха: мин	5	1	1	1	
P0001_03	Температура воздуха: средняя	5	1	1	1	
P0893_15	Лед сплоченность	2	5			108

5. Вычисление новых параметров

Важным функцией любого приложения является возможность вычисления новых параметров по данным исходных ИР. Например, расчет направления и скорости ветра на основе их составляющих (табл.4) или получение статистических характеристик. Аналогично можно подключить модули расчета других параметров.

6. Поиск данных

Недостатками первого варианта поиска ресурсов был сложный интерфейс, относительно большое время доступа к данным, достаточно сложная визуализация. Последняя версия приложения по поиску данных, во-первых, работает с БИД, в которую в соответствии с установленным регламентом загружаются распределенные, неоднородные ресурсы. Во-вторых, средства визуализации ИР настраиваются на атрибуты метаданных, что позволяет автоматически выбирать шаблоны представления данных (таблица, карта, множество графиков) в зависимости от значений таких атрибутов как периодичность наблюдения (обобщения), пространственное и вертикальное разрешения, тип платформы.

Для того чтобы пользовательский опыт был более комфортным, чтобы не приходилось много раз щелкать ссылки, пока найдешь ИР, сделан вариант, когда наборы ИР выделены в отдельные экземпляры портлета.

Приложения по поиску данных должны более точно выдавать данные из БИД. Список возможных значений поисковых атрибутов может быть большим, поэтому предлагается более универ-

сальный подход по формированию списков синхронизированных значений атрибутов. Для этого можно использовать поиск по двум или нескольким синхронизированным атрибутам. Например, если ищутся объекты, расположенные в разных странах, принадлежащие разным ведомствам, организациям, то необходимо сначала из БИД выбрать список уникальных значений стран, для которых существует необходимый объект. Затем после выбора пользователем одной из стран по ее значению создается список уникальных значений ведомств или организаций для выбранной страны. На основе выбора пользователем необходимого ведомства или организации происходит запуск запроса на выборку данных для объектов, находящихся в данной стране, ведомстве или организации (рис.1). В результате всегда выдается хотя бы одна запись. Такой поиск используются в ЕСИМО для визуализации метаданных (например, сведений о рейсах НИС по стране и организации).

Несмотря на наличие в системе достаточно подробных метаданных на каждый ИР, при обработке данных выявлена необходимость иметь дополнительные сведения о возможности вывода картографических объектов на карте, отображения параметров во времени, построения графиков определенного типа для различных параметров, представления параметров на горизонтах, использования классификаторов (табл.5).

7. Визуализация данных

Средства визуализации информационных ресурсов настраиваются на атрибуты метаданных, что позволяет автоматически выбирать шаблоны представления данных в зависимости от значений таких атрибутов метаданных как периодичность наблюдения (обобщения), пространственное и вертикальное разрешения, тип платформы.

В зависимости от типа системы хранения данных (структурированные данные, объектные файлы, приложения, географические сервисы, аналитическое представление) используется свой шаблон представления ИР. Возможными вариантами представлений структурированных ресурсов могут быть точки - классический или псевдо временной ряд, профиль, регулярные сетки.

Возможными вариантами представлений структурированных ресурсов могут быть [1]:

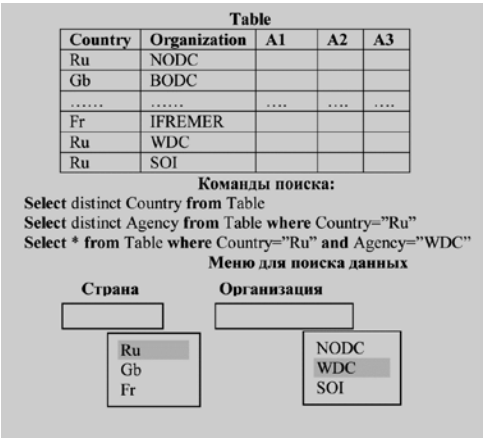


Рис. 1 Поиск данных по синхронизированным значениям атрибутов.

- **Точки:** фиксированная в пространстве, с регулярными измерениями во времени - **классический временной ряд;** случайные точки в пространстве (траектории) с регулярными или нерегулярными измерениями во времени; точка, относящаяся к определенной площади (квадрату, региону, области), с предварительным обобщением по любому временному масштабу и с возможными пропусками, требующие предварительной интерполяции и идентификации места – **псевдо временной ряд**; точка в виде псевдо временного ряда, приведенного к классическому виду путем интерполяции во времени; **данные в виде временного ряда, полученные на основе сеточных данных** – выборка для одной точки за весь период времени на одном уровне;
- **Профили: случайный** в виде исходных или вычисленных значений на различных горизонтах со случайными измерениями во времени; **случайный на стандартных горизонтах** – вычисленные и интерполированные данные; **серия профилей во времени**, полученных на основе случайных профилей, агрегированных для квадрата (области) с интерполяцией во времени (рейдовый пункт); **профили, отнесенные к центру района** в виде типового профиля - многолетние обобщения**; псевдо профили, полученные на основе сеточных данных** – выборка для одной точки для всех уровней в один момент времени.
- **Регулярные сетки** - нерегулярные данные, интерполированные в пространстве для одного момента времени, могут быть двухмерные (широта, долгота), трехмерные (широта,

долгота, уровень), четырехмерные (широта, долгота, уровень, время).

Правила выделения шаблонов для структурированных данных представлены ниже.

1) Если Временное представление: 12-часов, или 3-часа, или 6-часов. или год, или полугодие, или ежеквартально, или ежемесячно, или еженедельно, или ежесуточно, или ежечасно (далее - регулярное) и Пространственное разрешение: фиксированная точка и Платформа: неподвижная наблюдательная платформа на суше или в прибрежной зоне и Вертикальное представление: поверхность, **то это классический временной ряд.** Для этих ресурсов можно на основе карты выбрать точку или район, построить графики временного хода, представить данные в виде таблицы.

2) Если Временное представление: регулярное и Пространственное разрешение: район, то это **псевдо временной ряд**. Для этих ресурсов можно на основе карты выбрать точку или района, построить графики временного хода, представить данные в виде таблицы. Строить график нужно без соединения точек или в виде диаграммы.

3) Если Временное представление: регулярное и Пространственное разрешение: район и Вертикальное разрешение: слой или стандартные уровни, то это **псевдо временной ряд на одном из уровней.** Для этих ресурсов можно на основе карты выбрать точку или района, построить графики временного хода, представить данные в виде таблицы. Строить график нужно без соединения точек или в виде диаграммы. Примером являются данные обобщений по квадратам в океане.

4) Если Временное представление любое и пространственное разрешение район и Уровень обработки данных: наблюденные значения, то это **временной ряд с нарастающим итогом**. Для него производиться выборка отдельных лет в отдельные колонки произвольного ресурса, на основе которого строятся графики с нарастающим итогом.

5) Если Временное представление: не известно, или не планируется, или нерегулярно, или по необходимости и Пространственное разрешение: не определено или траектория и Вертикальное распределение: поверхность, или стандартные, или нестандартные уровни или не определено и Платформа: подвижная, то это случайный профиль на океанографической станции. Для профиля дается список для выбора платформы, строится карта пространственного распределения, график вертикального распределения гидрофизических параметров, таблица данных.

6) Если Временное представление: любое и Пространственное разрешение: траектория и Вертикальное разрешение: поверхность, Платформа: подвижная, то это **временной ряд для траектории**, для него строится график временного хода для одной платформы с возможностью фильтрации по времени. Примером являются траектории судов, циклонов.

7) Если Временное представление: регулярное и Пространственное разрешение: сетка и Вертикальное представление: стандартные уровни или поверхность и Уровень обработки данных: обобщение в рамках конкретного года или обобщение за многолетний период, то это **псевдо временной ряд**. Для него строится карта с возможностью выбора точки, построения графика временного хода для одного параметра, в точке, на одной поверхно-

сти, таблица данных. Примером являются данные в узлах регулярной сетки.

8) Если Временное представление: регулярное и Пространственное разрешение: сетка и Вертикальное представление: стандартные уровни, то это **псевдо профиль.** Для него можно построить график вертикального изменения параметров по глубине (высоте) на основе выбора точки по карте, а также лучить таблицу.

9) Если Временное представление: не известно, или не планируется, или нерегулярно, или по необходимости и Пространственное разрешение: не определено, то это **случайная точка**, по ней строится диаграмма. Примерами таких ресурсов является количество оборудования по портам, количество наблюдательных платформ (прибрежных станций, буев, научно-исследовательских судов) по организациям.

10) Если ресурс типа Приложение, то требуется только выход на приложение по ссылке.

11) Если ресурс типа Географический сервис, то выдается список слоев геосервиса и карта с выбранными слоями.

12) Если ресурс типа прикладная задача Аналитического комплекса, то выдается список аналитический представлений, входящих в одну из прикладных задач. На основе списка делается выбор соответствующего аналитического представления и отображение элементов выбранного представления.

13) Если ресурс типа Объектные файлы (изображения/документы), то выдается таблица ссылок на объектные файлы.

Примеры одного шаблона визуализации информационных ресурсов представлен на рис.2.

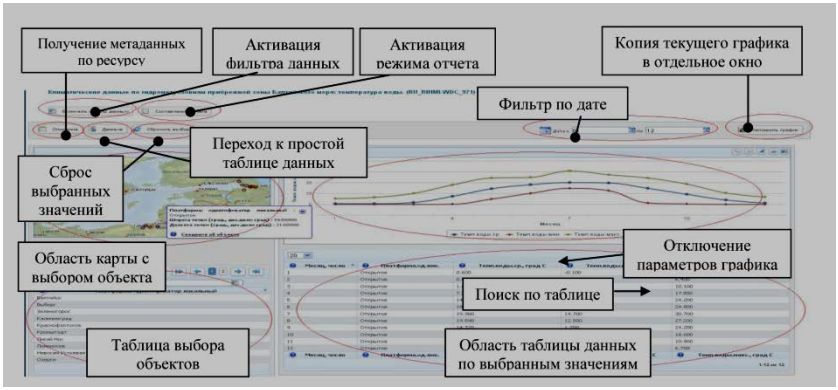


Рис.. 2. Пример одного из шаблонов визуализации информационных ресурсов.

8. Интерактивная продукция

Средства ЕСИМО позволяет создавать визуальные представления ресурсов, с которыми могут в дальнейшем работать любые пользователи по типовому адресу http://esimo.ru/dataview/viewresource?resourceId=RU_Hydrometcentre_52. В этом адресе resourceID=RU_Hydrometcentre_52 является идентификатором ресурса. При переходе на другой ресурс достаточно изменить этот идентификатор. В текст визуализируемого документа включены интерактивные элементы – графики и диаграммы, строящиеся по клику на объект карты, средства поиска данных в ресурсе (см. рис.2). Результаты визуализации представляют собой «живую» информационную продукцию [3], актуальную во времени и которую можно включать в цифровые учебные пособия, научные статьи, разного рода презентации и другие электронные документы. Графики и таблицы перестраиваются после ввода новых данных. Перемещая вертикальную полосу на

графике, можно выбрать часть графика для более детального анализа. Кроме того, в БИД имеются возможности расчета значений новых параметров, что позволяет их автоматически включить в интерактивную продукцию.

9. Монитор гидрометеорологической обстановки

В статье [2] был представлен вариант реализации монитора обстановки для точки, к сожалению, реализовать пока его не удалось. Предлагается проводить индикацию на уровне пока ИР. То есть, если значение какого-либо параметра превышает критическое, то ссылка на ресурс, в котором находится этот параметр, индицируется цветом (желтый, красный, малиновый), рис.3. Далее при просмотре ИР выявленное критическое значение параметра можно посмотреть в виде графика с индикацией его значения, рис.4. При этом можно построить тренд изменения параметра.

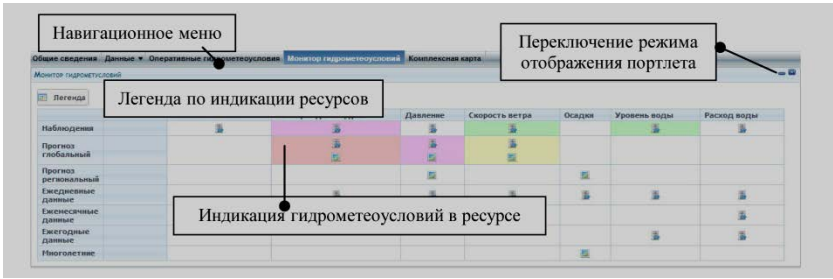


Рис. 3. Монитор гидрометеорологической обстановки.

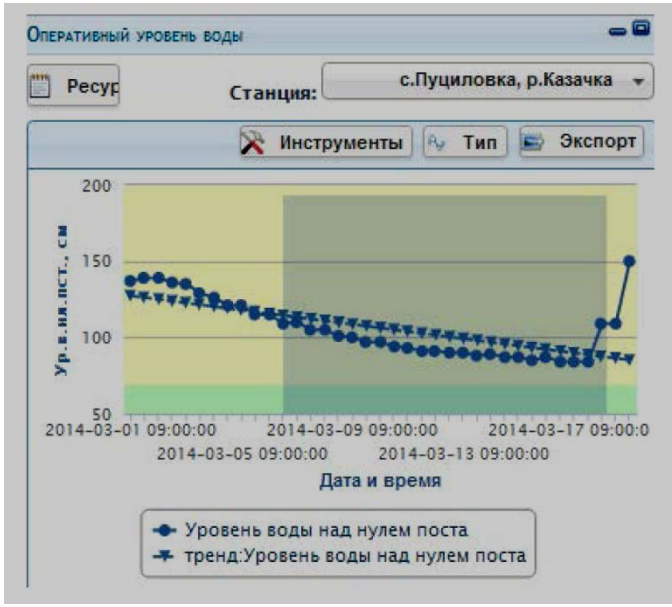


Рис. 4. Монитор гидрометеорологической обстановки.

Средства визуализации ИР настраиваются на атрибуты метаданных, что позволяет автоматически выбирать шаблоны представления данных в зависимости от значений атрибутов метаданных: периодичность наблюдения (обобщения), пространственное и вертикальное разрешения, тип платформы. В зависимости от типа системы хранения данных (структурированные данные, объектные файлы, приложения, географические сервисы, аналитическое представление) используется свой шаблон визуализации ИР.

10. Выводы

Система позволяет значительно сократить сроки доставки информации до пользователей с нескольких суток до минут, обеспечить межведомственный обмен данными и повысить качество информационного обслуживания за счет предоставления комплексной аналитической информации. Внедрение ЕСИМО способствует снижению затрат пользователей на оказание услуг.

Список литературы

1. Вязилов Е.Д. О стандартизации структур данных в области морской среды // Электронный журнал «Новости ЕСИМО». – 2007. Вып.30. [Электронный ресурс]. – Режим доступа: http://esimo.oceaninfo.ru/system/instance/instanceInfo.jsp?instanceId=74235/, свободный. – Загл. с экрана.
2. Вязилов Е.Д., Михайлов Н.Н., Мельников Д.А., Чуняев Н.В. Подходы по визуализации данных ЕСИМО // Российский научный Электронный журнал "Электронные библиотеки". - 2014. – Вып.3. Том 17.
3. Формат вычисляемых документов. - Компания Wolfram. - 2014. [Электронный ресурс]. – Режим доступа: http://www.wolfram.com/cdf/uses-examples/automated-reports.html, свободный.
4. Michael Franklin, Alon Halevy, David Maier. From Databases to Dataspaces: A New Abstraction for Information Management // SIGMOD Record. Dec. 2005. Vol. 34. No. 4.

ТЕХНОЛОГИЯ РАЗРАБОТКИ ИНТЕЛЛЕКТУАЛЬНЫХ НАУЧНЫХ ИНТЕРНЕТ-РЕСУРСОВ, ОРИЕНТИРОВАННАЯ НА ЭКСПЕРТОВ ПРЕДМЕТНОЙ ОБЛАСТИ

Загорулько Ю.А.

Федеральное государственное бюджетное учреждение науки Институт систем информатики им. А.П. Ершова
Сибирского отделения Российской академии наук, Новосибирск
zagor@iis.nsk.su

В докладе рассматриваются основные положения технологии разработки тематических интеллектуальных научных интернет-ресурсов (ИНИР), обеспечивающих содержательный доступ к систематизированным научным знаниям и информационным ресурсам определенной области знаний и средствам их интеллектуальной обработки. Важным преимуществом ИНИР является то, что он позволяет исследователям значительно сократить время, требуемое для обеспечения доступа к необходимой информации и ее анализа, за счет аккумуляции описаний релевантных интернет-ресурсов (в том числе, web-сервисов) непосредственно в своем контенте. Особенностью данной технологии является ориентация на широкой круг специалистов, являющихся экспертами в предметных областях, для которых создаются системы, и использование онтологического подхода.

Ключевые слова: интеллектуальный научный интернет-ресурс, онтология, семантическая сеть, тезаурус, семантический web-сервис, технология, специализированная программная оболочка.

The paper discusses fundamentals of a technology of development of intelligent scientific Internet resources (ISIR) providing the content-based access to the systematized scientific knowledge and information resources related to certain subject domain and to their intelligent processing facilities. An important feature of ISIR is its ability to appreciably reduce time of access to information required by researchers and processing it due to accumulation of descriptions of relevant Internet resources directly in the ISIR content. A main feature of the technology is orientation to experts, i.e. specialists in certain subject domain, and use of ontology-based approach.

Keywords: intelligent scientific Internet resource, ontology, semantic network, thesaurus, semantic web service, technology, specialized program shell.

Введение

В мире накоплено огромное количество информации по различным областям знаний, причем значительная ее часть представлена непосредственно в сети Интернет, но, несмотря на это, проблема эффективного обеспечения научного сообщества информацией по интересующим его тематикам пока не имеет удовлетворительного решения.

Нерешенной остается и проблема удобного доступа к средствам обработки данных, собранным по этим тематикам. Даже уже реализованные и представленные в Интернет в виде web-сервисов методы обработки информации остаются недоступными широкому кругу пользователей из-за отсутствия содержательной информации о них.

Это, с одной стороны, объясняется особенностями представления научных знаний в Интернет, которые слабо формализованы, недостаточно систематизированы и распределены по различным Интернет-сайтам, электронным библиотекам и архивам.

Другая причина такого положения в том, что современные информационные системы используют довольно ограниченный набор методов представления, поиска и интерпретации информации. Как правило, в них данные и знания представляются в виде текстовых документов (в корпоративных информационных системах) или множеством информационных ресурсов (в интернет-каталогах и порталах). В то время как наиболее естественной и удобной формой подачи информации для человека является представление ее в виде сети взаимосвязанных фактов. Такой способ интерпретации информации облегчает ее восприятие, по-

зволяет осуществлять содержательный поиск и удобную навигацию.

Для решения указанных проблем была предложена концепция тематического интеллектуального научного интернет-ресурса (ИНИР), который предназначен для информационной и аналитической поддержки научной и производственной деятельности в определенной области знаний.

Ввиду высокой потребности в системах такого класса нами разрабатывается технология создания и сопровождения ИНИР, ориентированная не на программистов, а непосредственно на специалистов в областях знаний, для которых такие ресурсы разрабатываются, т.е. экспертов. Эта технология является развитием технологии построения порталов научных знаний, разработанной нами раньше и успешно применявшейся при создании научных интернет-ресурсов для многих предметных областей [1].

1. Тематический ИНИР

Тематический ИНИР представляет собой доступную через Интернет информационную систему, обеспечивающую систематизацию и интеграцию научных знаний и информационных ресурсов определенной области знаний, содержательный эффективный доступ к ним (поиск и навигацию) и поддерживающую их использование при решении различных научных и производственных задач за счет предоставления соответствующих интерфейсов и сервисов.

ИНИР позволяет исследователям значительно сократить время, требуемое для обеспечения доступа к необходимой информации и ее анализа, за счет аккумуляции описаний релевантных интернет-ресурсов (в том числе, web-сервисов) непосредственно в контенте ИНИР.

1.1. Система знаний ИНИР. Ядром системы знаний ИНИР является онтология [2], которая наряду с описанием моделируемой области знаний содержит соотнесенное с ним описание структуры и типологии интегрируемых информационных ресурсов и методов интеллектуальной обработки содержащихся в них данных.

Семантическая сеть [3], структура которой определяется онтологией ИНИР, играет роль интеллектуального хранилища данных, в котором накапливается информация о релевантных научных информационных ресурсах и web-сервисах, реализующих методы обработки содержащихся в них знаний и данных.

Кроме онтологии и семантической сети система знаний ИНИР включает тезаурус, который содержит термины моделируемой области знаний, т.е. слова и словосочетания, с помощью которых понятия онтологии представляются в текстах и пользовательских запросах. Тезаурус также задает смысл понятий, причем не столько с помощью определений, сколько посредством соотнесения одних понятий с другими понятиями, используя для этого семантические отношения. Благодаря этому тезаурус может применяться как при обработке пользовательских запросов, так и при поиске и аннотировании информационных ресурсов, интегрируемых в ИНИР.

На основе онтологии и семантической сети организуется удобная навигация по научным знаниям и информационным ресурсам, интегрированным в ИНИР, а также содержательный поиск требующихся данных и средств их интеллектуальной обработки.

Формально система знаний ИНИР описывается четверкой:

$KS = \langle O, Th, SN, IRs \rangle$, где

O – онтология ИНИР,

Th – тезаурус области знаний ИНИР,

SN – семантическая сеть, служащая для представления информационного содержания (контента) ИНИР,

IRs – интегрируемые в ИНИР информационные ресурсы и средства их интеллектуальной обработки (web-сервисы).

1.2. Архитектура ИНИР. ИНИР имеет традиционную для информационных систем трехуровневую архитектуру (см. рис. 1), включающую уровень доступа к информации, уровень обработки информации и уровень хранения информации (базовый уровень).

Рис.1. Архитектура тематического ИНИР.

Первый уровень обеспечивается пользовательским интерфейсом, главными функциями которого являются представление пользовательских запросов и результатов поиска и решений задач, а также обеспечение управляемой онтологией навигации в информационном пространстве ИНИР.

На уровне обработки информации обеспечиваются все информационные потоки в ИНИР – от конструирования онтологии и управления контентом ИНИР до обработки пользовательских запросов. Он включает модуль поиска информации в контенте ИНИР, средства аналитической обработки найденной информации. средства разработки/редактирования онтологии, тезауруса и управления контентом ИНИР, а также подсистему сбора онтологической информации (метаданных) об интернет-ресурсах.

Базовый уровень обеспечивает выполнение функций хранения и манипулирования данными (контентом ИНИР) и знаниями (онтологией и тезаурусом) ИНИР с использованием средств стандартных СУБД, технологий Semantic Web и семантических web-сервисов [4].

2. Технология построения ИНИР

В настоящее время разрабатывается технология построения ИНИР. Ее особенностью является ориентация на широкой круг пользователей – экспертов, т.е. специалистов в определенных областях знаний. Это обусловлено тем, что обеспечить массовое производство ИНИР для различных областей знаний можно только путем привлечения к их разработке самих специалистов, для которых разрабатываются такие ресурсы. Такая технология позволяет им собирать и систематизировать в рамках единого информационного пространства обширные знания и данные в требуемой области знаний, а также средства их интеллектуальной обработки.

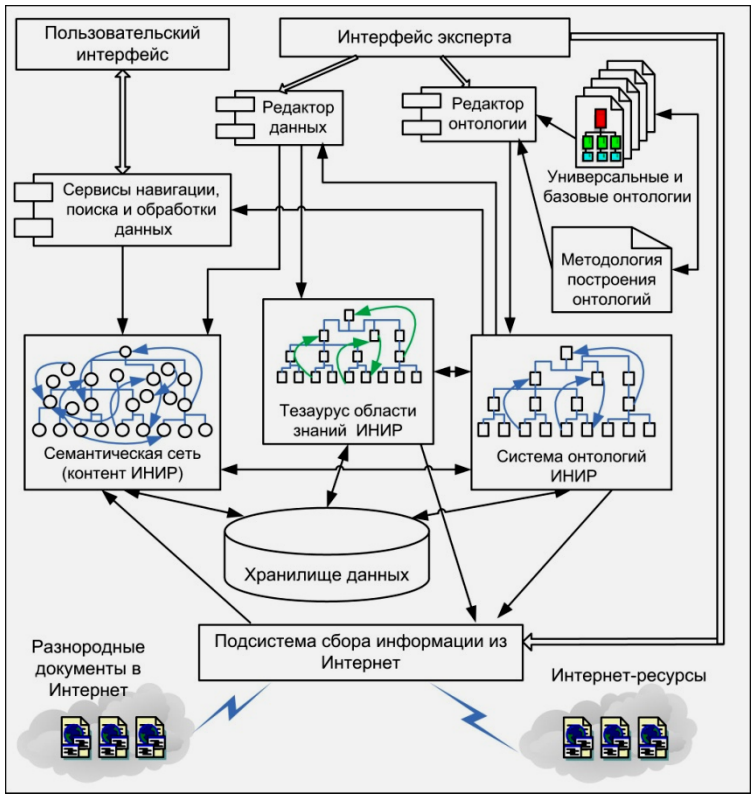


Рис.2. Средства технологической поддержки создания тематического ИНИР.

Основными элементами данной технологии являются (см. рис.2):

(1) методология построения онтологий вместе с набором универсальных и базовых онтологий,

(2) интерфейс эксперта, обеспечивающий доступ к программным средствам, поддерживающим построение онтологий, тезаурусов и управление контентом ИНИР,

(3) подсистема сбора онтологической информации об интернет-ресурсах,

(4) пользовательский интерфейс, обеспечивающий содержательный доступ к контенту ИНИР и средствам аналитической обработки информации,

(5) хранилище данных, обеспечивающее универсальные структуры для согласованного хранения онтологии и контента ИНИР.

2.1. Методология построения онтологии ИНИР. Наиболее важным компонентом рассматриваемой технологии является методология построения онтологии, так как онтология составляет основу тематического ИНИР. Рассмотрим ее подробнее.

Онтология конкретного ИНИР строится в соответствии с методологией, главными принципами которой являются:

- структурирование онтологии ИНИР на ряд относительно независимых онтологий;
- построение всех онтологий ИНИР на основе базовых онтологий путем их достройки и развития.

Использование такой методологии значительно упрощает создание онтологии ИНИР и ее дальнейшее сопровождение.

Как было сказано выше, онтология ИНИР кроме описания понятий и отношений моделируемой области знаний включает соотнесенное с ним описание структуры и типологии интегрируемых информационных ресурсов и методов интеллектуальной обработки содержащихся в них данных и знаний. В связи с этим онтология состоит из трех взаимосвязанных онтологий, отвечающих за представление указанных выше компонентов знаний,

а именно: онтологии области знаний ИНИР, онтологии научных информационных ресурсов и онтологии задач и методов.

В качестве базовых онтологий предложены онтология научной деятельности и онтология научного знания, на основе которых строится онтология области знаний ИНИР, базовые онтологии научных информационных ресурсов и задач и методов, а также онтология представления тезауруса

Следует заметить, что все базовые онтологии не зависят от области знаний ИНИР.

Онтология научного знания содержит классы, задающие структуры для описания понятий конкретных областей знаний, такие как *Раздел науки, Метод исследования, Объект исследования, Научный результат* и др. Эта онтология также включает отношения, связывающие между собой объекты указанных выше классов.

Онтология научной деятельности базируется на онтологии, предложенной в [5] для описания научно-исследовательских проектов и расширенной для применения к более широкому классу задач. Эта онтология включает классы понятий, относящиеся к организации научной и исследовательской деятельности, такие как *Персона, Организация, Событие, Научная деятельность, Проект, Публикация* и др.

Онтология научной деятельности включает также отношения, связывающие между собой как понятия данной онтологии между собой, так и с понятиями онтологии научного знаний. Выбор этих отношений осуществлялся не только исходя из полноты представления области знаний ИНИР, но и из удобства навигации по его информационному пространству и поиска информации.

Базовая онтология научных информационных ресурсов включает класс *Информационный ресурс* в качестве основного класса. Этот класс служит для описания, релевантных области знаний информационных ресурсов (в том числе, представленных в сети

Интернет). Набор атрибутов и связей этого класса основан на стандарте Dublin core [6]. Его атрибутами являются: *название ресурса, язык ресурса, тематика ресурса, тип доступа к ресурсу* и т.п. Объекты этого класса могут быть связаны семантическими отношениями с другими информационными объектами, представляющими в контенте ИНИР организации, персоны, публикации, события, разделы науки и т.д.

Базовая онтология задач и методов включает такие классы как *Задача, Метод* и *Web-сервис*, а также отношения, связывающие эти понятия между собой и понятиями других базовых онтологий. С помощью понятий и отношений данной онтологии могут быть описаны задачи, для решения которых предназначен ИНИР, методы их решения, а также реализующие их web-сервисы

В онтологии задач и методов, как правило, описываются web-сервисы, которые реализуют методы обработки информации, содержащейся в интегрируемых в ИНИР информационных ресурсах. При этом описания web-сервисов базируются на онтологии OWL-S [7], предназначенной для описания семантических web-сервисов. Благодаря этому с web-сервисом связываются не только описание его интерфейса в терминах типов входных и выходных данных, но и описание его семантики, т.е. того, что сервис делает, его предметной области, ограничений на область применения и качество сервиса и т.п. Причем все его свойства, функциональность и интерфейсы кодируются в однозначной поддающейся машиной обработке форме.

Наличие семантического описания у web-сервисов обеспечивает не только реализацию их поиска и корректного использования (исполнения), но и возможность композиции из них новых сервисов с целью получения функциональности, требуемой для решения пользовательских задач. Кроме того, наличие содержательных описаний у web-сервисов создает предпосылки и для их успешной интеграции в ИНИР. При этом будет обеспечиваться

содержательный доступ к ним не только для программных агентов, но и для человека, желающего найти необходимые для решения его задач средства интеллектуальной обработки информации.

Онтология представления тезауруса [8] включает набор базовых понятий и отношений, присутствующих в любом тезаурусе. В частности, она содержит классы, описывающие следующие сущности тезауруса: термины, которые подразделяются на дескрипторы (предпочтительные термины) и аскрипторы (текстовые входы, которые при поиске и индексировании документов могут быть заменены на соответствующие дескрипторы), источники терминов (web-ресурсы, текстовые документы или коллекции текстов, в которых встречаются или определяются термины) и области/подобласти знаний, с которыми могут быть соотнесены термины. В онтологии также представлены отношения, связывающие объекты перечисленных выше классов между собой.

2.2. Управление системой знаний ИНИР. Для поддержки процесса настройки и управления системой знаний ИНИР технология предоставляет редакторы онтологий и данных. Эти редакторы реализованы как web-приложения, поэтому обеспечивают удаленную настройку и поддержку системы знаний ИНИР авторизованными экспертами через Интернет.

Для построения онтологий и управления ими служит редактор онтологий. Этот редактор проектировался таким образом, чтобы им могли пользоваться не только инженеры знаний, но и эксперты, не являющихся специалистами в области информатики и математики.

Управление информационным контентом ИНИР осуществляется с помощью редактора данных. Этот редактор работает под управлением онтологии, что позволяет не только значительно

облегчить ввод данных, но и обеспечить их логическую целостность.

Тезаурус области знаний ИНИР строится на базе ядра тезауруса, построенного на основе онтологии представления тезауруса и изначально включенного в систему знаний ИНИР. Ядро тезауруса содержит описание понятий базовых онтологий, включая описание терминов, с помощью которых они представляются в интернет-ресурсах.

Редактирование содержания тезауруса осуществляется с помощью редактора данных, работающего под управлением онтологии представления тезауруса, что позволяет обеспечить логическую целостность его терминологической системы.

Для того чтобы ИНИР был полезным ресурсом его система знаний должна содержать достаточно полную информацию о моделируемой области знаний и выполняемой в рамках ее научной и/или производственной деятельности. Создание такого ИНИР – довольно трудоемкая задача, требующая значительных усилий разработчиков. Для ее автоматизации разрабатывается подсистема сбора онтологической информации об интернет-ресурсах.

Сбор информации для ИНИР предполагает поиск релевантных области знаний ИНИР интернет-ресурсов и документов, извлечение информации из найденных интернет-ресурсов и документов и занесение полученной информации в контент ИНИР. В соответствии с этим подсистема сбора информации включает модуль поиска релевантных интернет-ресурсов, модуль извлечения информации из интернет-ресурсов, модуль занесения найденной информации в контент ИНИР, а также базу данных ссылок на интернет-ресурсы (БД СИР).

При настройке ИНИР на область знаний выполняется заполнение БД СИР ссылками на релевантные, по мнению экспертов, интернет-ресурсы. При этом для каждой ссылки указывается класс онтологии, объекты которого описывает соответствующий ей ресурс.

Список ссылок может пополняться не только вручную, но и автоматически – модулем поиска интернет-ресурсов, который выполняет сбор ссылок на релевантные интернет-ресурсы по поисковым запросам, сформированным на основе названий классов онтологии и терминов тезауруса, представляющих понятия моделируемой области знаний. Он запускается с заданной при настройке ИНИР периодичностью. При этом модуль поиска обращается к поисковым системам Google, Яндекс и Bing через их программные интерфейсы, т.е. использует механизм метапоиска с последующей фильтрацией дубликатов и нерелевантных ссылок [9].

Для заполнения контента ИНИР собирается информация из таких источников, как порталы знаний, электронные библиотеки и журналы, сайты организаций, ассоциаций, проектов и конференций, новостные ленты, социальные научные сети, вики-ресурсы, реестры (каталоги) веб-сервисов и др. Как было сказано выше, из этих источников извлекается информация о проектах, организациях, персонах, конференциях и публикациях, т.е. обо всех объектах базовых классов онтологии научной деятельности, а также об объектах класса *Информационный ресурс* онтологии научных информационных ресурсов.

Для каждого из этих классов создается свой метод извлечения информации, включающий набор шаблонов. В шаблонах для каждого типа извлекаемой информации указываются обработчики, реализующие алгоритмы обхода и анализа соответствующих фрагментов интернет-страниц или документов. Указанные шаблоны генерируются на основе онтологии. Для повышения полноты извлечения информации увеличивается вариативность этих шаблонов за счет использования в них альтернативных терминов из тезауруса (синонимов и гипонимов).

В настоящее время реализован ряд компонентов подсистемы сбора информации из сети Интернет, а именно: модуль поиска релевантных интернет-ресурсов, модуль извлечения информации,

база данных ссылок на интернет-ресурсы. На данный момент разработан метод извлечения данных о проекте, включая сопутствующие ему шаблоны и обработчики, реализующие извлечение информации о персонах и публикациях.

2.3. Пути развития технологии построения ИНИР. Использование онтологии в качестве основы ИНИР, создает предпосылки для того, чтобы технология построения ИНИР стала действительно массовой. С одной стороны, онтология является удобным средством формирования и фиксации общего разделяемого экспертами-разработчиками знания о данной предметной области, обеспечивая при этом возможность переиспользования знаний, что упрощает и ускоряет разработку новых приложений. С другой стороны, базирование средств описания области знаний ИНИР, как и создания и сопровождения его контента, на онтологии делает их доступными для использования непосредственно экспертами, так как представление знаний и данных в виде объектов и отношений между ними, принятое в онтологии, является наиболее естественным для человека.

Чтобы данной технологией мог воспользоваться широкий круг экспертов, в ее рамках должны быть разработаны программные оболочки разных типов ИНИР, отличающиеся набором базовых онтологий и, возможно, программных компонентов. Такие специализированные оболочки будут представлять собой «пустые» ИНИР, т.е. в них будут представлены все необходимые структурные компоненты будущего ИНИР, но не достроены нижние уровни онтологии области знаний и не заполнен контент.

Примером специализированной оболочки является оболочка портала научных знаний [10], включающая рассмотренные выше базовые онтологии – онтологию научной деятельности и онтологию научного знания, с помощью которых эксперт может, не прибегая к помощи инженеров знаний и программистов, построить онтологию требуемой области знаний.

В качестве другого примера можно привести оболочку для построения многоязычных тезаурусов [8], включающую онтологию представления тезауруса, содержащую набор базовых понятий и отношений, присутствующих в любом тезаурусе.

Была создана еще одна специализированная оболочка, предназначенная для разработки ИНИР, ориентированных на области знаний, в которых используются математические методы. Эта оболочка дополнительно включает редактор математических формул и веб-сервис, реализующий методы математического программирования.

Заключение

В докладе представлен подход к построению тематических ИНИР, обеспечивающих систематизацию и интеграцию информационных ресурсов определенной области знаний и средств интеллектуальной обработки содержащейся в них информации, а также содержательный эффективный доступ к ним и их использование при решении различных задач.

Основу ИНИР составляет онтология. Семантическая сеть, структура которой определяется онтологией, играет роль интеллектуального хранилища данных, в котором накапливается информация о релевантных научных информационных ресурсах и web-сервисах, реализующих интеллектуальную обработку содержащейся в них информации.

Важное преимущество подхода: ИНИР позволяет упростить и ускорить доступ к затребованной пользователем информации и сократить время ее анализа благодаря аккумуляции описаний релевантных интернет-ресурсов и методов их обработки непосредственно в семантической сети ИНИР.

Использование онтологии в качестве основы ИНИР, создает предпосылки для того, чтобы технология построения ИНИР стала действительно массовой.

Чтобы данной технологией мог воспользоваться широкий круг экспертов, в ее рамках разрабатываются программные оболочки

разных типов ИНИР, отличающиеся набором базовых онтологий и небольшой части программных компонентов.

В настоящее время разработаны и используются три специализированные оболочки: (1) оболочка портала научных знаний, (2) оболочка многоязычного тезауруса и (3) оболочка ИНИР, ориентированных на области знаний, использующих математические методы вычислений.

С помощью первой оболочки были построены научные интернет-порталы по археологии [11] и компьютерной лингвистике [12], выполнявшихся на базе ИСИ СО РАН. Целью этих проектов было обеспечение содержательного доступа к систематизированным знаниям и информационным ресурсам соответствующей области научных знаний.

Пользователи созданных порталов могут получить представление не только о моделируемой области знаний в целом, но и найти информацию о выполняемой в ней научной деятельности. В первую очередь это информация об ученых, организациях, исследовательских группах и их деятельности. Важным компонентом информационного контента таких систем является описание интернет-ресурсов: сайтов организаций, проектов, конференций, тематических порталов и каталогов, а также отдельных страниц с материалами графического, мультимедийного или текстового типа.

Вторая оболочка применялась для создания русско-английского тезауруса по компьютерной лингвистике [13].

На основе третьей оболочки был построен портал знаний для информационно-аналитической поддержки разработчиков СППР.

Работа выполнена при финансовой поддержке РФФИ (проект № 13-07-00422).

Литература

1. Загорулько Ю.А. Технология построения порталов научных знаний: опыт применения, проблемы и перспективы // Труды 21-й Международной Крымской конференции «СВЧ-техника и телекоммуникационные технологии» – КрыМиКо-2011 –Севастополь, Крым, Украина, изд. Севастополь: Вебер, 2011. -Т.1. -С.51-54.
2. Guarino N. Formal Ontology in Information Systems // Proceedings of FOIS'98 (Trento, Italy, 1998). Amsterdam: IOS Press, 1998. pp.. 3-15.
3. Осипов Г.С. Построение моделей предметных областей. Неоднородные семантические сети // Известия АН СССР. Техническая кибернетика. –1990. – №5. – с. 32–45.
4. McIlraith S.A., Son T.C., Zeng H. Semantic Web Services. Intelligent Systems, IEEE, 2001, 16(2). pp. 46-53.
5. Benjamins V.R. and Fensel D. Community is Knowl-edge! in (KA)2 . Proc. of 11th Banff Knowledge Acquisition for Knowledge-based Systems workshop, KAW'98 (Banff, Canada, April 1998). – Calgary: SRDG Publications, Department of Computer Science, University of Calgary, 1998.
6. Hillmann D. Using Dublin Core. Available at: http://dublincore.org/documents/usageguide/ (accessed 8 May 2014).
7. OWL-S: Semantic Markup for Web Services. Available at: http://www.w3.org/Submission/OWL-S/ (accessed 8 May 2014).
8. Загорулько Ю.А., Боровикова О.И. Программная оболочка для построения многоязычных тезаурусов предметных областей, ориентированная на экспертов // Труды 13-й национальной конференции по искусственному интеллекту с международным участием КИИ-2012. – Белгород: Изд-во БГТУ, 2012. -Т.4. -С. 76-83.
9. Ахмадеева И.Р., Загорулько Ю.А., Саломатина Н.В., Серый А.С., Сидорова Е.А., Шестаков В.К. Подход к формированию тематических коллекций текстов на основе интернет-ресурсов // Вестник НГУ. Серия: Информационные технологии. 2013. Том.11, выпуск 4.
10. Загорулько Ю.А., Боровикова О.И. Подход к построению порталов научных знаний // Автометрия. № 1, 2008, т. 44. –с. 100–110.
11. Андреева О.А., Боровикова О.И., Булгаков С.В. и др. Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии // Тр. 10-й национальной конференции по искусственному интеллекту с международным уча-

стием КИИ-2006 (25–28 сентября 2006 г., Обнинск). М.: Физматлит, 2006. Т. 3. С. 832–840.
12. Боровикова О.И., Загорулько Ю.А., Загорулько Г.Б. и др. Разработка портала знаний по компьютерной лингвистике // Тр. 11-й национальной конференции по искусственному интеллекту с международным участием КИИ-2008 (Дубна, 2008 г.). М.: ЛЕНАНД, 2008. Т. 3. С. 380–388.
13. Загорулько Ю.А., Боровикова О.И., Кононенко И.С., Соколова Е.Г. Подход к разработке русско-английского тезауруса по компьютерной лингвистике // Труды XIII Всерос. науч. конференции RCDL'2011 «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Воронеж, 19-22 октября 2011 г. – Воронеж: Издательско-полиграфический центр Воронежского гос. университета, 2011. – С.27–34.

ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ СИСТЕМЫ ИНТЕГРАЦИИ ТЕПЛОФИЗИЧЕСКИХ ДАННЫХ НА ОСНОВЕ ОНТОЛОГИЧЕСКОЙ МОДЕЛИ ПРЕДМЕТНОЙ ОБЛАСТИ[1]

В.А. Серебряков[a], *К.Б. Теймуразов*[a], *Р.И. Хайруллин*[b], *А.О. Еркимбаев*[c], *В.Ю. Зицерман*[c], *Г.А. Кобзев*[c], *М.С. Трахтенгерц*[c]

[a]Вычислительный центр им. А.А. Дородницына РАН, [b]МФТИ, [c]ОИВТ РАН

Выполненная работа завершает определенный этап в решении достаточно сложной задачи интеграции теплофизических данных. Анализ предметной области позволил «отстроить» детализированную онтологию для задач, выходящих за рамки возможностей традиционных БД. Формализуя предметную область, онтология обеспечивает строгое и унифицированное описание терминов и понятий. В рамках этой системы онтология позволяет поддерживать все виды логических и математических связей, например связей между термодинамическими функциями или ограничений в отнесении тех или иных свойств в определенных фазовых состояниях. Наконец, применительно к целям интеграции разнородных данных важную роль играет возможность пользователя наращивать списки веществ, свойств, единиц измерения и прочих элементов набора данных.

Ключевые слова: Интеграция данных, онтология, теплофизика, свойства веществ, связанные открытые данные.

Performed work completes a definite step in solving a rather complex task of integrating thermal data. Domain analysis allowed the "rebuild" a detailed ontology for tasks that go beyond the capabilities of traditional database. Formalizing the domain, the ontology provides a rigorous and unified description of the terms and concepts. Under this system, the ontology allows

[1] Работа выполнена при поддержке РФФИ, грант №13-07-00218.

to support all kinds of logical and mathematical relationships, such as relationships between the thermodynamic functions or constraints in assigning of certain properties in certain phase states. Finally, for the purposes of integrating heterogeneous data plays an important role the user's ability to build lists of substances, properties, units and other elements of the data set.

Keywords: data integration, ontology, thermal physics, material properties, linked open data.

1. Интеграция данных и роль онтологий

В предыдущих работах авторов [1, 2] был выявлен потенциал *Semantic Web* в задачах интеграции научных данных и, прежде всего, данных по теплофизическим свойствам веществ, к которым относят сжимаемость, энергетические и транспортные характеристики, а также данные, определяющие химическое и фазовое равновесие, диаграмму состояний и т.п. Ключевая роль теплофизических свойств в объяснении и расчете множества природных и технологических процессов определяют крайнюю актуальность процедур распространения и обмена данными между многочисленными гетерогенными источниками, в роли которых выступают базы данных (БД), информационно-вычислительные системы, электронные издания. В самом общем виде процесс интеграции гетерогенных источников предусматривает их соединение в рамках унифицированного представления и в терминах единой модели данных. В работе [3] на многих примерах было показано, что одним из наиболее эффективных подходов в задаче интеграции, особенно применительно к данным по свойствам вещества, является **онтологическое моделирование**, посредством которого можно адекватно формализовать понятийный багаж предметной области, обеспечив одновременно доступ к словарям и онтологиям общенаучного содержания, во множестве представленных в сети Интернет. Рассматривая онтологию как базовый элемент *Semantic Web*, можно публиковать стандартизо-

ванные наборы данных, обеспечивая возможность их связывания с тематически родственными ресурсами среды Linked Open Data (LOD).

Одно из многих определений онтологии в сфере компьютерных наук трактует ее как семантически точное и машинно-обрабатываемое определение *вещей* и их связей. Из этого определения следует, что посредством онтологического моделирования можно достичь автоматической интеграции данных и реализовать процедуры логических рассуждений, контролирующих достоверность данных и их соответствие физическим принципам.

В рамках данной работы решается более ограниченная задача – строится онтология для выбранной предметной области («теплофизические свойства веществ»), гарантирующая формализацию основных понятий и выполнение связей и ограничений, обусловленных физическими принципами. Задача состоит в том, чтобы в согласии с онтологией обеспечить проектирование БД, способной хранить и интегрировать данные из разнородных источников, контролируя выполнение требуемых логических и математических ограничений. Работа предваряет создание системы публикации в LOD, поскольку в существующей технологии именно реляционные данные используются как первичный массив, конвертируемый в RDF-формат, необходимый при связывании тематически родственных данных.

2. Теплофизические данные – особенности, методы распространения и интеграции

По многим причинам в теплофизике ключевую роль играет работа с численными данными, включая их накопление, обработку и систематизацию. Повышенное внимание к первичным данным с детальным изучением их достоверности, согласованности, воспроизводимости повторяемых экспериментов связано с ограниченными возможностями теории обеспечить априорное про-

гнозирование свойств и закономерностей. Как следствие, развитие теплофизики сопровождает нарастающий масштаб производства новых данных, публикуемых в десятках журналов различного профиля: физического, химического, инженерного и др., как например, J. Chem. & Engn. Data, J. Chem. Thermodynamics, Fluid Phase Equilibria. По данным [4] за десятилетний период (1998 – 2007 гг.) объем данных по теплофизическим свойствам, опубликованных в этих журналах, вырос втрое. Издавна в теплофизике организовывались проекты создания многотомных справочников, авторы которых формировали коллекции критически оцененных данных, компилируя и обрабатывая сотни ранее опубликованных источников.

Современный этап работ в теплофизике характеризуется повсеместным переходом от печатной формы справочников к компьютерным БД. Соответственно методы обмена неоднородными данными, различающимися форматом и структурой, возникла в теплофизике задолго до того момента, когда проблема интеграции приобрела актуальность для информационного сообщества. Так один из первых стандартов обмена термодинамическими данными, получивший название COSTAT (Codata STAndard Thermodynamics), был разработан Термодинамическим Исследовательским Центром США в течение 1985-1987 гг. под эгидой Международной комиссии по численным данным (CODATA) [5].

Как и в других областях знания, интеграция данных призвана преодолеть проблемы, связанные с многообразием стилей и форматов представления данных в отсутствие общепринятых рецептур и стандартов записи [2]. В дополнение к этому, теплофизике присущи и внутренние причины различий структуры данных в разных источниках: разнообразие используемых физических моделей (например, уравнений состояния или моделей растворов); зависимость описания и набора характеристик от диапазона параметров; различия в способах представления данных и методах оценки неопределенности и проч. Среди наиболее известных примеров – разброс в использовании начал отсчета, температурных шкал, единиц измерения, зависимость номенклатуры пара-

метров от принятой модели. Разрабатываемая здесь онтология позволяет в определенной мере сгладить различия в использовании моделей и терминов. При этом, важный элемент, присущий онтологическому моделированию – возможность наращивания (по мере расширения предметной области) новых понятий, в основном за счет расширения перечня веществ и номенклатуры свойств.

Учитывая крайнее многообразие веществ и свойств, для которых создается онтология, формализующая понятия и ограничения в их использовании, имеет смысл провести по некоторым критериям ее «сужение» с целью отработки методов концептуализации и программных средств. Выбор этих критериев соответствовал тем ограничениям, которые много лет назад (в 1973 г.) были приняты Теплофизическим Центром ОИВТ РАН [6] при создании государственного информационного фонда: чистые (однокомпонентные) вещества; преимущественная ориентация на неорганические вещества, включая нестехиометрические соединения; сужение круга органических соединений веществами, содержащими группы не более, чем из двух атомов углерода (простейшими углеводородами, фреонами и т.п.); отказ от рассмотрения материалов, свойства которых зависят от способа получения и метода обработки. Нетрудно видеть, что эти ограничения, помимо сокращения списка веществ, упрощают правила их идентификации. При отказе от рассмотрения материалов, смесей, растворов основным дескриптором является стехиометрическая формула (H2O, CO2, CH4 …), дополненная перечнем тривиальных или номенклатурных названий, а отказ от включения в фонды сложной «органики» позволяет обойтись без нотаций, кодирующих структуру и топологию многоатомной молекулы (например, The IUPAC International Chemical Identifier, InChI - [7]). Помимо упрощения идентификации, исключение смесей позволяет сократить число независимых переменных, исключив из рассмотрения концентрации.

При определенных условиях можно пойти еще на одно ограничение, а именно исключить из параметров состояния одну из

переменных, давление. Это связано с тем, что во множестве практических задач барическая зависимость свойств проявляется слабо. Прежде всего, это относится к твердой и жидкой фазе, если последняя удалена от критической области. Что касается газовой фазы, часто для учета неидеальности допустимо ограничиться вириальным разложением по плотности, при том, что сами вириальные коэффициенты являются функциями температуры. Точно также и транспортные свойства, такие как вязкость или теплопроводность слабо зависят от плотности или давления. Никоим образом, не предполагая универсальность этого положения, можно принять, что подавляющее большинство публикуемых или компилируемых теплофизических данных сводится к представлению некоторого числа температурных функций, как например, теплоемкость, энтальпия, вязкость и т.п., а также численных констант, например критических постоянных вещества или энтальпий их образования.

Таким образом, если ввести указанные ограничения при осознанном сужении предметной области, мы приходим к обозримой задаче – интеграции относительно однотипных данных, представимых константами и функциями одной переменной, с доминированием табличной формы передачи данных. При таком сужении предметной области задача интеграции данных делается относительно обозримой, хотя и здесь приходится учитывать крайнее многообразие вариантов, связанное с идентификацией вещества, принимаемой номенклатурой свойств, фазовым многообразием и т. п.

Построение онтологии должно упростить принятие некоторых дополнительных соглашений о форме представления данных, подлежащих хранению и распространению. Предполагается, что общий поток данных разбивается на порции, называемые наборами данных. Каждый набор содержит данные для одного вещества и произвольного комплекса свойств. Набор свойств включает несколько констант и несколько функций одной переменной, как правило, температуры. Все функции заданы для одних и тех же значений независимой переменной. Набор данных включает

также сведения: о фазовом состоянии вещества, единицах измерений свойств, неопределенности и источнике данных. Существенно при этом, что основные списки - веществ, свойств, фазовых состояний, единиц измерений и т.д. считаются открытыми, что позволит в рамках той же онтологии обеспечить подстройку под новые типы данных, если они удовлетворяют принятым ограничениям.

3. Концептуализация предметной области

Построению онтологии с ее записью на языке OWL предшествует так называемая *концептуализация*, то есть строгое описание понятий предметной области, их связей и отношений средствами естественного языка. Для предметной области, выделенной в разделе 2 с учетом принятых там ограничений, в основу концептуализации следует положить два понятия: имя/название вещества и названия тех величин, которые представляют значения свойств. Используя их, упрощенную схему представления данных можно составить из трех элементов: (1) перечень/словарь веществ; (2) перечень/словарь свойств и (3) формат/шаблон для представления набора численных данных. С каждым элементом из списка веществ связаны одно или несколько названий и формульное обозначение. С каждым элементом из списка свойств связаны его название, обозначение и единица измерения. Шаблон данных для определенного документа фиксирует: название вещества, для которого представлены данные по свойствам; названия свойств, рассматриваемых в данном контексте как константы; названия свойств, рассматриваемых в данном контексте как функции.

Первые два элемента (перечни веществ и свойств) могут иметь древовидный (иерархический) характер. Например, в перечне веществ возможно выделение таких классов как элементы, окислы, гидриды и т.п. Определенные сложности возникают и с названием веществ. Всегда есть неопределенность в выборе назва-

ния из некоторого множества общеупотребительных; скажем простейшей формуле O_2 соответствуют такие названия как кислород, дикислород, oxygen. Из этого следует, что под **названием** вещества в общем случае подразумевается блок **идентификации**, включающий набор терминов или классификационных номеров (например, CAS Number [8]), допускающих однозначное определение вещества с учетом особенностей его состава и структуры.

Схема, включающая три элемента (вещества, свойства, шаблон данных), заметно упрощена в сопоставлении с реальными данными. Дополнительно для веществ указывается состояние, в котором оно находится. Выделяются агрегатные состояния (**solid, liquid, gas**), причем отдельно указывается состояние идеального газа. Кроме того, для состояния **solid** детализируется кристаллическая фаза, например кубическая, тетрагональная, гексагональная и др. Вещество может находиться также в состояниях, отвечающих пограничным линиям (**solid-liquid, solid-gas, liquid-gas**), а также окрестности критической точки (**critical state**). Указание в документе на состояние, в котором находится вещество, накладывает определенные связи и ограничения при выборе свойств, которые привлекаются для его характеристики. Например, такое свойство как **viscosity** может использоваться лишь для веществ, находящихся в состояниях **liquid** или **gas.** Специально выделяются также те данные, которые относятся только к межфазным границам, например **enthalpy of vaporization** (энтальпия испарения) по физическому смыслу характеризует вещество, находящееся только на линии **liquid-gas**.

Шаблон данных включает текстовой фрагмент с определением вещества, названия свойств в виде констант и названия свойств в виде функций. Помимо этого, он должен содержать указания на состояние, в котором находится вещество, и характеристики окружения (environment). К ним относятся, прежде всего, то свой-

ство, которое является аргументом функций и те дополнительные характеристики, которые принимают фиксированные значения. Пример типовой формы документа показан на рис. 1 и 2, где приведены записи в БД ИВТАНТЕРМО по термодинамическим свойствам индивидуальных веществ. Первый фрагмент содержит название и формулу вещества, указатель состояния и фиксированного внешнего параметра ($p_{ref} = 1\,\text{atm}$), а также значения большого набора физических констант. Второй фрагмент дает в табличной форме значения 7 функций, выделяет аргумент (температуру) и указывает его значения. Каждое из свойств в наборе данных (константы, аргумент и функции) идет со ссылкой на единицу измерения, а набор в целом имеет некоторую ссылку на неопределенность данных. На рис. 1 это обобщенный показатель достоверности, обозначенный как precision class.

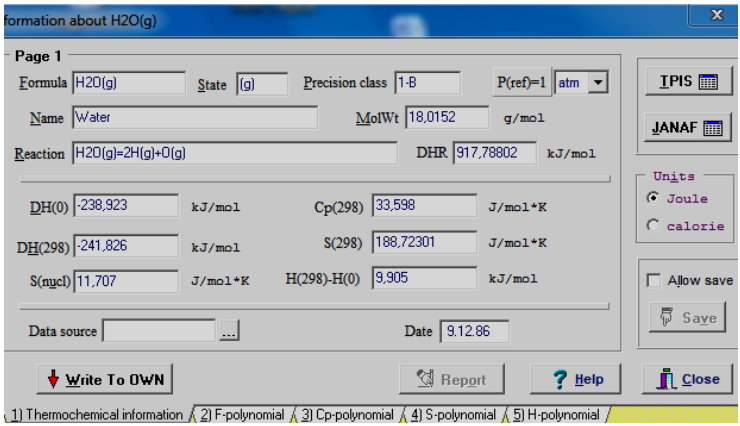
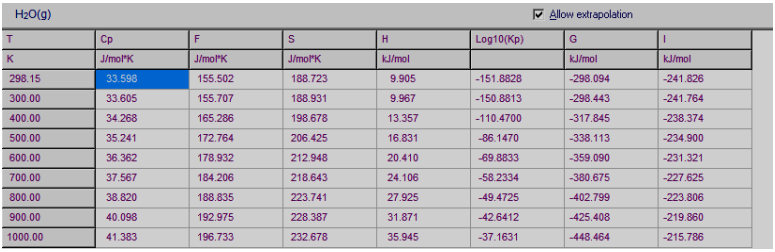


Рис. 1. Типовая выдача констант в БД ИВТАНТЕРМО.



$H_2O(g)$ — Allow extrapolation

T	Cp	F	S	H	Log10(Kp)	G	I
K	J/mol*K	J/mol*K	J/mol*K	kJ/mol		kJ/mol	kJ/mol
298.15	33.598	155.502	188.723	9.905	-151.8828	-298.094	-241.826
300.00	33.605	155.707	188.931	9.967	-150.8813	-298.443	-241.764
400.00	34.268	165.286	198.678	13.357	-110.4700	-317.845	-238.374
500.00	35.241	172.764	206.425	16.831	-86.1470	-338.113	-234.900
600.00	36.362	178.932	212.948	20.410	-69.8833	-359.090	-231.321
700.00	37.567	184.206	218.643	24.106	-58.2334	-380.675	-227.625
800.00	38.820	188.835	223.741	27.925	-49.4725	-402.799	-223.806
900.00	40.098	192.975	228.387	31.871	-42.6412	-425.408	-219.860
1000.00	41.383	196.733	232.678	35.945	-37.1631	-448.464	-215.786

Рис. 2. Типовая выдача функций в БД ИВТАНТЕРМО.

В общем случае оценка неопределенности требует привлечения довольно большой и разнородной информации. Прежде всего, эта оценка может относиться к набору в целом, к отдельным свойствам, включенным в набор или быть отнесена к отдельным числам в таблице, то есть различаться для разных значений аргумента. Наряду с отнесением погрешностей, приходится учитывать и их тип, как например, среднеквадратическая, неопределенность при заданном уровне значимости (обычно 95%), комбинированная неопределенность, учитывающая вклад от неопределенности аргумента, интегральные оценки достоверности (см. рис. 1) и т.п. Соответственно, здесь приходится выделить метаданные, указывающие тип неопределенности, и сами данные – значения неопределенности. Перечень возможных типов неопределенности должен быть представлен в виде словаря, по аналогии с перечнем свойств или единиц измерения.

Еще один важный элемент в представлении данных – информация об источнике, стандартизованная (или нестандартизованная) библиографическая запись. Точно так же, как и в представ-

лении неопределенности, данные об источнике могут относиться к набору в целом, к отдельным свойствам из набора, и в принципе, различаться для отдельных строк в таблице данных. Последнее оправдано, когда в набор включены результаты измерений, поведенных разными авторами при разных значениях аргумента. Простейший способ учесть многовариантность в задании источников – составить их априорный список, включая ссылки на номер источника в соответствующих элементах набора данных. Существенно также, что при ссылке на источник обычно в публикациях или БД указывают и статус представленных в источнике данных, различая данные эксперимента, расчетные данные и справочные (критически оцененные).

Наконец, в определенных случаях набор данных включает сведения об условиях эксперимента, послужившего источником данных, включая данные о методе измерения, подготовке образца и проч. В тех же случаях, когда набор данных получен в результате критического анализа совокупности данных, он может включать и текстовые фрагменты с описанием принципов отбора данных и метода их обработки. Пример записи из БД NIST для кислорода [9] (рис. 3) иллюстрирует широкое аннотирование данных со ссылками на источники и процедуры. Наконец, должен быть сформирован открытый список единиц измерения, предусмотрев необходимость при задании свойства соотнести его с определенной единицей, использованной в источнике данных.

Phase change data

Go To: Top, References, Notes / Error Report

Data compiled as indicated in comments:
TRC - Thermodynamics Research Center, NIST Boulder Laboratories, M. Frenkel director

Quantity	Value	Units	Method	Reference	Comment
T_{boil}	90.2	K	N/A	Streng, 1971	Uncertainty assigned by TRC = 0.2 K; *TRC*
Quantity	**Value**	**Units**	**Method**	**Reference**	**Comment**
T_{fus}	54.8	K	N/A	Streng, 1971	Uncertainty assigned by TRC = 0.2 K; *TRC*
Quantity	**Value**	**Units**	**Method**	**Reference**	**Comment**
T_{triple}	54.33	K	N/A	Henning and Otto, 1936	Uncertainty assigned by TRC = 0.06 K; temperature measured with He gas thermometer; *TRC*
Quantity	**Value**	**Units**	**Method**	**Reference**	**Comment**
T_p	154.58	K	N/A	Pentermann and Wagner, 1978	Uncertainty assigned by TRC = 0.0015 K; *TRC*
T_c	154.58	K	N/A	Wagner, Ewers, et al., 1976	Uncertainty assigned by TRC = 0.0015 K; *TRC*
T_c	155.15	K	N/A	Cardoso, 1915	Uncertainty assigned by TRC = 0.3 K; 4 determinations with same result; *TRC*
Quantity	**Value**	**Units**	**Method**	**Reference**	**Comment**
P_c	50.43	bar	N/A	Wagner, Ewers, et al., 1976	Uncertainty assigned by TRC = 0.005 bar; Vapour-pressure measurements give pc=5.04332 MPa at Tc from L A Weber, 1970 PRT, IPTS-68, PP+ differential pressure transducer.; *TRC*
P_c	50.0343	bar	N/A	Cardoso, 1915	Uncertainty assigned by TRC = 0.3039 bar; *TRC*
P_c	49.9228	bar	N/A	Cardoso, 1915	Uncertainty assigned by TRC = 0.3039 bar; *TRC*
P_c	49.8519	bar	N/A	Cardoso, 1915	Uncertainty assigned by TRC = 0.3039 bar; *TRC*
Quantity	**Value**	**Units**	**Method**	**Reference**	**Comment**
ρ_c	13.60	mol/l	N/A	Pentermann and Wagner, 1978	Uncertainty assigned by TRC = 0.014 mol/l; from density measurements 65 to 300 K, Tc from Weber, 1970; *TRC*

Рис. 3. Запись из БД NIST, определяющие значения свойств на линии равновесия *liquid-gas*

4. Построение онтологии

Вторым этапом построения онтологии является *спецификация* с выделением классов и записью всех связей и отношений на языке OWL. Построенная онтология включает 12 основных классов и 2 класса потомка. Диаграмма на рис. 4 показывает связи классов, после чего приведен перечень классов с указанием назначения и основными атрибутами каждого из классов. Первая четверка классов определяет ключевые понятия для представления набора данных: вещества, свойства, состояния, численные данные. Смысл большей части атрибутов достаточно понятен из их названия. Так атрибут *InConditoins* в классе *Substances* отсылает к перечню состояний и внешних условий, перечни которых даны в соответствующих классах. Использован также ряд атрибутов для таких понятий как единицы измерения и неопределенность.

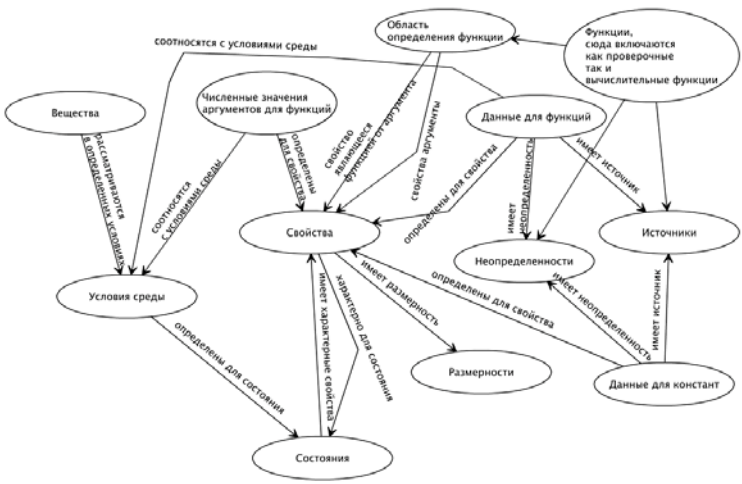


Рис. 4. Классы онтологии и их связи.

Класс *Substances*: Определяет вещества, для которых приводятся данные

Класс *States*: Определяет агрегатные состояния вещества.

Класс *Properties*: Определяет свойства вещества.

Класс *NumericalData*: Определяет набор численных данных для функциональной зависимости свойств вещества в определенных условиях от аргументов; в данной онтологии, свойства зависят от температуры T и давления P.

Substances	
1.	String **Name**
2.	String **Subst_Formula**
3.	EnvironmentalConditions **inConditions**
States	
1.	String **Name**
2.	States **ParentStates**
3.	Properties **inProperty**
Properties	
1.	String **Name**
2.	String **PropDesignation**
3.	States **inState**
4.	Dimensions **Dimension**
NumericalData	
1.	Decimal **Value**
2.	Decimal **UncertaintyValue**
3.	Uncertainties **UncertaintyType**
4.	Properties **Property**
5.	EnvironmentalConditions **inConditions**

Следующая группа классов определяет базовые понятия, сопровождающие физические величины: единицы измерения (размерности), неопределенность данных и внешние условия. К условиям в данной работе относятся и константы, связанные с веществом, например, энтальпия образования при 298.15 К (DHF(298)). С точки зрения онтологии это не совсем корректно, но в целях упрощения схемы принято временное решение об отнесении констант к условиям. Соответственно, класс *EnvironmentConditions* содержит указание на агрегатное состояние вещества и ссылок на набор классов *NumericalData***,** в которых указываются численные значения свойств с заданной погрешностью и условиями, характерными для данной среды.

Класс *ConstantsOfSubstance*: Определяет набор численных значений констант в определенных условиях среды. Хотя классы *ConstantsOfSubstance* и *NumericalData* имеют одинаковую структуру, они кардинально отличаются по содержанию; соответственно с введением этих классов мы разделили содержание *Property* на константы, переменные и функции от этих переменных .

Класс *EnvironmentConditions*: Задает перечень свойств, определяющих условия среды, в которых находится вещество, численные значения свойств вычисляются с помощью класса *NumericalData*. Приводимые далее два класса (*Data, Data_Source*) вводят данные, например, справочную информацию по молекулярным весам, а также сведения о публикациях, откуда приняты наборы данных.

Класс *Uncertainties*: Определяет тип погрешности физической величины.

Класс *Dimensions*: Определяет размерности физических величин.

Класс *Data*: Определяет перечень данных из справочников физических величин.

Класс *Data_Source*: Определяет источники данных для классов *NumericalData*, *Data* и *Functions*.

ConstantsOfSubstance	
1.	Decimal **Value**
2.	Decimal **UncertaintyValue**
3.	Uncertainties **UncertaintyType**
4.	Properties **Property**
5.	EnvironmentConditions **inConditions**
EnvironmentalConditions	
1.	Substances **Substance**
2.	States **State**
3.	Constants_of_Substance **Constants**
Uncertainties	
1.	String **Name**
Dimensions	
1.	String **Name**
Data	
1.	Decimal **Value**
2.	Decimal **UncertaintyValue**
3.	Publication **Datasource**
4.	Properties **inPtroperty**
5.	Uncertainties **Uncertainty**
DataSource	
1.	String **Name**

Наконец, последняя группа состоит из двух основных классов (*Functions*, *DomainOfFunctions*) и двух потомков класса *Functions*: *ControlFunc* и *ComputableFunc*. В совокупности они решают задачу вычисления свойств по формулам при контроле допустимой области изменения аргумента и функций, а также заранее установленных соотношений между различными свойствами, которые в экспериментальных данных выполняются с точностью до некоторой погрешности. Функции делятся на два типа: «вычислительные функции» и «контрольные функции». Вычислительные функции дополнительно содержат указание на вычисляемое свойство, тип и величину погрешности. Результатом вычисления функции является значение свойства, помещаемое в БД. Контрольные функции являются булевскими и отвечают на вопрос, выполняется ли заданное соотношение с при допустимой погрешности или нет.

Класс *DomainOfFunctionDefinition*: Определяет перечень аргументов и ограничений физических свойств для функций. Атрибут *inStates* определяет для каких состояний характерна данная зависимость.

Класс *Functions*: Определяет функции для вычисления и проверки корректности

значений физических величин.

Класс *ControlFunc* (Подкласс класса *Functions*): Определяет перечень проверочных функций, которые определяют, выполняется ли заданное соотношение с при допустимой погрешности.

Класс *ComputingFunc* (Подкласс класса *Functions*): Определяет перечень функций для вычисления значений свойств.

DomainOfFunctionDefinition	
1.	Properties **Calculated_Property**
2.	Properties **Argument_Property**
3.	Decimal **Lower-Range_Of_Definition**
4.	Decimal **Upper-Range_Of_Definition**
5.	Decimal **Lower-Range_Of_Variation**
6.	Decimal **Upper-Range_Of_Variation**
7.	States **inStates**
Functions	
1.	String **FuncFormula**
2.	Publication **Datasource**
3.	DomainOfFunctionDefinition **Domains**
ConrolFunc	
1.	Decimal **RequiredUncertainty**
ComputingFunc	
1.	Decimal **UncertaintyValue**
2.	Uncertainties **Uncertainty**

В итоге, построенная онтология формализует предметную область до уровня, позволяющего для «суженной» предметной области охватить практически все виды представляемых в литературе или компьютерных средах данных по свойствам как в виде таблиц, так и математических выражений. При этом возможно произвольное расширение на новые виды свойств, единицы измерения, способы задания неопределенности и прочие элементы, сопровождающие набор экспериментальных или справочных данных. Окончательная задача – разработка приложения, обеспечивающего экспорт данных из БД или публикаций в форме, соответствующей разработанной онтологии.

5. Программная реализация

На основании построенной онтологии посредством использования стека технологий *Hibernate ORM* и *Spring MVC* генерируется БД в реляционной СУБД *PostgreSQL*. Соответственно 12 классам онтологии создано 12 java классов, которые отображаются посредством *Hibernate* на таблицы реляционной БД данных, схема которой приведена на рис 5.

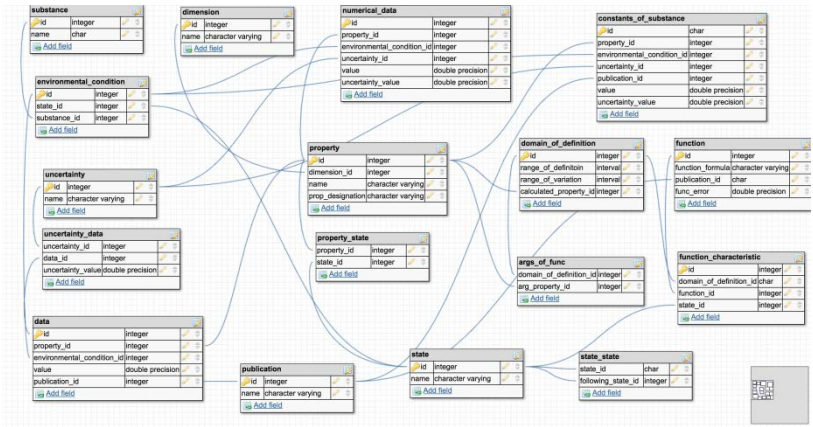


Рис. 5. Схема реляционной БД.

Для системы загрузки данных разработан комплекс, позволяющий анализировать документы и загружать данные из них, причем будет выполняться проверка на соответствие содержания документа онтологической модели и выполнение ограничений, определяемых физическими законами. Логика работы приложения, связанная с вводом исходных данных и проверкой их на соответствие онтологии, показана на рис. 6.

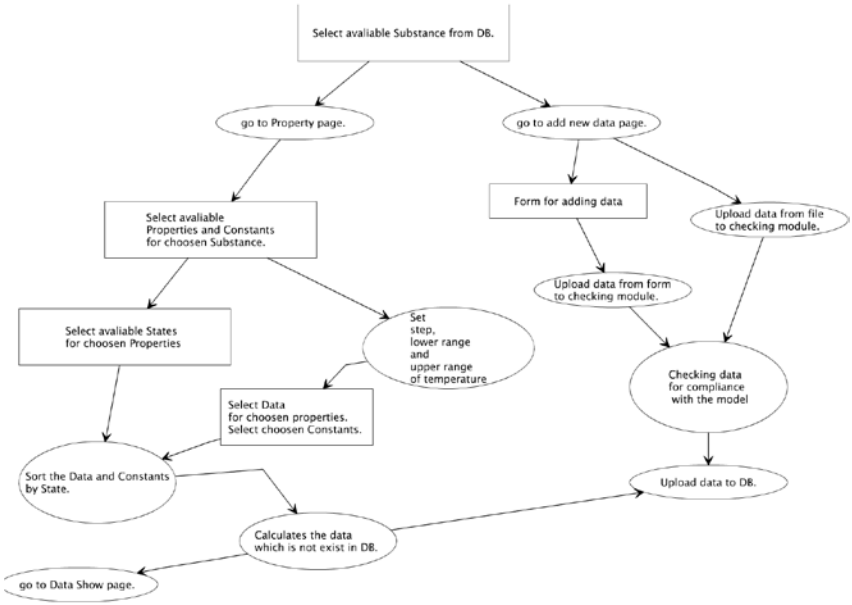


Рис. 6. Логика работы приложения.

На рис. 7–9 показаны отдельные элементы интерфейса для диалога с пользователем: формирование списка веществ (рис. 7), формирование списка свойств с подготовленной формой для выбора определенных свойств и ввода данных (рис. 8), и наконец, сформированный набор данных (рис. 9).

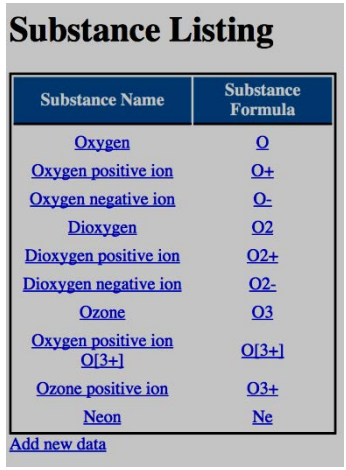


Рис. 7. Список веществ введенных в БД.

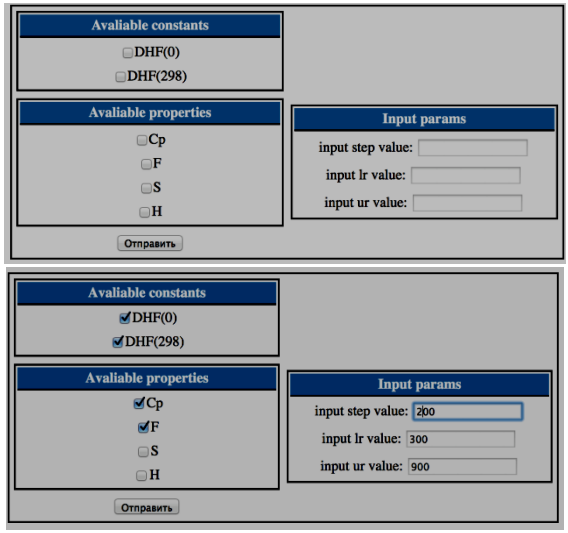


Рис. 8. Список свойств и форма для ввода.

Data Listing

STATE: idgas

Data Source
ИСТОЧНИК Гурвич Л.В., Вейц И.В., Медведев В.А. и др. Термодинамические свойства индивидуальных веществ. Том 1, книга 1 и 2. Москва, Наука, 1978

Constant Name	Constant Value	Constant Dimension	Constant Uncertainty	Uncertainty Type
DHF(0)	246.795	kJ/mol	0.0	Precision class 1
DHF(298)	249.17999	kJ/mol	0.0	Precision class 1

Data Source
ИСТОЧНИК Гурвич Л.В., Вейц И.В., Медведев В.А. и др. Термодинамические свойства индивидуальных веществ. Том 1, книга 1 и 2. Москва, Наука, 1978

T	Uncertainty		Cp	Uncertainty		F	Uncertainty	
K	Value	Type	J/mol*K	Value	Type	J/mol*K	Value	Type
300.0	0.0	Precision class 1	21.9	0.0	Precision class 1	138.533	0.0	Precision class 1
500.0	0.0	Precision class 1	21.27	0.0	Precision class 1	149.952	0.0	Precision class 1
700.0	0.0	Precision class 1	21.035	0.0	Precision class 1	157.353	0.0	Precision class 1
900.0	0.0	Precision class 1	20.932	0.0	Precision class 1	162.82	0.0	Precision class 1

вернуться к списку веществ

Рис. 9. Сформированный набор данных.

Задача системы не ограничивается хранением введенных данных, предполагая возможность ряда вычислительных операций, в том числе: для контроля корректности численных данных, основанного на физических принципах; расчета физического свойства в произвольной точке, например, при произвольно заданной пользователем температуры; при добавлении пользователем нового свойства в виде формулы или программного кода.

Так как разработка собственной системы машинной математики является достаточно трудоемкой, решено использовать одну из наиболее популярных систем *Wolfram Mathematica*. Ее преимущество перед аналогичными состоит в том, что она позволяет проводить расчеты функций (интегрирование, дифференцирова-

ние) в аналитическом виде, предоставляя при этом доступ к вычислительным средствам прямо изнутри программного комплекса, путем подключения к ядру. Для того, чтобы посчитать функцию *Mathematica,* достаточно передать ее текстовый вид, детализируя вычисления. Например, чтобы посчитать интеграл, необходимо указать подынтегральную функцию, переменную и пределы интегрирования, и затем обратиться к ядру; система сама выбирает алгоритм расчета перед выдачей ответа. Система также поддерживает графическую интерпретацию результатов, дополняя выдачу таблиц графиками.

После выбора свойств, осуществляется запрос для выбора данных, при этом заполняется форма для отображения, которая хранит в себе пару **<Property_value, Temperature_value>**, затем все наборы пар классифицируются по состояниям, и далее идет сопоставление **<Temperature_value, Prop_1_val, Prop_2_val, Prop_3_val,…>.** Если некоторой пары **<Prop_k_val, Temperature_value>** в БД не оказывается, то в зависимости от области определения свойства может использоваться аналитическая функция с передачей запроса к ядру *Mathematica.* Запрос задается с определенным значением параметра <**Temperature_value>**, в ответ на что *Mathematica* возвращает пару **<Prop_k_calculated_val, Temperature_value>** с возможностью ее загрузки в БЛ и отображении на странице.

6. Заключение

Выполненная работа завершает определенный этап в решении достаточно сложной задачи интеграции теплофизических данных [1-3, 10]. Анализ предметной области позволил «отстроить» детализированную онтологию для задач, выходящих за рамки возможностей традиционных БД. Формализуя предметную область, онтология обеспечивает строгое и унифицированное описание терминов и понятий. В рамках этой системы онтология позволяет

поддерживать все виды логических и математических связей, например связей между термодинамическими функциями или ограничений в отнесении тех или иных свойств в определенных фазовых состояниях. Наконец, применительно к целям интеграции разнородных данных важную роль играет возможность пользователя наращивать списки веществ, свойств, единиц измерения и прочих элементов набора данных. Тем самым удается резко усилить потенциал созданной на основе онтологии реляционной БД. Реализованная система рассматривается авторами как базовый элемент при создании архитектуры, использующей принципы и технологии Semantic Web.

Литература

1. О.М. Атаева, А.О. Еркимбаев, В.Ю. Зицерман, Г.А. Кобзев, В.А. Серебряков, К.Б. Теймуразов, Р.И. Хайруллин. Представление данных по теплофизическим свойствам веществ с использованием концепций и методов Semantic Web //Третий Всероссийский Симпозиум "Инфраструктура научных информационных ресурсов и систем". Сухум, 30.09 – 03.10 2013 г.
2. О.М. Атаева, А.О. Еркимбаев, В.Ю. Зицерман, Г.А. Кобзев, В.А. Серебряков, К.Б. Теймуразов, Р.И. Хайруллин. Интеграция данных по теплофизическим свойствам веществ методами онтологического моделирования. Электронные библиотеки: перспективные методы и технологии, электронные коллекции. XV Всероссийская научная конференция. Ярославль, Россия, 4-17 октября 2013 года. – Ярославль: ЯрГУ, 2013.- 422 с. ISBN 978-5-8397-1004-7
   URL: http://rcdl2013.uniyar.ac.ru/doc/full_text/rcdl_ataeva_i_dr.pdf
3. А.О. Еркимбаев, В.Ю. Зицерман, Г.А. Кобзев, В.А. Серебряков, Л.Н. Шиолашвили. Интеграция данных по свойствам веществ и материалов на основе онтологического моделирования предметной области // Журнал «Электронные библиотеки», 2013, том 16, № 6 URL:

http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2013/part6/EZKSS
4. M. Frenkel. Global Information Systems in Science: Application to the Field of Thermodynamics. J. Chem. Eng. Data 2009, 54, 2411–2428.
5. R. C. Wilhoit, K.N. Marsh. COdataSTAndardThermodynamics. Rules for Preparing COSTAT Message for Transmitting Thermodynamic Data, Report to CODATA Task Group on Geothermodynamic Data and Chemical Thermodynamic Tables, Paris (1987).
6. Горгораки Е.А., Краевский С.Л., Трахтенгерц М.С. Швальб В.Г., Шпильрайн Э.Э., Якимович К.А. Автоматизированная информационно-поисковая система Теплофизического Центра ИВТАН // Обзоры по теплофизическим свойствам веществ. Москва 1977. №4.
7. The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/home/publications/e-resources/inchi.html
8. CAS REGISTRY – The gold standard for chemical substance information. http://www.cas.org/content/chemical-substances
9. The NIST Chemistry WebBook. http://webbook.nist.gov
10. Хайруллин Р.И. Метод проектирования систем хранения и интеграции теплофизических данных на основе онтологической модели предметной области. Дипломная работа. МФТИ, кафедра системного программирования, 2014.

РАСПРЕДЕЛЕННАЯ НАУЧНАЯ СРЕДА ДЛЯ КОМПЛЕКСНОЙ ПОДДЕРЖКИ РАЗРАБОТЧИКОВ ИНТЕЛЛЕКТУАЛЬНЫХ СППР[1]

Загорулько Г.Б., Загорулько Ю.А.

Федеральное государственное бюджетное учреждение науки Институт систем информатики им. А.П. Ершова
Сибирского отделения Российской академии наук, Новосибирск
gal@iis.nsk.su, zagor@iis.nsk.su

В докладе рассматривается распределенная информационно-вычислительная среда, предназначенная для помощи разработчикам систем поддержки принятия решений (СППР), создаваемых в различных, в том числе слабо формализованных, областях.

Рассматриваемая среда строится на основе онтологии задач и методов ППР (поддержки принятия решений), которая выступает концептуальной основой комплексной поддержки разработчиков. Для обеспечения информационно-аналитической поддержки, необходимой на начальных этапах проектирования СППР, в среду включен интернет-ресурс, в котором представлена систематизированная в соответствии с онтологией информация о поддержке принятия решений. На этапе реализации большую роль играет компонентная поддержка разработчиков, обеспечивающая возможность выбора готовых программных компонентов (модулей), что может существенно облегчить и ускорить процесс создания СППР. Средством решения этой задачи является репозитарий МППР. Данный репозитарий реализован в виде распределенной системы, содержащей набор программных сервисов, реализующих описанные на интернет-ресурсе методы, сценарии задач, решаемых этими методами, и инфраструктуру, позволяющую исполнять готовые сценарии как на локальных, так и на удаленных серверах, а также создавать новые методы и сценарии, в том числе путем композиции уже существующих.

---

[1]Работа выполнена при финансовой поддержке РФФИ (проект № 13-07-00422).

Ключевые слова: система поддержки принятия решений**,** комплексная поддержка разработчиков, онтология задач и методов ППР, портал по ППР, репозитарий методов ППР.

The paper presents the distributed information-computing environment designed to provide a comprehensive support for developers of decision support systems (DSS). A conceptual basis of such support is the ontology of tasks and methods of decision support (DS). To provide information and analytical support, the online resource presenting information about DS domain systematized in accordance with the ontology is included into the environment. To provide a component support the environment contains a repository of DS methods implemented as a distributed system.

Keywords: decision support system, a comprehensive support of developers, DS task and methods ontology, portal on DS, repository of DS methods.

Введение

При создании таких сложных программных систем, какими являются системы поддержки принятия решений (СППР), возникает ряд методологических и технических проблем.

Первая проблема связана со сложностями междисциплинарного взаимодействия. В разработке СППР участвуют специалисты разных типов – эксперты той предметной области (ПО), для которой создается СППР, инженеры знаний, являющиеся специалистами в области представления знаний и поддержки принятия решений, и программисты. Все они говорят на своем профессиональном языке, используют свою терминологию. Для их успешной совместной работы необходим общий концептуальный базис, единая система понятий.

Следующая проблема вытекает из необходимости междисциплинарного взаимодействия. Поскольку разработчики СППР работают над общими задачами, они зачастую вынуждены осваивать смежные специальности. Инженерам знаний приходится стать

немного экспертами в области знаний создаваемой СППР и освоить навыки программиста. Эксперты и программисты должны иметь хотя бы общее представление о методах и средствах представления знаний и поддержки принятия решений. При этом в рамках теории принятия решений разработано большое количество методов [1-3], и в них очень трудно ориентироваться разработчику, не имеющему специальной подготовки. Чтобы повысить скорость и эффективность создания СППР, необходима качественная информационная поддержка, позволяющая разработчику лучше уяснить стоящую перед ним задачу и проанализировать доступные методы и средства для ее решения.

Третья проблема связана с переиспользованием готовых разработок. В настоящее время в свободном доступе находится большое количество программных продуктов – библиотек, пакетов, приложений. Разработчик, казалось бы, имеет хорошую возможность использовать при создании своей системы готовые компоненты. Однако зачастую оказывается, что их непросто найти на просторах Интернета. Кроме того, для решения задач может понадобиться несколько компонентов. И тут возникают новые сложности – готовые программы имеют свои стандарты, форматы входных - выходных данных, работают на удаленных серверах. Поэтому их интеграция может оказаться очень непростой.

В докладе описывается распределенная среда, которая включает средства решения указанных проблем. Эти средства призваны оказать комплексную поддержку разработчикам СППР, чтобы облегчить и ускорить процесс их работы, позволить создавать более качественные и эффективные системы.

В качестве концептуального базиса комплексной поддержки была разработана онтология задач и методов ППР (поддержки принятия решений). Для информационной поддержки создан информационный ресурс, а для компонентной поддержки – репозитарий методов ППР.

1. Концептуальный базис комплексной поддержки

В настоящее время онтологии используются на всех этапах жизненного цикла, практически, любой интеллектуальной системы и играют в ней очень важную роль [4]. Рассматриваемая среда содержит большой объем знаний о предметной области «Поддержка принятия решений» и основывается на онтологии задач и методов ППР. Рассмотрим, какие возможности обеспечивает данная онтология.

Поскольку в разработке СППР принимают участие разные типы специалистов (эксперты, инженеры знаний, программисты) и каждый из них использует свою профессиональную терминологию, свести ее к общей системе понятий и помогает онтология ЗиМППР. Для этого она должна иметь так называемую «вертикально-горизонтальную» структурную организацию [5], что позволит представлять интересующую область знаний в двух измерениях. Первое измерение («вертикальная структуризация») представляется в виде традиционной иерархической структуры, на каждом уровне которой интересующая область знаний описывается с разной степенью детализации. Второе измерение («горизонтальная структуризация») задает описание области знаний с точки зрения разных типов специалистов, участвующих в процессе создания и использования СППР.

В качестве общего ядра, на котором строится вертикально-горизонтальная онтология ЗиМППР, выступает метаонтология задач и методов (см. рис.1).

Рис. 1. Вертикально-горизонтальная организация онтологии ЗиМППР.

Данная метаонтология содержит описание таких базовых понятий поддержки принятия решений, как Задача, Метод, Модуль, Решатель, Входные данные, Результат, Ситуация, Проблемная ситуация, Альтернатива, Этап принятия решений, а также отношения между ними.

Онтология ЗиМППР тесно связана с понятиями предметных областей, для которых создаются СППР. Для того чтобы дать представление о задачах и методах, не вдаваясь в конкретику ПО, описания этих областей должны быть одинаково устроены. Это задается путем использования в качестве основы для построения онтологии конкретной предметной области метаонтологии ПО, включающей базовые понятия, являющиеся общими для всех ПО. При специализации СППР на определенную предметную область и для облегчения работы экспертов и инженеров знаний могут разрабатываться онтологии базового уровня, которые затем используются для построения онтологии конкретной ПО.

2. Организация информационной поддержки

Для оказания информационной поддержки был разработан портал знаний, который на основе онтологии задач и методов ППР систематизирует имеющуюся в данной области информацию, а также предлагает средства для ее просмотра, анализа и практического использования. Рассматриваемый портал представляет собой интернет-ресурс, построенный с использованием технологии разработки порталов научных знаний [6] и дополненный новыми возможностями. На рис. 2 показаны его основные содержательные и функциональные компоненты. Знания о предметной области портала, которой является поддержка принятия решений, представлены в виде онтологии задач и методов ППР, которая помимо понятий и отношений ПО содержит такие концепты, как информационные ресурсы и проблемно-

ориентированные сервисы. Информационное наполнение портала, его контент, образуют экземпляры онтологии – конкретные задачи, методы, ресурсы и другие объекты. Таким образом, контент представляет систематизированную в соответствии с онтологией информацию о поддержке принятия решений.

Рис.2. Компоненты портала.

Вся работа на портале осуществляется с помощью системных сервисов. Сервисы управления онтологией и контентом позволяют создавать новые понятия, отношения и объекты, а также редактировать и удалять ранее созданные. Средства поиска, навигация, фильтрации и визуализации позволяют получать, просматривать и анализировать необходимую пользователю информацию. И, наконец, средства управления проблемно-ориентированными сервисами дают возможность запустить эти сервисы и получить результаты их работы.

Качественная информационная поддержка предполагает наличие развитого пользовательского интерфейса. Он должен включать средства поиска требующейся информации, навигации по ее структурным элементам и удобного их просмотра. Необходимо также наличие широкого спектра аналитических инструментов, обеспечивающих такие формы «подачи» и визуализации просматриваемой информации, которые способствуют улучшению ее восприятия.

Имеющиеся на портале сервисы поиска позволяют осуществлять как поиск информации по ключевым словам, так и расширенный семантический поиск с использованием ограничений, задаваемых в терминах онтологии.

Сервис навигации позволяет, просматривая описание понятия или объекта, переходить к описанию связанных с ним понятий и объектов, представляя таким образом его семантическую окрестность.

В качестве аналитических инструментов на портале используются средства фильтрации и визуализации объектов и понятий. Фильтрация позволяет из большого списка выбрать объекты, значения атрибутов которых удовлетворяют указанным ограничениям. Сервис визуализации [7] предоставляет следующие возможности:

1. Отображение полной системы понятий в виде графа (рис. 3., справа).
2. Оценка «мощности» понятий с помощью круговой диаграммы. Под мощностью подразумевается число объектов, относящихся к данному понятию (рис. 3, слева).
3. Отображение полной семантической сети объектов
4. Отображение сети объектов, связанных выбранным отношением.

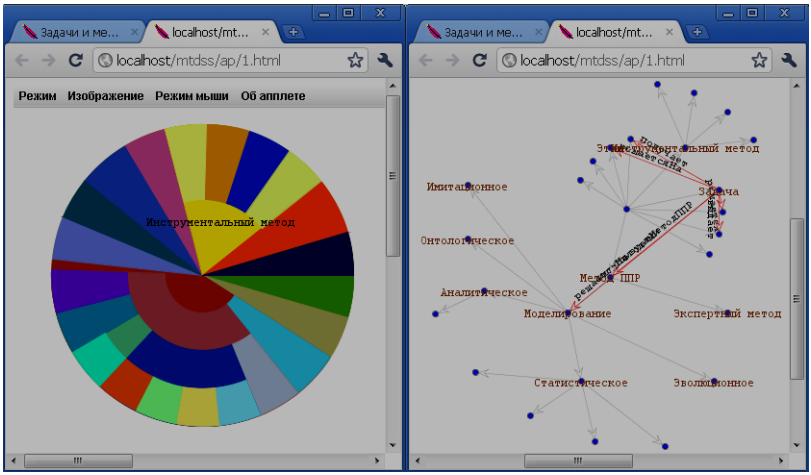


Рис.3. Средства визуализации.

Средства управления проблемно-ориентированными сервисами позволяют создавать и исполнять эти сервисы непосредственно из портала. Добавление таких средств на портал поднимает возможности оказания информационной поддержки на качественно новый уровень – пользователь не просто получает информацию об интересующем его методе или ссылку на реализацию метода. Он может тут же, на портале, посмотреть примеры использования метода, сам запустить его, проанализировать его работу с разными входными данными.

Для подробного описания методов и реализующих их алгоритмов в интерфейс портала был встроен текстовый редактор. Описания методов, задаваемые или просматриваемые с помощью данного редактора, являются значением атрибута «Подробное описание» соответствующего метода. На рис. 4 представлено описание метода внутренней точки для незарегистрированного

пользователя. Для зарегистрированных пользователей подробное описание отображается в режиме редактирования с широким набором инструментальных средств, включая средства задания математических формул.

Для создания проблемно-ориентированных сервисов был разработан web-сервис, который на данном этапе реализует несколько методов математического программирования. Этот web-сервис снабжен пользовательским интерфейсом, позволяющим просматривать спецификации реализуемых им методов.

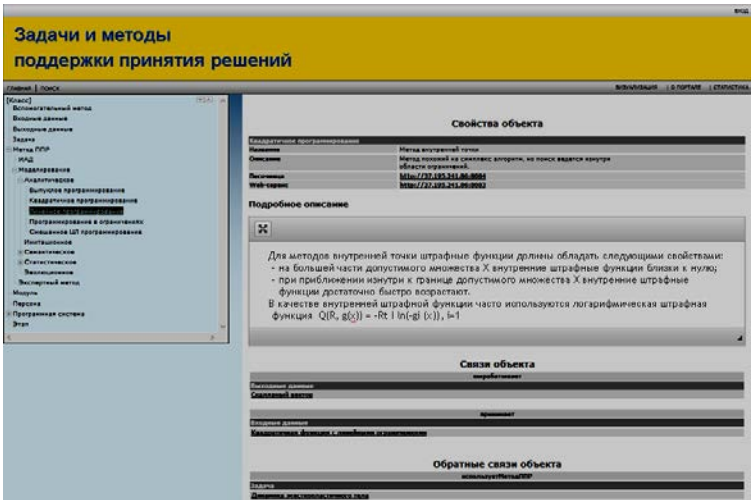


Рис. 4. Представление описания метода пользователям портала знаний по ППР.

Возможность практического ознакомления с работой методов поддержки принятия решений реализована в виде web-ресурса, являющегося своего рода «песочницей», позволяющей задать входные данные, запустить исполнение метода и посмотреть результаты его работы.

3. Организация компонентной поддержки

В процессе реализации СППР большую помощь разработчикам может оказать репозитарий – пул готовых к использованию методов вместе с методикой и средствами их исполнения и композиции.

Рассмотрим, какие функциональные возможности должен иметь такой репозитарий:

1. Представительный корпус методов ППР, организованный на основе их онтологии.
2. Возможность исполнения любого метода с передачей ему данных и просмотром результатов его работы. При этом если методы физически развернуты на удаленных машинах, от пользователя должны быть скрыты детали их распределенной работы.
3. Возможность создания новых методов путем компоновки нескольких готовых методов.
4. Наличие пользовательского интерфейса для просмотра информации о методах, для их компоновки, запуска и просмотра результатов.
5. Возможность публикации новых методов и предоставления доступа к ним.
6. Возможность работы как с desktop-, так и с online-версиями репозитария.

Для продвинутых пользователей полезной возможностью будет наличие у методов программных интерфейсов (API) для обеспечения их встраивания в другие методы и/или приложения.

При работе с репозитарием прикладной пользователь не должен изучать дополнительное программное обеспечение, форматы описания, команды запуска методов, передачи данных между отдельными компонентами разрабатываемой системы и т.п. Репозитарий должен включать средства, поддерживающие инфраструктуру. Во многих случаях пользователю будет достаточно

свести исходную задачу к одной или нескольким задачам известных классов, для которых существуют готовые методы. При необходимости разработки новых методов потребуются услуги продвинутого пользователя с квалификацией системного программиста. Однако после того как метод создан, он может многократно переиспользоваться другими пользователями, не имеющими специальной подготовки.

При разработке репозитария использовался подход, сочетающий достоинства распределенных вычислительных систем, в частности, сервис-ориентированных научных сред [8], и возможности средств, предлагаемых для их построения.

4. Разработка репозитария

4.1. Сервис-ориентированный подход

Сервис-ориентированная научная среда (СОНС) представляет собой распределенную систему, предназначенную для поддержки научных исследований. В СОНС можно выделить три уровня (Рис. 5): уровень вычислительных ресурсов, уровень сервисов и уровень приложений. Вычислительные ресурсы образуют сервера и хранилища данных. Агрегированные в СОНС сервисы делятся на системные и прикладные. Системные сервисы осуществляют непосредственную работу с ресурсами, тогда как прикладные сервисы предназначены для решения проблемно-ориентированных задач пользователей. Приложения могут использовать все доступные сервисы, и в отличие от последних, предоставляют пользовательский интерфейс. Все программные компоненты СОНС могут быть распределены на разных серверах, а методы – реализованы в виде локальных или web-сервисов.

Развертывание СОНС в облачной среде позволяет разработчикам прикладных сервисов и пользователям не задумываться о том, где находятся нужные им сервисы и приложения, а исполь-

зовать их так, как если бы они находились на локальном компьютере.

В работе [8] подчеркивается, что СОНС должны базироваться на вычислительной архитектуре грид-систем, используя ее для проведения сложных вычислений и хранения больших массивов данных. При этом акцент смещается от агрегации вычислительных ресурсов на решаемые с их помощью задачи.

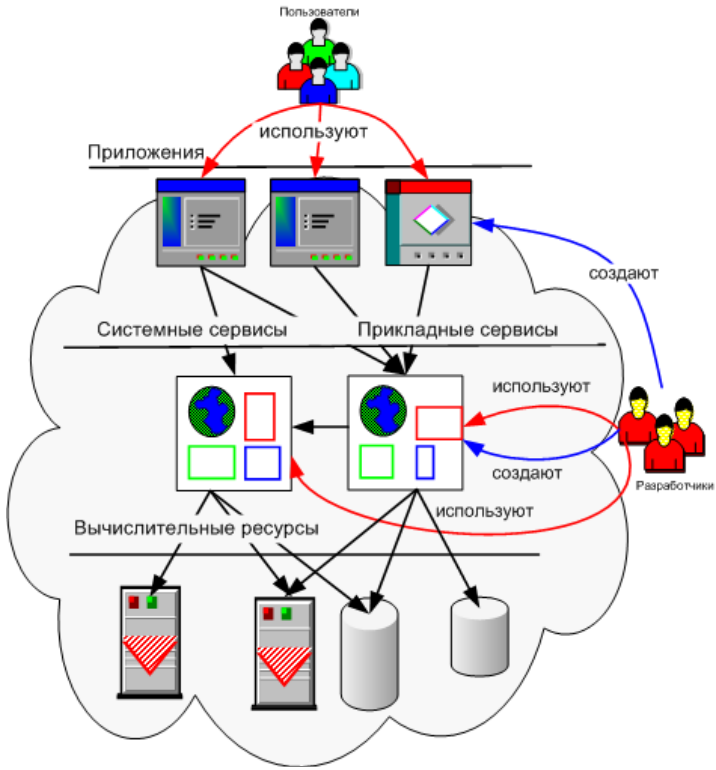


Рис.5. Сервис-ориентированная научная среда.

Разработка репозитария как распределенной среды позволяет решить задачи, поставленные перед его разработчиками, и обеспечить требуемую функциональность. Имеющиеся в свободном доступе методы ППР, реализованные в виде библиотек, пакетов, приложений, легко могут быть представлены в виде сервисов, а средства, предоставляемые доступными инструментариями, существенно упростят эту процедуру и ускорят процесс разработки репозитария.

4.2. Платформа для построения репозитария

В настоящее время активно разрабатываются подходы к созданию распределенных систем нового поколения – облачных платформ, COHC, систем управления сценариями. Существует ряд платформ, поддерживающих разработку таких систем [9]. Они решают как технологические, так и организационные проблемы, связанные с вовлечением исследователей в процесс создания проблемно-ориентированных сервисов.

Основные функциональные возможности и свойства платформ в совокупности представлены на рис. 6.

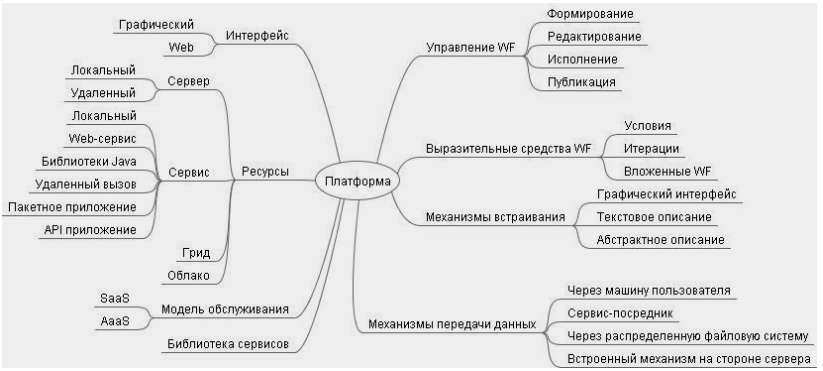


Рис. 6. Функциональные возможности и свойства платформ для построения распределенных систем нового поколения.

Системы, разрабатываемые с помощью подобных платформ, ориентированы на прикладного пользователя и предоставляют графический desktop или web-интерфейс. Такие инструментарии поддерживают широкий спектр ресурсов. Они могут быть развернуты на локальных или удаленных серверах, использовать грид- или облачную инфраструктуру, предоставлять разнообразное программное обеспечение (сервисы) – web-сервисы, локальные или удаленные вызовы модулей из командной строки, библиотеки Java, пакетные или API приложения. Платформы ориентированы на поддержку моделей обслуживания SaaS (Software as a Service) или AaaS (Application as a Service). Одной из наиболее интересных особенностей таких систем является возможность публикации сервисов и их повторного использования для разработки композитных приложений на основе концепции workflow (WF) – потоков работ или сценариев. Для управления WF (разработки, редактирования, исполнения, преобразования в новый сервис) инструментальные системы предоставляют редакторы с графическим интерфейсом или специализированным языком. Для передачи данных между модулями WF могут быть задействованы разные механизмы: через машину пользователя или распределенную файловую систему, либо с использованием сервисов-посредников или встроенных механизмов на стороне сервера. Наличие ориентированных на конечного пользователя средств для разработки композитных приложений и их дальнейшего совместного использования делает возможным массовое использование распределенных систем для научных исследований и создания сервис-ориентированных научных сред.

После анализа доступных платформ для разработки репозитария была выбрана система Taverna [10], разрабатываемая университетом Манчестера и распространяемая под лицензией LGPL 2. Архитектура данной системы включает в себя автономное при-

ложение, возможность управления WF на локальной машине с возможностью доступа к серверу системы Taverna. В качестве вычислительных ресурсов данная система позволяет использовать как локальную машину, так и удаленные сервера и грид. В качестве сервисов – модулей, из которых конструируются WF, в Taverna можно использовать web-сервисы (WS), описанные в формате WSDL и REST, скрипты на языках R и Beanshell, библиотеки Java, локальный и удаленный вызов команд. Модель workflow в этой системе представляет собой направленный граф, для построения и редактирования которого предлагается графический редактор. Сохранять workflow можно в форматах SCUFL (XML) или SCUFL2 (OWL). При построении WF можно использовать такие конструкции, как условия, циклы, итерации, вложенные WF. Новые WF можно публиковать в виде web-сервисов.

4.3. Структура репозитария

Репозитарий методов ППР состоит из 3 удалённых взаимодействующих между собой серверов и клиентского фреймворка:

1. Сервер баз данных. Удалённый компьютер с установленным MySQL сервером. Обеспечивает поддержку использования БД при создании СППР.
2. WorkFlow-сервер. Удалённый компьютер с установленным Taverna Server. Хранит набор готовых WF, которые могут потребоваться разработчику СППР.
3. Tool-сервер. На этом сервере хранятся методы ППР, библиотеки, модули, программные пакеты и приложения, которые могут потребоваться для разработчика СППР. Вызов определённого метода происходит путём обращения к этому серверу по безопасному протоколу SSH.
4. Taverna Workbench. Представляет собой framework для создания и управления workflow.

Для работы с репозитарием предложена методика, которая позволяет использовать в сценариях базы данных, выполнять различные SQL-запросы, вызывать методы, приложения, модули, хранящиеся на удалённом компьютере, работать с программами, написанными на Java и хранящимися в виде JAR-файлов, работать с различными Windows-приложениями, хранящимися в репозитарии в виде EXE-файлов.

В репозитарий были подключены следующие методы и инструменты для поддержки принятия решений: интегрированная программная среда для построения систем, основанных на знаниях SempN [11], фреймворк для разработки методов принятия решений на основе прецедентов Project.J [12], сервисы целочисленного линейного программирования GLPK [13], универсальный решатель математических задач Unicalc [14], набор методов для выявления трендов и аномальных значений во временном ряде.

На рис. 7 представлен сценарий, разработанный с помощью Taverna Workbench, и результаты его работы.

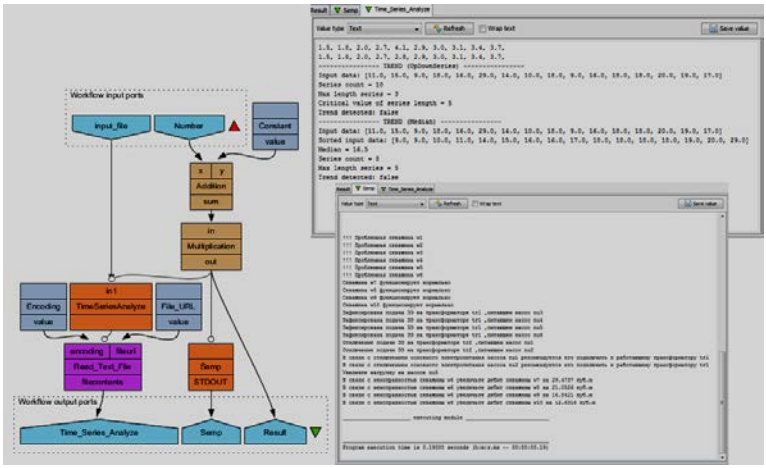


Рис. 7. Исполнение сценария в системе Taverna.

Заключение

В докладе описывается сервис-ориентированная научная среда, предназначенная для комплексной поддержки разработчиков на всех этапах создания СППР. Эта среда включает Интернет-ресурс, в котором хранится структурированное на основе онтологии описание области знаний «поддержка принятия решений», задач, решаемых в данной области, методов и данных, используемых для решения задач. Ресурс оказывает информационную поддержку, дает общее представление о задачах и методах и позволяет детально познакомиться с интересующими методами. Такая поддержка очень ценна на этапе проектирования.

Для оказания помощи разработчику на этапе реализации СППР был разработан репозитарий задач и методов ППР. Разработка осуществлялась на основе системы Taverna. В начальное наполнение репозитария были включены следующие методы: рассуждение на основе прецедентов, рассуждение на основе экспертных правил, анализ временных рядов, аналитическое моделирование с использованием метода недоопределенных вычислений, методы целочисленного линейного программирования. Репозитарий реализован в виде распределенной сервис-ориентированной системы, позволяющей использовать ранее разработанные методы в виде сервисов и создавать новые методы, в том числе, путем композиции уже имеющихся в репозитарии методов.

Дальнейшее развитие среды предполагает расширение начального наполнения репозитария набором методов, описанных в онтологии задач и методов ППР и реализацию взаимодействия Интернет-ресурса и репозитария.

Литература

1. Ларичев О.И. Теория и методы принятия решений, а также Хроника событий в Волшебных странах.– М.: Логос, 2000.

2. Петровский А.Б.. Теория принятия решений. – М.: Издательский центр «Академия», 2009.

3. Саати Т. Л. Принятие решений. Метод анализа иерархий. – М.: Радио и связь, 1993.

4. Загорулько Ю.А., Загорулько Г.Б. Использование онтологий в экспертных системах и системах поддержки принятия решений // Труды Второго симпозиума «Онтологическое моделирование» (Казань, октябрь 2010 г.) – Москва: ИПИ РАН, 2011. – С. 321-351.

5. Загорулько Г.Б. Обеспечение информационной поддержки разработчиков СППР// Труды XVIII Байкальской Всероссийской конференции «Информационные и математические технологии в науке и управлении». Часть III. – Иркутск: ИСЭМ СО РАН, 2013. – С. 137–142.

6. Загорулько Ю. А., Боровикова О.И. Подход к построению порталов научных знаний // Автометрия. 2008. Т. 44. № 1. С. 100 –110.

7. Апанович З.В., Винокуров П.С., Кислицина Т.А. Методы и средства визуализации информационного наполнения больших научных порталов // Вестник НГУ Серия: Информационные технологии. 2011. – том 9, выпуск 3, Редакционно-издательский центр НГУ. – С. 5-14.

8. Сухорослов О.В. Архитектура и реализация сервис-ориентированной научной среды MathCloud/ XIII Российская конференция с участием иностранных ученых "Распределенные информационные и вычислительные ресурсы" (DICR'2010): материалы конф. – Электрон. дан. – Новосибирск: ИВТ СО РАН, 2010. – 1 электрон. опт. диск (CD-ROM).

9. Князьков К.В. Технология разработки композитных приложений с использованием предметно-ориентированных программных модулей. Автореферат. СПб, 2012. – 19 с.

10. Taverna Workflow Management System.
www.taverna.org.uk

11. Загорулько Ю.А., Попов И.Г.. Представление знаний в интегрированной технологической среде Semp-ТАО. // Проблемы представления и обработки не полностью определенных знаний. –Москва-Новосибирск, 1996. – С.59–74.

12. Загорулько Г.Б., Шмаков Е.С. Онтологический Подход к Разработке Интеллектуальных СППР на Основе Прецедентов // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–2013), 8-10 октября 2013 г., Новосибисрк. – Новосибирск: Институт математики им. С.Л. Соболева СО РАН, 2013. – Т. 1. – С. 157-164.

13. Glpk for Windows. Glpk: GNU Linear Programming Kit / Version 4.34. URL: http://gnuwin32.sourceforge.net/packages/glpk.htm (дата обращения: 26.06.2014).

14. Semenov A.L. Solving Optimization Problems with Help of the UniCalc Solver // Applications of Interval Computations. R.B. Kearfott and V. Kreinovich (Eds). Kluwer Academic Publishers. 1996. – P. 211–225.

РАЗРАБОТКА ИНТЕРНЕТ-ПОРТАЛА ПО ТЕПЛОФИЗИЧЕСКИМ СВОЙСТВАМ ХИМИЧЕСКИХ ВЕЩЕСТВ

Г.Б. Загорулько[a], *Ю.И. Молородов*[b]

[a] Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск

[b] Институт вычислительных технологий СО РАН, Новосибирск

gal@iis.nsk.su, yumo@ict.sbras.su

Рассматривается интернет-портал данных, обеспечивающий систематизацию и интеграцию знаний и информационных ресурсов по изучению теплофизических свойств металлов и сплавов в широком диапазоне температур. Использование онтологии для описания предметной области портала позволяет семантически структурировать его информационное наполнение и организовать в нем навигацию и содержательный поиск информации. Для разработки данного портала используется технология, ориентированная на экспертов предметных областей.

Ключевые слова: портал знаний, неорганические вещества, металлы, сплавы, информационные ресурсы, онтология, содержательный доступ, управляемая онтологией навигация.

The Internet portal of data providing systematization and integration of knowledge and information resources on studying of heatphysical properties of metals and alloys in the wide range of temperatures is considered. Ontology use for the description of subject domain of a portal allows to structure semantic its information filling and to organize in it navigation and substantial information search. For development of this portal the technology focused on experts of subject domains is used.

Keywords: Knowledge portal, inorganic substances, metals, alloys, information resources, ontology, content-based access, ontology-driven navigation.

Введение

Исследования теплофизических свойств металлов при высоких температурах представляют важную научную проблему, имеющую большую практическую значимость.

Они служат не только основой для дальнейшего развития высокотемпературной физики твердого тела, но и позволяют определить области практического использования новых материалов.

Практическая значимость таких работ определяется стремительным развитием техники высоких температур, созданием новых материалов, обладающих уникальными характеристиками.

Развитие теплофизики сопровождается нарастающим производством новых данных, публикуемых в десятках журналов различного профиля. Современный этап характеризуется выделением систематизации данных в самостоятельное направление, наряду с экспериментом и теорией, а также повсеместным переходом от печатной формы к базам данным (БД).

В связи с этой задачей, необходимо обрабатывать и осмысливать огромные массивы данных, полученных при проведении экспериментов, опубликованных в научной литературе и справочниках. Современные научные достижения в области информационно-вычислительных технологий, в частности веб-ориентированные информационно-вычислительные системы, дают основу для решения этих проблем.

Становится актуальным организация эффективного доступа не только к публикациям, описывающим методы и подходы к исследованию свойств неорганических и органических веществ, но и разного рода справочникам, программным компонентам и алгоритмам, обеспечивающим решение различных задач по работе с данными исследований.

1. Предметная область

Теплофизика – одна из дисциплин, в которых центральное место занимает работа с численными данными. При работе с ними приходится учитывать, что в публикациях и БД используют несколько типовых форм, а именно: табличную, графическую и математическую (в виде хранимых формул или программных кодов). Графическая форма иллюстрирует характер зависимостей, рассеяние опытных точек и т.п. Табличная форма наиболее надежна при передаче данных, легко контролируема в отношении пропусков, ошибок в знаке или порядке величины и т.п. Математическая форма, избавляя от интерполяции, требует повышенной тщательности в обнаружение ошибок, легко вылавливаемых в табличной форме. Доминирующей формой в экспериментальных работах и справочниках является именно табличная форма.

Развитие теплофизики сопровождается нарастающим производством новых данных, публикуемых в десятках журналов различного профиля. Современный этап характеризуется выделением систематизации данных в самостоятельное направление, наряду с экспериментом и теорией, а также повсеместным переходом от печатной формы к базам данных (БД) [9].

Переход к созданию Интернет-ресурсов базирующихся на использовании БД, предполагает проведение экспертной оценки существующих массивов данных экспериментальных измерений. Это позволяет обосновать выбор базовых значений и оценить их погрешности, провести обработку разнородных данных с помощью физически обоснованных и термодинамически согласованных моделей, и, что особенно важно – построить информационно-аналитические системы, которые позволят оперативно хранить либо выбирать данные по свойствам веществ при произвольных значениях температур, давлений и составов.

Такая специализированная проблемно-ориентированная информационная система (ИВС) позволит решить проблему хранения эмпирического материала, его обработку и интерпретацию. ИВС позволят перевести работу с этими данными на качественно более высокий уровень, открывающий перспективы для постановки и эффективного решения новых научных и практических задач.

Информационную систему по теплофизическим свойствам химических веществ предлагается разрабатывать в виде портала научных знаний с использованием технологии, хорошо зарекомендовавшую себя в ряде гуманитарных дисциплин [5,6], и в которую были добавлены возможности для хранения и обработки специфических для данной области численных данных.

Чтобы портал знаний мог предоставлять пользователям описанные выше возможности, он должен не только иметь гибкие средства представления разнородной информации и содержательного доступа к ней, но и обеспечивать оперативное управление своим информационным наполнением (контентом). Этим целям служит информационная модель портала знаний [8], которая объединяет модели его предметной и проблемной областей, а также описывает типы представляемой в его контенте информации.

2. Построение модели предметной области

Разработка моделей информационных систем и алгоритмов поиска функциональных зависимостей в массивах данных предполагает, прежде всего, построение модели предметной области. В качестве модели предметной области обычно выступает ее *онтология* [6]. Онтология является ядром, базовым компонентом информационной модели портала. Она не только описывает систему знаний портала, но и задает формальные структуры для

представления его контента. Онтология содержит понятия моделируемой области, связывающие их отношения, атрибуты понятий и отношений, ограничения на значения атрибутов, а также аксиомы, определяющие семантику понятий и отношений.

Формализм, используемый в технологии построения порталов научных знаний обеспечивает описание понятий проблемной и предметной областей портала и разнообразных семантических связей между ними, а также выстраивание понятий в иерархию «общее-частное и поддержку наследования свойств по этой иерархии.

При построении любого портала научных знаний его онтология строится на основе двух базовых технологий: онтологии научного знания и онтологии научной деятельности.

Онтология научного знания, по своей сути, является метаонтологией. Она содержит метапонятия, задающие структуры для описания предметной области (области знаний) портала, такие как *Раздел науки, Предмет исследования, Объект исследования, Метод исследования, Научный результат*, позволяющие выделить в данной науке значимые разделы и подразделы, задать типизацию предметов, объектов и методов исследования, описать результаты научной деятельности.

Онтология научной деятельности является онтологией верхнего уровня и включает базовые понятия, относящиеся к организации научно-исследовательской деятельности, такие как *Научный результат*, *Объект исследования*, *Персона, Публикация*, используемые для описания результатов научной деятельности, мероприятий, научных программ и проектов, различного типа публикаций. В эту онтологию также включено понятие *Информационный ресурс*, которое служит для описания информационных ресурсов, представленных в сети интернет.

Свойства каждого понятия описываются с помощью атрибутов и ограничений, наложенных на область их значений. Понятия

базовых онтологий связаны между собой ассоциативными отношениями, выбор которых осуществлялся не только исходя из полноты представления проблемной и предметной областей портала, но и из удобства навигации по его информационному пространству и поиска информации.

Понятия онтологии подраздела научной дисциплины, изучающего теплофизические свойства веществ (ТСВ) являются реализациями метапонятий онтологии научного знания и организованы в несколько иерархий «общее-частное», каждая из которых соответствует одному из метапонятий, представленных в этой онтологии. Все эти иерархии связаны между собой посредством ассоциативных отношений, часть которых наследуется из базовых онтологий, а часть отражает специфику данной предметной области. На рис. 1. представлена онтология ТСВ.

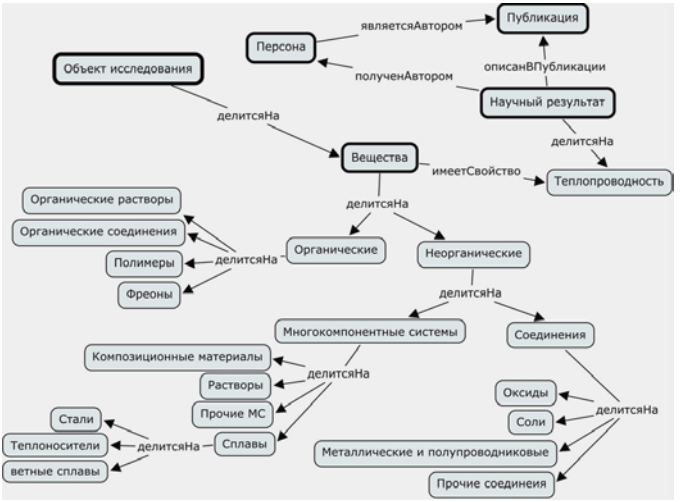


Рис. 1. Онтология портала знаний по теплофизическим свойствам химических веществ.

В качестве базовых *Объектов исследования* данной области рассматриваются химические *Вещества*. Предметом исследования являются такие свойства, как *Теплопроводность, Температура, Давление* и т.д.

Иерархия *Методов исследования* служит для систематизированного описания инструментов исследования, применяемых при исследовании теплофизических свойств…. В этой иерархии выделены следующие подклассы: *Методы измерений теплопроводности (температуропроводности):* стационарные и нестационарные.

Большинство исследований теплофизических величин выполнено стационарной группой методов из-за простоты конструкции экспериментальных аппаратов, достаточно высокой точности измерений, и очевидностью расчетных формул [11].

Измерения стационарной группой методов затрачивает много времени на проведение единичного эксперимента, что является еще одним недостатком, по сравнению с нестационарными методами. Именно по этим причинам широкую популярность получили нестационарные методы, имеющие практически ту же точность измерений, что и стационарные, и требующие гораздо меньше времени для проведения измерений. Основным их недостатком является более сложная приборная конструкция и методика обработки экспериментальных результатов, которые преодолеваются современными измерительными и вычислительными техниками. Все стационарные методы измерения теплопроводности, согласно [12, 13], можно разделить на две большие группы, в зависимости от того, как нагревается образец. К первой группе относятся методы, использующие наружные электрические нагреватели, и методы, использующие нагрев образца электронным или лазерным пучком. Ко второй группе следует отнести те ме-

тоды, где образец непосредственно нагревается проходящим через него электрическим током.

Нестационарные методы измерения тепло - и температуропроводности, согласно работам [12 – 14] принято разделять на две большие группы. К первой группе относятся методы регулярного теплового режима (нестационарного теплового режима), а ко второй – методы стационарного теплового режима. Под регулярными тепловыми режимами понимаются стадии нестационарных тепловых процессов, характеризуемые независимостью пространственно-временного изменения температуры от начальных условий. Методы первой группы позволяют определять температурную зависимость температуропроводности.

В последнее время все большую популярность приобретают импульсные методы измерения коэффициента температуропроводности (a), зная который, можно рассчитать коэффициент теплопроводности (λ) по известным формулам [14], используя более легко получаемые данные о теплоемкости (C_p) и плотности (ρ).

Данный класс методов обладает рядом достоинств:

- нет необходимости измерять непосредственно тепловой поток Q, что позволяет меньше заботиться о тепловых потерях за счет радиационного и конвективного переноса тепла;
- используя такие методы, можно исследовать широкий спектр различных материалов, как в твердом, так и в жидком состоя–ниях;
- малое время проведения единичного измерения (менее 1 с) в комплексе с малым градиентом температур в образце (не более 2…5 К), позволяет измерять молекулярную теплопроводность жидкостей, избавляясь от свободно - конвективных течений.

К нестационарным методам измерения можно отнести: метод температурных волн Ангстрема, метод продольного теплового потока тепла, метод Кольрауша, метод узкой перемычки, сравнительный метод, метод лазерной вспышки и калориметрические методы измерений.

В основе иерархии *Разделов* науки о ТСВ лежит классификация базовых теоретических и прикладных направлений исследования теплофизических свойств материалов. В качестве основных можно выделить разделы, развивающие следующие направления:

Уравнения состояния, фазовые переходы и критические явления.

- Термодинамические свойства. Базы данных.
- Экстремальные состояния вещества.
- Наноматериалы, наножидкости, межфазные явления.
- Транспортные, оптические и радиационныесвойства.
- Техника теплофизических измерений.

В иерархии Научных результатов представлены взаимные зависимости теплофизических свойств веществ, например зависимость теплопроводности от температуры.

Вводя формальные описания понятий области знаний портала в виде классов объектов и отношений между ними, онтология задает структуры для представления реальных объектов и связей между ними. В соответствии с этим данные на портале представлены как множество разнотипных информационных объектов (ИО) и связей между ними, которые в совокупности и образуют контент портала.

3. Разработка и использование портала знаний по теплофизическим свойствам химических веществ

При разработке порталов научных знаний используется онтологический подход, который предполагает двухэтапную процедуру формирования информационного наполнения портала. На первом этапе инженером знаний совместно с экспертом формируется онтология; на втором этапе эксперт формирует контент в рамках системы понятий и структур, определенных в онтологии.

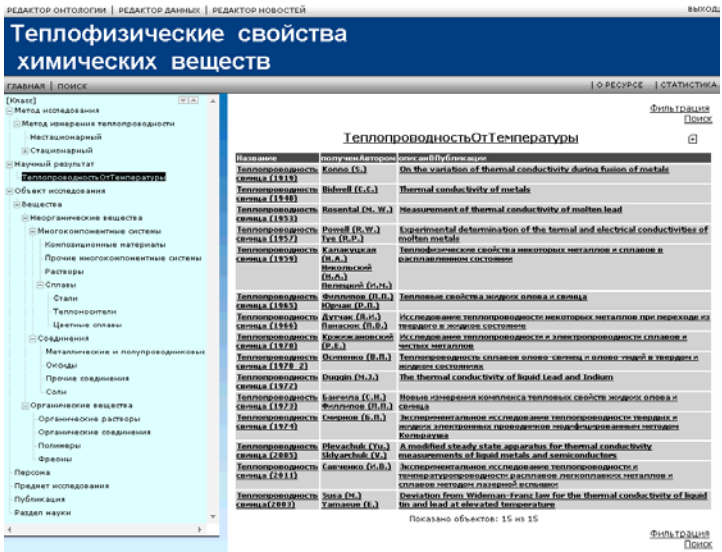


Рис. 2. Просмотр списка экземпляров одного понятия.

Технология построения порталов предлагает разработчикам средства, ориентированные на экспертов – специалистов предметных областей. Для описания онтологии и ввода контента пре-

доставляются редакторы с интуитивно понятным веб-интерфейсом.

Пользователям портала предлагается ряд средств и аналитических инструментов, позволяющих получить как общее представление о предметной области портала, так и детальное описание отдельного понятия или объекта, а также провести анализ интересующей пользователя информации.

На рис. 2 показан пользовательский интерфейс портала. В левом верхнем углу представлена онтология ТСВ, а в основном окне – список экземпляров понятия ТеплопроводностьОтТемпературы, т.е. объектов, представляющих зависимости от температуры теплопроводности свинца, полученные в разное время разными авторами. Каждый элемент такого списка является гиперссылкой, позволяющей перейти на страницу с подробным описанием соответствующего объекта.

На рис. 3 представлена зависимость, полученная, как следует из названия, в 1965 г. Атрибут «Значение» имеет табличный тип и по желанию пользователя может быть просмотрен в виде таблицы. На странице данного объекта также представлены связи с другими объектами. Данный научный результат описан в публикации «Тепловые свойства жидких олова и свинца» и получен авторами Филипповым и Юрчаком.

Атрибуты и связи объекта описывают его свойства образуют контекст или семантическую окрестность данного объекта. Связи, представленными гиперссылками, позволяют перейти к подробному описанию соответствующего объекта, осуществляя таким образом навигацию по информационному наполнению портала.

Рис. 3. Зависимость теплопроводности свинца от температуры.

В качестве аналитических инструментов используются средства фильтрации и визуализации объектов и понятий. Фильтрация позволяет из большого списка выбрать объекты, значения атрибутов которых удовлетворяют указанным ограничениям. Визуализация [16] предоставляет такие возможности, как отображение полной системы понятий в виде графа, оценка «мощности» понятий с помощью круговой диаграммы, отображение полной семантической сети объектов.

Технология также позволяет подключить к порталу другие сервисы, позволяющие обрабатывать представленную на портале информацию.

Заключение

Рассмотренный в статье портал знаний обеспечивает систематизацию и интеграцию знаний и доступных информационных ресурсов, относящихся к изучению свойств органических и неорганических соединений в единое информационное пространство, и содержательный доступ к ним. Благодаря тому, что системати-

зация и структуризация таких знаний и информационных ресурсов выполнена на основе онтологии, доступ к ним осуществляется путем навигации по дереву понятий онтологии и контенту портала, а также через средства поиска в терминах его предметной области.

При создании портала использовались программные средства, методология и технология разработки порталов научных знаний предложенные в [2, 5, 6].

Ближайшими целями авторов является пополнение контента портала новыми результатами и области исследования теплофизических свойств материалов. Планируется подключение к порталу знаний развитых средств графической визуализации, что позволит представлять в виде графа не только иерархии понятий онтологии, но и весь контент.

Список литературы

1. Gruber T. Toward Principles for the Design of Ontologies Used for Knowledge Sharing // International Journal of Human-Computer Studies. – 1995. – V. 43. – Iss. 5–6. – P. 907–928.
2. Загорулько Ю.А., Боровикова О.И. Информационная модель портала научных знаний // Информационные технологии. – 2009. – № 12. – С. 2–7.
3. Using Dublin Core: [сайт]. [1995–2011]. URL: http://dublincore.org/documents/usageguide/ (дата обращения: 10.01.2011).
4. Protйgй. Web site: [сайт]. [2010]. URL: http://protege.stanford.edu/ (дата обращения: 10.01.2011).
5. Загорулько Ю.А. Автоматизация сбора онтологической информации об Интернет-ресурсах для портала научных знаний // Известия Томского политехнического университета. – 2008. – Т. 312. – № 5. – С. 114–119.
6. Загорулько Ю.А., Боровикова О.И. Подход к построению порталов научных знаний // Автометрия. – 2008. – Т. 44. – № 1. – С. 100–110.

7. С.В. Станкус, Р.А. Хайрулин, В.Г. Мартынец, Ю.И. Молородов Плотность перфторгексана в окрестности критической точки испарения // Вестник НГУ. Серия: Физика. 2013. Том 8, выпуск 1. С.73-77.
8. В.Б. Барахнин, Ю.И. Молородов, С.В. Станкус, А.М. Федотов Информационные технологии для задач теплофизических свойств веществ. //Информатика и системы управления. Автоматизированные системы и комплексы. 2013, № 4(38).- С. 149-157
9. А.О. Еркимбаев, В.Ю. Зицерман, Г.А. Кобзев, В.А. Серебряков, Л.Н. Шиолашвили. Интеграция данных по теплофизическим свойствам веществ методами онтологического моделирования
10. T. Ashino. Materials ontology: an infrastructure for exchange materials information and knowledge. Data Science Journal, Volume 9, 8 July 2010, pp. 54-61.
11. Д.А. Самошкин. Экспериментальное исследование теплоемкости и температуропро-водности твердых переходных металлов в широком интервале температур. Выпускная квалификационная магистерская диссертация.- Новосибирск, НГУ, 2014, 70 с.
12. Кондратьев Г.М. Регулярный тепловой режим. – М.: Гостехиздат, 1954. – 408 с.
13. Кондратьев Г.М. Тепловые измерения. – М.-Л.: Машгиз, 1957. – 244 с.
14. Пономарев С.В. Теоретические и практические основы теплофизических измерений. – М.: ФИЗМАТЛИТ, 2008. – 408 с.
15. Филлипов Л.П. Измерение тепловых свойств твердых и жидких металлов при высоких температурах. – М.: Изд-во МГУ, 1967. – 325 с.
16. Апанович З.В., Винокуров П.С., Кислицина Т.А. Методы и средства визуализации информационного наполнения больших научных порталов // Вестник НГУ Серия: Информационные технологии. 2011. – том 9, выпуск 3, Редакционно-издательский центр НГУ. – С. 5-14.

НОВЫЕ ПОДХОДЫ К НОРМАЛИЗАЦИИ СЛОВАРЕЙ И УСТАНОВЛЕНИЮ ИДЕНТИЧНОСТИ СУЩНОСТЕЙ ПРИ ОБОГАЩЕНИИ КОНТЕНТА НАУЧНЫХ БАЗ ЗНАНИЙ[1]

З.В. Апанович, А.Г. Марчук
ИСИ СО РАН, НГУ, Новосибирск
apanovich@iis.nsk.su, mag@iis.nsk.su

В данной работе описаны подходы к решению проблем нормализации словарей, установления идентичности сущностей и фильтрации данных, возникающих в процессе использования данных из облака LOD для обогащения контента научных баз данных и знаний. В качестве тестовых примеров использовались данные открытого Архива СО РАН, структурированные при помощи онтологии ОНС, а также различные наборы библиографических данных, как *структурированные, так и слабо структурированные, включая текстовые.*

Ключевые слова: Связанные Открытые Данные, выравнивание онтологий, установление идентичности сущностей.

This paper describes approaches to the vocabulary normalization , identity resolution , and data filtering problems arising during the use of the LOD datasets to enrich the content of scientific knowledge bases. The dataset of the Open Archive of the Russian Academy of Sciences, as well as different bibliographic datasets are used as test examples. Texts of publications in natural language are used as an additional source of information.

Keywords: Linked Open Data, ontology alignment, identity resolution.

[1] Работа выполнена при финансовой поддержке РФФИ (проект № 4-07-00386) и проекта РАН 15/10.

Введение

В связи с бурно развивающимся направлением Semantic Web и его новой ветвью LOD (Связанные Открытые Данные) в Интернете становятся доступными большие объемы информации, посвященной различным научным направлениям. Облако LOD содержит в настоящий момент более 50 миллиардов троек RDF. С одной стороны, эти данные могут быть использованы для обогащения имеющихся семантических баз данных, с другой стороны, имеющиеся базы данные могут быть также полезны для уточнения информации, хранящейся в облаке LOD.

Один из проектов, осуществляемых в Институте систем информатики Сибирского отделения Российской академии наук (ИСИ СО РАН) направлен на обогащение Открытого архива СО РАН (http://duh.iis.nsk.su/turgunda/Home,http://duh.iis.nsk.su/Virtuoso Endpoint/Home/Samples) [1] данными облака Открытых Связанных Данных (Linked Open Data, LOD) [4]. Фрагмент страницы Открытого Архива посвященной академику А.П. Ершову, показан на рис. 1.

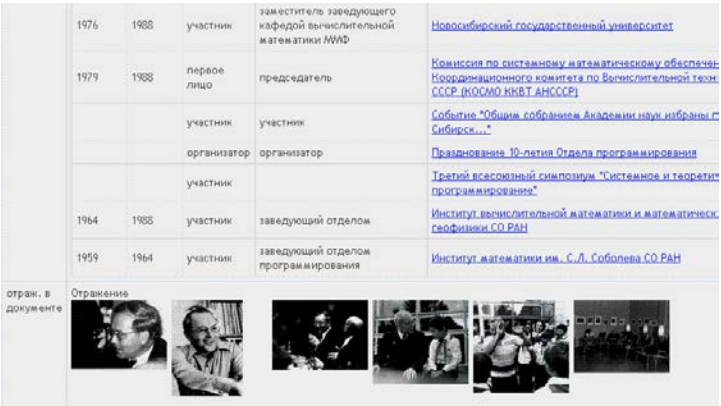


Рис. 1. Фрагмент страницы Открытого Архива, посвященной академику А.П. Ершову.

В работе [7] предложена четырехшаговая стратегия интеграции Связанных Данных в приложения. Помимо проблем, специфических для конкретного приложения, требуется решить проблему доступа к связанным данным (1), проблему нормализации словарей (2), установления идентичности сущностей (3) и фильтрации данных (4). Способы решения этих проблем варьируют в диапазоне от ручных до автоматизированных [6, 10, 8]. При этом такие проблемы как проблема установления соответствия между онтологиями, а также проблема объединения данных из разных наборов «еще находятся в детском состоянии» [9]. В работах [2, 3] подробно рассмотрены методы исследования онтологий и контента семантических систем при помощи различных методов визуализации разработанных в ИСИ СО РАН. В данной работе будут рассмотрены проблемы (2), (3) и (4) и продемонстрированы методы решения этих проблем.

В качестве тестовых примеров использовались данные открытого Архива СО РАН, структурированные при помощи онтологии ОНС [1] и различные наборы библиографических данных портала RKBExplorer.com, структурированные при помощи AKT Reference онтологии [11], а также источники текстовых данных такие как электронный архив А.П. Ершова

(http://ershov.iis.nsk.su/ershov/english/scient.html)

и электронная библиотека SpringerLink

(http://link.springer.com/).

1. Эксперименты по выравниванию онтологий

Одним из шаблонов построения онтологий в приложениях Semantic Web является то, что сущности, описанные, с помощью отношений в одних онтологиях могут быть описаны как экземпляры классов в другой онтологии. Этот шаблон называется «qualified relation" и компенсирует отсутствие атрибутов у предикатов RDF. Такой шаблон используется в онтологии ОНС и позволяет

описывать такие факты как "академик А.П. Ершов был главой отдела в Институте Математики СО АН СССР с 1959 по 1964 и руководителем отдела в Вычислительном центре СО АН СССР с 1964 по 1988 год " (см. Рис. 1). Для этого существует специальный класс *онс:participation,* соответствующий отношению *akt:has-affiliation* онтологии AKT Reference, экземпляры которого имеют атрибуты *from-dat*e и *to-date*. По тем же причинам, такие классы как *онс:dating*, *онс:naming*, *онс:authorship*, используются в онтологии ОНС вместо таких предикатов, как *akt:has-author, akt:has-date* или *akts:has-pretty-name*. Платой за расширение выразительных возможностей является усложнение проблемы интеграции данных, поскольку систематически возникает необходимость в установлении соответствия между различными группами классов и отношений двух онтологий с различными структурными свойствами. А именно, возникает необходимость установления соответствия между группой вида "Class1-relation1- Class2" одной онтологии и одной или несколькими группами вида "Class3- relation2-Class4-relation3-Class5" другой онтологии. Такая трансляция может быть осуществлена при помощи запроса SPARQL 1.1. Упрощенная версия шаблона этого запроса имеет следующий вид:

```
PREFIX iis:<http://iis.nsk.su#>
PREFIX akt:<http://www.aktors.org/ontology/portal#>
PREFIX akts:<http://www.aktors.org/ontology/support#>
CONSTRUCT {
_:p a iis:Class4.
_p: iis:relation2 ?instance1.
_p: iis:relation3 ?instance2.
}
WHERE {
        ?instance1 akt:relation1  ?instance2.
        ?instance1 a akt:Class1.
        ?instance2 a akt:Class2.
        }
```

Для упрощения задачи пользователя по написанию такого рода запросов нами разработана программа, которая позволяет генерировать SPARQL-запросы на основе визуализации онтологии. Пример установления такого соответствия показан на рис. 2. Сначала в интерактивном режиме устанавливается соответствие между двумя наборами классов и отношений, а затем автоматически генерируется шаблон SPARQL-запроса, осуществляющий трансляцию данных.

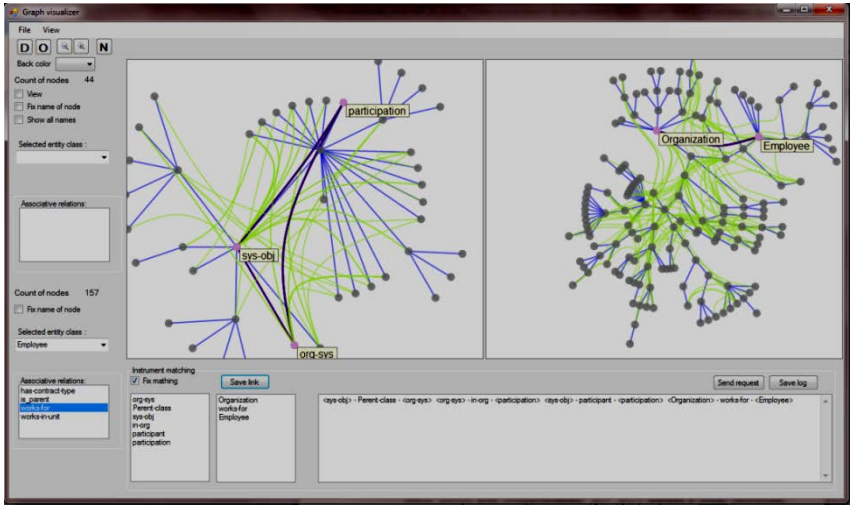


Рис. 2. Интерактивное установление соответствия между классами и отношениями двух онтологий.

2. Установление идентичности сущностей и фильтрация результатов

Важным этапом пополнения одной базы знаний при помощи другой является этап установления идентичности сущностей, то есть, генерация отношений вида *owl:sameAs*. В нашем случае необходимо правильно сопоставить персонам, описанным в От-

крытом архиве СО РАН информацию про эти персоны, взятую из других семантических систем. Проблема осложняется тем, что в случае Открытого архива используются русскоязычные имена персон, а в большинстве систем, с которыми мы работали, используются англоязычные имена тех же самых персон. Конечно, может возникнуть вопрос, почему бы не воспользоваться русскоязычным источником данных, например данными научной электронной библиотеки elibrary.ru? Эта библиотека представляет персонифицированную информацию по российским исследователям, но, к сожалению, срок развития этой электронной библиотеки значительно уступает периоду времени, охватываемому фотоархивом СО РАН. Поэтому elibrary.ru может быть весьма полезной при идентификации персон и их публикаций за последние 10-15 лет, но она становится мало полезной при изучении публикаций таких персон, как академик Андрей Петрович Ершов. Нам не удалось обнаружить в elibrary.ru информации про таких людей, как А.П. Ершов, Б.А. Трахтенброт, В.Е. Котов и многих других исследователей, заложивших основы советской и российской информатики.

Про используемые наборы данных. Первоначально эксперименты по интеграции осуществлялись на наборах данных Открытого архива и различных наборах данных RKBExplorer.com, который интегрирует данные из многих известных научных электронных библиотек. Так, например, подмножество dblp.rkbexplorer.com соответствует набору данных DBLP Computer Science Bibliography (http://www.informatik.uni-trier.de/~ley/db/index.html), и содержит информацию о публикациях с 1936 по настоящее время. Набор данных acm.rkbexplorer.com соответствует набору данных Association for Computing Machinery (ACM), а публикации, описанные в этом множестве, взяты из цифровой библиотеки dl.acm.org. Набор данных

citeseer.rkbexplorer.com взят из цифровой библиотеки CiteSeerx (http://citeseer.ist.psu.edu). Набор данных ieee.rkbexplorer.com содержит информацию про публикации IEEE (http://ieeexplore.ieee.org/Xplore/guesthome.jsp) и т.д. Следует отметить, что количество наборов данных RKBExplorer.com постоянно расширяется. В последнее время туда добавлена информация о сотрудниках и о публикациях многих европейских университетов, о европейских проектах (cordis.rkbexplorer.com) и много другой интересной и полезной информации.

Важно также отметить, что информация, хранимая в наборах данных rkbexplorer.com, не полностью совпадает с информацией, хранимой в библиотеках-прообразах этих наборов данных. Например, набор данных dblp.rkbexplorer.com содержит те же самые публикации, что и его прообраз база данных DBLP Computer Science Bibliography, но эти наборы данных используют разные эвристики при идентификации синонимов и омонимов на уровне персон. Тем не менее, многие ошибки идентификации персон, первоначально обнаруженные нами в базе данных RKBExplorer.com, имели в своей основе ошибки идентификации, совершенные в ее базе-прообразе DBLP Computer Science Bibliography. Более того, в процессе наших экспериментов мы научились обнаруживать в хорошо известных иностранных системах поддержки научных исследований систематические ошибочные данные, касающиеся публикаций различных персон с русскоязычными именами, в том числе, публикаций такого известного человека, как академик Андрей Петрович Ершов. Для понимания масштаба проблемы достаточно посмотреть страницу, посвященную персоне по имени Andrei P. Ershov электронной библиотеки Microsoft Academic Search (http://academic.research.microsoft.com/Author/1905518/andrei-p-ershov).

Поэтому остановимся на вопросе сопоставления персон, описанных в Открытом архиве СО РАН с данными dblp.rkbexplorer.com подробнее. В Открытом архиве все персоны описаны при помощи атрибута *онс:name*. Он имеет формат <Фамилия, Имя Отчество> и два варианта написания: русскоязычный и (иногда) англоязычный вариант. При этом возможно много вариантов латинского написания русских имен. Русская фамилия Ершов может писаться как Yershov так и Ershov. Для повышения полноты сравнения русскоязычных вариантов фамилий с их латинскими эквивалентами необходимо генерировать все возможные варианты транслитерации русских букв в латинские. Вторым источником многообразия имен является то, в латинских информационных системах имена не всегда нормализованы и имеют много вариантов. Это может быть <Имя Фамилия >, <Имя Первая буква Отчества Фамилия> <Первая буква имени Первая буква Отчества Фамилия > и др. В качестве первого шага необходимо по нормализованному русскому имени сгенерировать все возможные варианты транслитерации, а затем соответствующие им варианты атрибута full-name и сравнить эти варианты с теми именами, которые содержатся в базе данных RKBExplorer.com.

Конечно, методы сравнения строковых значений не являются новой областью исследования и имеют обширную библиографию [5, 15, 16]. Поэтому достаточно сказать, что для сравнения вариантов имен используется токенизированный вариант метрики Jaro-Winkler [5]. То есть, сначала сравниваются фамилии из данных RKBExplorer и открытого архива с очень высоким пороговым значением, а затем с более низким порогом сравниваются варианты имени и отчества. Не точное совпадение имен на данном этапе является необходимым условием, обеспечивающим полноту поиска, поскольку в латинских вариантах данных постоянно обнаруживаются не предусмотренные правилами варианты

транслитерации. Так в ходе этого эксперимента в базе данных RKBExplorer.com были обнаружены публикации, соответствующие персоне A. P. Yershóv.

После завершения этапа генерации кандидатов программно строится SPARQL-запрос, который по списку фамилий, идентифицированных как совпавшие в обеих базах знаний, выдает список идентификаторов RKBexplorer и атрибутов *akt:full-name*, соответствующих каждому варианту атрибута *онс:name*, и список публикаций, связанных с конкретным идентификатором. При этом каждому из этих имен Открытого архива соответствует несколько разных идентификаторов персон, каждый из которых имеет свой собственный атрибут *akt:full-name* и отдельный список публикаций.

Так, например, для персоны «Андрей Петрович Ершов» из Открытого архива в базе данных dblp.rkbexplorer.com было обнаружено 18 экземпляров персон с атрибутом *akt:full-name* равным «Andrei P. Ershov», при этом большинству найденных идентификаторов соответствовало по одной публикации. Помимо этого обнаружились другие персоны с похожими атрибутами *akt:full-name* и своими наборами публикаций. Две публикации соответствовали двум персонам с разными идентификаторами и атрибутом *akt:full-name* равным “A. P. Yershóv”, одна публикация – персоне с атрибутом *akt:full-name* равным “A. Ershov”, одна публикация – персоне с атрибутом *akt:full-name* , равным “A. Yershov”, две публикации - персоне с атрибутом *akt:full-name* равным “Andrew Ershov”.

Возникает два вопроса:

1) Какие из вышеупомянутых идентификаторов соответствуют одному и тому же физическому объекту, и, стало быть, могут быть связаны отношением *owl:sameAs*, а какие из них описывают разные физические объекты?

2) Все ли публикации, приписанные персоне с одним идентификатором, принадлежат одному и тому же физическому объекту?

Очевидно, что ответы на эти вопросы имеют существенное влияние, например, на подсчет индекса цитирования. Наши эксперименты показали, что, как правило, публикации персон, имеющихся в Открытом архиве, бывают разбросаны по нескольким разным персонам с разными идентификаторами из базы данных RKBExplorer, и иногда одному и тому же идентификатору приписываются публикации разных физических персон. Аналогичный феномен наблюдается и в тех электронных библиотеках, из которых были получены наборы данных RKBExplorer.com.

Стандартным способом идентификации персон считается их идентификация не только по имени, но и по адресу электронной почты или по персональной странице в Интернете. К сожалению, для многих персон, составляющих наполнение Открытого архива, такой информации не существует. Кроме этого, эти данные могут меняться вместе с изменением места работы. Зато в Открытом архиве поддерживается информация о местах работы персон, с указанием периода работы в конкретной организации, как это показано на примере А.П. Ершова. Эту информацию можно сравнивать с информацией о месте работы автора публикации, если таковая доступна. Еще одна эвристика, используемая для идентификации одной и той же персоны, являющейся автором нескольких публикаций, основана на объединении в одну персону всех авторов, имеющих одних и тех же соавторов [14]. Такую проверку можно осуществлять, зная только авторов и названия публикаций.

Кроме этого, авторы научных публикаций часто ссылаются в списках литературы на свои прежние публикации, поэтому возникает желание попробовать объединить автора текущей публи-

кации и автора с похожим именем, указанного в списке литературы, в одну персону. В AKT Reference ontology имеется отношение *akt:cites-publication-reference* между экземплярами классов *akt:Person* и *akt:Publication-Reference*, которое потенциально можно использовать для такого способа идентификации персон. Хотя информация этого типа присутствует в наборах данных RKBExplorer.com, она оказалась очень неполной. Для многих персон из Открытого Архива, редко имеется информация более чем о двух ссылках из списка литературы каждой из статей. Поэтому сеть цитирования, сгенерированная на основании этой информации, получается несвязной с большим количеством изолированных вершин. Ввиду неполноты имеющейся структурированной информации было принято решение попробовать извлекать информацию о списках цитирования из полу-структурированных и неструктурированных, текстовых источников публикаций. Поэтому в настоящий момент выполнена первая серия экспериментов по идентификации персон не только на основании информации из наборов данных Открытого архива и rkbexplorer.com, но также с использованием текстовой информации, доступной в Интернете. Выполняется два вида проверки с привлечением текстовой информации.

Проверка по месту работы. Дата каждой публикации, извлекается из RKBEplorer при помощи SPARQL-запроса, по этой информации находится место работы персоны, соответствующее указанному периоду в Открытом архиве, и найденное место работы сравнивается с местом работы, указанным в тексте публикации (если имеется).

Проверка по списку цитируемой литературы. Для идентификации всех публикаций, принадлежащих одной персоне, предлагается построить граф самоцитирований и выделить в нем связные кумпоненты. Упрощенная процедура состоит в том,

что варианты имени тестируемого автора каждой публикации сравниваются с именами авторов публикаций, извлекаемых из списка цитируемой литературы. Если обнаруживается приблизительное) совпадение имен, то текущая публикация объединяется в одно множество с названиями публикаций одноименного автора из списка цитируемой литературы. Затем та же самая процедура применяется к добавленным публикациям. Основной трудностью этого применения этого метода является не сама процедура построения графа самоцитирований, а поиск подходящих источников текстовой информации. Первые эксперименты были проведены с текстами публикаций, найденными вручную в Интернете. Следующая группа экспериментов была осуществлена с текстами электронной библиотеки SpringerLink(link.springer.com), которая предоставляет информацию, достаточную для наших экспериментов. Имя автора, название публикации, место работы автора и список цитирований. Поскольку на сайте ИСИ СО РАН также имеется библиографический указатель публикаций А.П. Ершова с текстами этих публикаций, у нас оказалась возможность проверить предлагаемый подход. Приведем несколько примеров, демонстрирующих нашу идею.

Пример 1. (Персоны A. Ershov, A. P. Yershóv)

Для персоны, обозначенной в наборе данных rkbexplorer.com идентификатором

http://dblp.rkbexplorer.com/id/people-caaa3c31d151bb7e83f5d6b37aa9de2e-855ed1acf622531e224916a6afa900fb

с атрибутом *akt:full-name*, равным «A. Ershov», указана единственная публикация:

A. Ershov, A. Nariniany, I. Mel'chuk «RITA - An Experimental Man-Computer System on A Natural Language Basis».

Принадлежит ли эта публикация академику А.П. Ершову, и, стало быть, потенциально она должна попасть в тот же самый список, что и все публикации, приписанные в базе данных dblp.rkbexplorer.com персоне (персонам) с атрибутом *akt:full-name*, равным Andrei P. Ershov? Поскольку эта публикация имеется в электронном архиве А.П. Ершова, для утвердительного ответа достаточно автоматического сравнения названия этой работы с названиями работ, указанными в архиве А.П. Ершова.

Более сложная процедура состоит в извлечении из текста публикации (доступной в Интернете) имени автора, названия работы, места работы, и списка цитирований. Название работы совпадает с названием, указанным в тексте, в качестве автора в тексте работы указано «A.P. Ershov», а в качестве места работы, «Computing Center, 630090, Novosibirsk, USSR», фигурирующий в Открытом архиве для периода публикации (1975). На основании этих данных также можно выдать утвердительный ответ на наш вопрос. Но анализ текста публикации имеет дополнительную ценность тем, что позволяет правильно идентифицировать еще одну работу, которая в базе данных rkbexplorer.com приписана персоне с идентификатором

http://dblp.rkbexplorer.com/id/people-5c092dbe2d3183e77182210f00d40b4-faefb1714b3e31775353e4fae216a87e

и атрибутом *akt:full-name*, равным «A. P. Yershóv» и публикацией «One View of Man-Machine Interaction». На основании информации о цитировании отношением *owl:sameAs* можно связать две персоны с разными идентификаторами и разными атрибутами имени, после чего сопоставить обе работы персоне «Ершов, Анд-

рей Петрович» Открытого архива. Проверка названия по электронному архиву А.П. Ершова подтверждает правильность такой гипотезы.

Пример 2. (Публикация, не указанная в архиве А.П. Ершова) На dblp.rkbexplorer.com для персоны с идентификатором http://dblp.rkbexplorer.com/id/people-e1ac8593dbc7db6ec5766ea313914be4-ea3c7cab911196a0b2f40f9f0a1242e2 и атрибутом *akt:full-name,* равным Andrei P. Ershov указана публикация

«Axiomatics for memory allocation».

Поиск этого названия в электронном архиве А.П. Ершова дает отрицательный результат. Зато проверка по электронной библиотеке link.springer.com показывает, что место работы автора этой статьи равно “Computing Center, 630090, Novosibirsk, USSR”, в списке цитирований указаны работы уже идентифицированные как принадлежащие академику А.П. Ершову, и помимо этого обнаруживается наличие цитирований этой работы в публикациях академика А.П. Ершова (как в базе link.springer.com так и в архиве А.П. Ершова). Например, название этой работы фигурирует в списке цитирований публикации «The Transformational Machine: Theme and Variations.», датированной 1981 годом и уже приписанной академику А.П. Ершову. На основании этой информации предлагается добавить эту публикацию к Открытому архиву, а персону с указанным идентификатором связать отношением *owl:sameAs* с другими персонами, идентифицированными ранее как академик А.П. Ершов. Схема этой ситуации показана на рис. 3. Самая левая колонка соответствует данным электронного архива А.П. Ершова, центральная колонка- данным link.springer.com, а самя правая колонка – набору данных dblp.rkbexplorer.com.

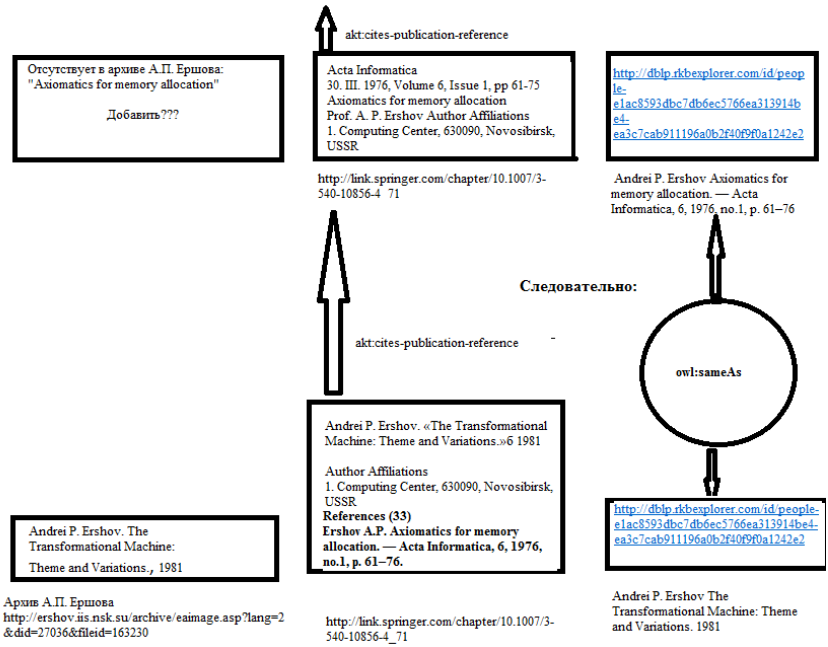


Рис. 3. Связь между двумя публикациями в трех базах данных.

Заключение

В данной работе рассмотрены проблемы обогащения научных баз знаний при помощи контента библиографических порталов из облака LOD и подходы к их решению. Соответствие между наборами данных, основанных на этих онтологиях, устанавливается при помощи SPARQL-запросов, которые генерируются на основе визуализации онтологий.

Также продемонстрировано, что обычные инструменты, применяемые для установления идентичности сущностей на основе

строковых метрик сходства, не позволяют различать синонимы и омонимы при идентификации персон. Предложены новые эвристики, одна из которых основана на выделение связных компонент в графе самоцитирований. Эксперименты показали, что эта процедура не всегда позволяет однозначно определить идентичность персон и в дальнейшем планируется ее развитие в нескольких направлениях. Во-первых, предполагается строить более разветвленную сеть цитирования, учитывающую не только автора, нуждающегося в идентификации, но и его соавторов. А во-вторых, включать в эту сеть цитирования ссылки на всех авторов, а не только на авторов с похожими именами.

В результате проведенных экспериментов удалось не только правильно расклассифицировать публикации А. П. Ершова, разбросанные по нескольким десяткам персон портала www.rkbexplorer.com, но и обнаружить несколько публикаций А. П. Ершова, не отраженные в библиографическом списке электронного архива А.П. Ершова.

Список литературы

1. Марчук А.Г., Марчук П.А. Особенности построения цифровых библиотек со связанным контекстом //Труды RCDL'2010. Двенадцатая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» Казань, Казанский университет , 2010. – С. 19-23.
2. Zinaida Apanovich and Alexander Marchuk Experiments on using the LOD cloud datasets to enrich the content of a scientific knowledge base P.Klinov and D.Mouromtsev (Eds.) KESW 2013, CCIS 394 pp. 1-14, Springer Verlag Berlin Heidelberg 2013
3. Apanovich Z. V., Vinokurov P. S. An extension of a visualization component of ontology based portals with visual analytics facilities. // Bulletin of NCC . – Issue 31. – 2010. – pp. 17-28.
4. Bizer, C., Heath, T. , Berners-Lee, T. Linked Data - The Story So Far. //Int. J. Semantic Web Inf. Syst., 5 (3). 2009. P. 1-22.

5. William W. Cohen, Pradeep D. Ravikumar, Stephen E. Fienberg: A Comparison of String Distance Metrics for Name-Matching Tasks.IIWeb 2003: 73-78.
6. Isele R., Jentzsch A., Bizer Ch. Silk Server - Adding missing Links while consuming Linked Data// 1st International Workshop on Consuming Linked Data (COLD 2010), Shanghai, November 2010.
7. Schultz A. et al. How to integrate LINKED DATA into your application //Semantic technology & Business Conference, San Francisco, June 5, 2012.
http://mes-semantics.com/wp-content/uploads/2012/09/Becker-etal-LDIF SemTechSanFrancisco.pdf.
8. Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges IEEE Transactions on Knowledge and Data Engineering, 25(1) pp. 158-176 (2013)
9. Tramp, S., Williams, H., Eck, K.: Creating Knowledge out of Interlinked Data: The LOD2 Tool Stack http://lod2.eu/Event/ESWC2012-Tutorial.html.
10. Ngomo A.-C. N., Auer S.: LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. //IJCAI 2011: Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011 pp. 2312-2317.
11. Описание AKT Reference ontology:
http://www.aktors.org/ontology.
12. Набор данных RKBExplorer rkbexplorer.com/.
13. Набор данных DBLP: http://dblp.rkbexplorer.com/.
14. Michael Ley: DBLP - Some Lessons Learned. PVLDB 2(2): 1493-1500 (2009)
15. Peter Christen A Comparison of Personal Name. Matching: Techniques and Practical. Issues. TR-CS-06
https://digitalcollections.anu.edu.au/bitstream/1885/44521/3/TR-CS-06-02.pdf
16. Ahmed K. Elmagarmid,Panagiotis G. Ipeirotis, Vassilios S. Verykios, Duplicate Record Detection: A Survey Journal IEEE Transactions on Knowledge and Data Engineering, Volume 19 Issue 1, January 2007, Pages 1-16.

DATA INTENSIVE SCIENCE AND E-INFRASTRUCTURE FOR ACCESS TO SCIENTIFIC DATA

Belov A.F.[a], *Kudashev E.E.*[a], *E.B. Kudashev*[a, b]
[a]Space Research Institute of Russian Academy of Sciences,
[b]Lomonosov Moscow State University.
kudashev@iki.rssi.ru

The paper offers a short review of some recent advances to the research of digital information and Earth Observation data from satellites. We consider evolution the Earth Observation Data Archives. The volume of Earth-Observation data from the European Space Agency's satellites passed three PB in 2007 and is in constant increase. We also consider the creation of e-Infrastructure for scientific data and present APARSEN Project. APARSEN stands for Alliance Permanent Access to the Records of Science in Europe Network. A key issue here is how APARSEN are developed its Virtual Centers of Excellence for Digital Preservation and how it will ensure its ongoing viability.

Keywords: digital content, scientific data, big data science, Earth-Observation data, big research data, satellite data archives, e-Infrastructure, Virtual Center of Excellence for digital preservation.

В статье предлагается обзор состояния современных исследований цифрового контента и проблемы больших данных из космоса. Рассматривается формирование е-Инфраструктуры для доступа к научным данным. Представлен проект APARSEN Европейской Комиссии, разрабатывающий информационную инфраструктуру, в структуру которой входят Виртуальные центры развития научных ресурсов.

Ключевые слова: цифровой контент, научные данные, проблема больших данных, данные Исследования Земли из космоса, архивы спутниковых данных, цифровая инфраструктура, Виртуальный Центр развития цифровых данных.

1. Introduction. Big Data Science

In the digital age there is a dramatic increase in the volume of digital content: the science is faced with a process of constant income of huge data and with an exponential growth of scientific data. In the informational society the role of information and data processing becomes a dominant factor. We begin to see what some have called a "*fourth paradigm of science*".

Science has entered the modern stage: development of e-Science. The main products of the contemporary industry are information and knowledge. The main informational challenge is how to gather and store data and how to exchange knowledge. Apart from the data amount issue, the diversity of the information and requirements to data accessibility have increased, too. Generalizing these requirements, Gartner [1] proposed a threefold definition encompassing the “three Vs”: **Volume** (remarks upon the increasing size of data) - **Velocity** (the increasing rate at which it is produced) - **Variety** (the increasing range of formats and representations employed). Gartner now suggests to include a fourth V: **Veracity** which includes questions of trust and uncertainty with regards to data and the outcome of analysis of that data.

Whereas the great astronomer Tycho Brahe (1546–1601) was happy with size of data of about 500Kb, Google now processes 24PB per day; the volume of social networks data is over PB threshold. Taking into account the *Volume* of created digital content, we can tell now that *Fourth Research Paradigm* today arises. This is Data-Intensive Science - beyond "Observation" (First research paradigm), "Theory" (Second paradigm) and "Simulation" (Third Research Paradigm: Computational Science). For the first time, large-scale and complex "whole body" solutions become possible for some of society's Grand Challenges of energy and water supply, global warming and healthcare [2].

The present situation affects various fields of knowledge operating with huge volumes of information. Above all, these are Physics of elementary particles and High Energy Physics (Open Grid); Astronomy and Astrophysics (Virtual Observatory); Earth Sciences (Earth Observation); Biology (Bioinformatics).

2. Earth Observation Satellite Data

Consider evolution the Earth Observation Data Archives. Remote Sensing from satellites allow a global perspective on observation of the Earth to be developed. After a long, slow rise since 1986, the volume of Earth-Observation data from the European Space Agency's satellites passed three PB in 2007and is in constant increase. New datasets, coming from the future Earth Explorer GMES missions will contribute to increase the volume in the years to come [3]. Prediction of GSCB (Ground Segment Coordination Body) is that this volume will exceed 20 PByte for 2020.

Example 1. Pacific Russian Centre for Satellite Data at Vladivostok. The Pacific Centre provides the studies on physics of the ocean and atmosphere and concludes reception, processing and archiving data from satellites AQUA, TERRA, MTSAT-1R, FY-1Dand NOAA. The total volume of EO data archives of Pacific Russian Centre in 2011 exceeds 10TB.

Example 2. German Aerospace Center (DLR) created the National Satellite Data Archive. The Earth Observation Center (EOC) at DLR is the Center of competence in Germany, providing expertise in Earth Observation research and development activities, as well as operational tasks for data reception, processing and archiving. The powerful and centralized archive at the DLR Earth Observation Center has proven its stability and flexibility to allow Long Term Data Preservation over more than 20 years with nearly exponentially growing data capacity. In 2012 input/output data rates have grown to be beyond 100

MB/s, but the disk drives and networks have also grown. Archiving capacity of National Satellite Data Archive which is Remote Sensing Data Center is 2,2 PB [4].

Where will this lead? **Growing satellite data USA**. The volumes of NASA and NOAA archives have grown from 1 PB in 2000 to 10 PB in 2011 annually. Total volume of NOAA Archives will exceed 100 PB [5].

3. e-Infrastructure for Access to Scientific Data

The EU is now very concerned in promoting and funding activities to organizations that are providing services and tools to the research, library, public and commercial communities. This information then allows us to put the services on the web site that research organizations can offer to both **Alliance for Permanent Access** (**APA**) Members, the APARSEN group (**Alliance for Permanent Access to the Records of Science in Europe Network)** and all global organizations engaged and interested in preserving their digital assets [6]. This enables all the organizations involved to promote their particular expertise and tools and provides them with an added revenue stream to a global audience through the promotional activities of the APA.

What is APARSEN?

A **Network of Excellence** in digital preservation. Funded by European Commission

- 7th Framework Programme — Digital Libraries and Digital Preservation
- January 2011 to December 2014 (4 years)
- Coordinated by Science and Technology Facilities Council (UK)

Fig. 1. Alliance Permanent Access to the Records of Science in Europe Network.

APARSEN stands for **Alliance Permanent Access to the Records of Science in Europe Network**. The APARSEN project involves a very broad set of organizations from academia, research laboratories, major national libraries, national membership organizations and industry. APARSEN defines four topics of Digital Preservation in which it has undertaken research in state of the art and gap analysis: as a first step in integration:

ACCESSIBILITY, USABILITY, TRUST, SUSTAINABILITY.

ACCESS. Without the ability to provide access to digital objects and data within a repository digital preservation fails. APARSEN provides some of the answers as to how access can be maintained through work on identifiers and citability, data policies and governance, and

digital rights. Why do we need Persistent Identifiers for digital resources? How can I refer to a digital resource in a sure, unique, stable and global way? What are the challenges related to the preservation of digital rights information? What are the recommendations in dealing with DRM-protected material? What kind of issues should be covered in data policies? What are the recommendations for governance structures and data policies?

USABILITY. APARSEN examined issues relating to interoperability, intelligibility and scalability to help develop a roadmap for how data and digital objects can remain useable and understandable in the long-term. Which approaches for interoperability currently exist? How is interoperability related to Digital Preservation? How can I plan for preservation systems at scale with the rapidly growing amount of data and complexity? How I can use this digital object? How can I perform this task? What are the recommendations in order to help to plan for scalability?

TRUST. The issue of trust is at the core of all digital preservation work [7]. How do we trust that a digital object or data is what it claims to be? APARSEN addresses the issues of trust through work on testing environments, authenticity and provenance, annotation and data quality, and repository certification. Has data been preserved properly? Is my digital repository trustworthy – a framework for audit and certification of digital repositories? Is being audited worthwhile? Is my data of a good quality? How can peer review of data be carried out? Has the data been changed in any way? Am I being directed to the right object?

SUSTAINABILITY. Digital repositories must be sustainable if they are to meet their preservation commitments in the medium to long-term. Work within APARSEN on topics such as cost/benefit analysis, preservation services, storage, and business cases brings together disparate research around this topic to help us build sustainable

repositories. What are the main requirements for economically sustainable digital preservation? How can I estimate the resources needed? How can I justify the resources needed for digital preservation? How can I keep the required resources under control? How can I plan to cope as the volume increases over time?

APARSEN has been working to collect, evaluate and develop key answers to these questions. The objective of this project may be simply stated, namely to look across the excellent work in digital preservation which is carried out in Europe and to try to bring it together under a common vision. The success of the project are based on the subsequent coherence and general direction of travel of research in digital preservation, with an agreed way of evaluating it and the existence of an internationally recognized Virtual Centre of Excellence. APARSEN brings together expertise from across Europe including 31 partners (Rutherford Applet Laboratory (UK), Alliance for Permanent Access, CERN, IBM, ESA, Airbus, IKI, British Library and other) and will bring coherence, cohesion and continuity to research into barriers to the long term accessibility and usability of digital information and data. It will defragment ideas about and create a Virtual Centre of Excellence for digital preservation.

4. Leading to a "Centre of Excellence"

4.1. The Virtual Centre of Excellence (CoE).

CoE primarily targets organizations that have digitally encoded information which they wish to or they must preserve [7]. In addition it will also support related stakeholder organizations including research institutions, private companies, third-party archives, professional societies working with domain and preservation experts to ensure that personnel are fully equipped with the technical skills needed for selecting, curating, and preserving materials.

The CoE provides **Training** , **Consultancy**, **Tools**, **Services** to support organisations which need to ensure that their digital resources remain understandable and usable.

4.2. Training events.

There are a number training options available from the Centre of Excellence in the form of:

- Face to face courses
- Workshops
- Clinics
- Summer school
- Online training portal

Training can be customized to meet participants learning objectives.

4.3. Online Training Portal.

Course material will be available online in the form of training modules via the APA's website. Introductory presentations are free of charge and are available on the APA website. A small fee is payable on accessing more detailed course material.

- Training Modules. Training modules are presented within specified course outlines as examples.
- Modules are accessible individually on the portal.
- Sample content is provided, for the full set of modules access the portal which will be fully populated by the end of 2014.

Example of Training Module is presented below.

4.4. Tools.

APARSEN has been used Toolkits from **SCIDIP-ES** (Science Data Infrastructure for Preservation – Earth Science) project including:

- Representation Information Toolkit
- Authenticity Toolkit

– Preservation Strategy Toolkit

... all consistent with OAIS model

plus other offerings of Centre of Excellence members offered through the marketing channels of the CoE.

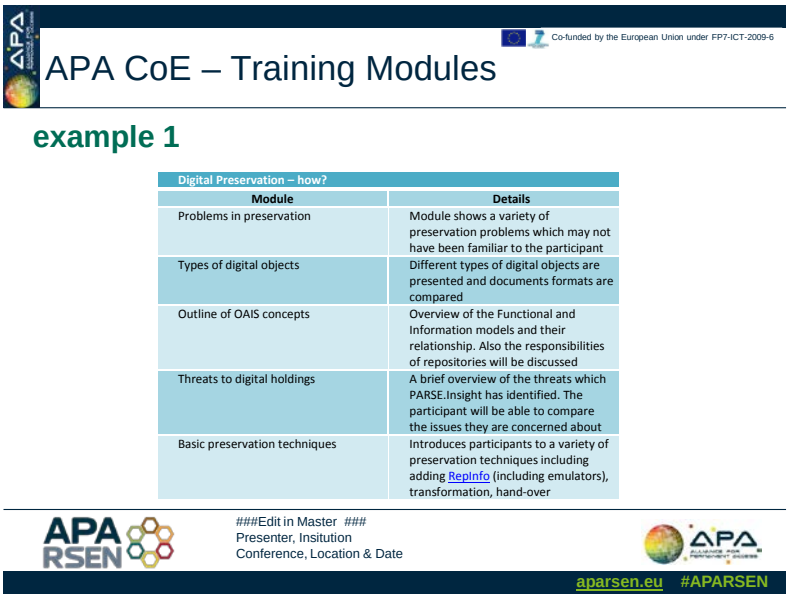


Fig. 2. Example of Training Module.

4.5. Centre of Excellence

- Is a membership organisation
- Provides expertise about digital preservation
- Obtains value from digitally encoded information

The unique selling point is that Centre of Excellence offering is coherent and consistent and should be applicable to any type of digital object and it is backed by the combined experience of the

Digital Preservation pioneers both in the research field and, more importantly, as worldwide earliest adopters of Digital Preservation practices**.**

The Centre of Excellence [8] primarily targets organizations that have digitally encoded information which they wish to or they must preserve. The CoE provides training, consultancy, tools and services to support organizations which need to ensure that their digital resources remain understandable and usable. In addition it will also support related stakeholder organizations including research institutions, private companies, third-party archives, professional societies working with domain and preservation experts to ensure that personnel are fully equipped with the technical skills needed for selecting, curating, and preserving materials, as well as fund internal preservation and access activities as core infrastructure.

- Centres of Excellence provide solutions for data management and long term preservation with big research data
- Data management training and planning
- Targeted workshops to find needs of Big Industry
- Visibility for members of CoE
- Help for audit and certification of trustworthy digital repositories
- Advice and consultancy on outsourcing, tools and services

5. Conclusion

Specific challenge: Development and deployment of integrated, secure, permanent, on-demand service-driven, privacy-compliant and sustainable.

e-Infrastructures incorporating advanced computing resources and software are essential in order to increase the capacity to manage, store and analyses extremely large, heterogeneous and complex datasets, including text mining of large corpora.

These e-infrastructures need to provide services cutting across a wide-range of scientific communities and addressing a diversity of computational requirements, legal constraints and requirements, system and service architectures, formats, types, vocabularies and legacy practices of scientific communities that generate, analyse and use the data.

- Services to ensure the quality and reliability of the **e-Infrastructure**, including certification mechanisms for repositories and certification services to test and benchmark capabilities in terms of resilience and service continuity of **e-Infrastructures**
- Increased availability of scientific data for scientific communities independently of them having already embraced or not e-science; this will be measured by cross-border data traffic over the research networks in Europe as a proxy.

Acknowledgement

The reported study was supported by RFBR project 11-07-90404-a and project No 14-07-00032.

References

17. Ward,J. S. and BarkerA.Undefined by Data:A Survey of Big Data Definitions (2013). See: *http://arxiv.org/pdf/1309.5821v1.pdf.*
18. Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High level Expert Group on Scientific Data. European Union (2010).
19. PinnaG.M., MbayeS. SAFE – Standard Archive Format for Europe // In: PV2007 Conference Abstracts. DLR Oberpfaffenhofen – Munih – Germany (2007).
20. C.Reck et all. Behind the Scenes at the DLR national Satellite Data Archive, a Brief History and Outlook of Long Term Data Preservation. In: Proc. of PV2011 Conf.. CNES. Toulouse, France (2011).

21. Ramapriyan H.K. Development, Operation and Evolution of EOSDIS - NASA's major capability for managing Earth science data. CENDI/NFAIS Workshop on Repositories in Science & Technology: Preserving Access to the Record of Science November 30.2011.
22. APARSEN Project (2011-2014).
See: http://www.allianpermanentaccess.org.
23. See http://www.trusteddigitalrepository.eu.
24. Kudashev E.B., Popov M.A. Towards Virtual Data Centers for Remote Sensing Data / /In: Infrastructure of Satellite GIS Resources and their Integration. – Kiev, National Academy of Sciences of Ukraine, 2013.

ИНТЕГРИРОВАННАЯ ИНФОРМАЦИОННО-ВЫЧИСЛИТЕЛЬНАЯ ИНФРАСТРУКТУРА ИРКУТСКОГО НАУЧНО-ОБРАЗОВАТЕЛЬНОГО КОМПЛЕКСА

Бычков И.В., Маджара Т.И., Ружников Г.М.

Институт динамики систем и теории управления СО РАН
(ИДСТУ СО РАН)
bychkov@icc.ru, taras@icc.ru, rugnikov@icc.ru

В статье рассматриваются предпосылки, история и перспективы развития одного из успешно реализованных подходов к построению региональной интегрированной информационно-вычислительной инфраструктуры научных организаций.

Ключевые слова: информационно-телекоммуникационные инфраструктуры поддержки междисциплинарных исследований, корпоративные сети и маршрутизация, IP-телефония, высокопроизводительные вычислительные системы, проблемно-ориентированные базы данных.

The article considers the background, history and prospects for the development of a successfully implemented approach to the construction of a regional integrated information infrastructure for research organizations.

Keywords: information-computing-telecommunication infrastructures for support of interdisciplinary research, corporate networks and routing, high-performance computing system, problem-oriented databases.

В институтах Иркутского научного центра (ИНЦ) СО РАН и вузах региона ведутся научные исследования, базирующиеся на уникальных проблемно- и предметно-ориентированных базах пространственных данных по ландшафтам и геосистемам, картографированию природы, хозяйства и населения Сибири, геологической среде и сейсмическим процессам, геохимии окружающей среды и осадочных бассейнов, электроэнергетическим и трубо-

проводным системам, биоразнообразию фауны и флоры оз. Байкал, физиологии растений, молекулярной биологии и экологии растительных организмов, дистанционному зондированию поверхности Земли и т.д.

Особенностью этих информационных ресурсов является их разноформатность, пространственно-временной характер, разнообразие используемых технологий обработки, отсутствие интеграции и удаленного доступа. Поэтому для перехода на новый технологический уровень проведения комплексных междисциплинарных исследований необходимо создание распределенного информационно-телекоммуникационно-вычислительного ресурса, построенного в рамках единого подхода на основе существующих международных стандартов, сопровождение и развитие которого могли бы поддерживать по своему профилю ведущие институты и вузы региона. При этом стоит задача не столько формирования новых ресурсов, сколько разработки технологии комплексирования и хранения уже имеющихся пространственных данных (ПД), а также создания технологии удалённого доступа к ПД, их интеллектуального анализа и внедрения высокопроизводительных информационно-вычислительных комплексов их обработки.

Организация обмена результатами научных исследований и предоставление доступа к общим информационным ресурсам, таким как тематические БД, электронная топографическая основа исследуемых территорий, геопорталы, каталоги метаданных, электронные библиотечные системы, данные удаленных полевых стационаров и данные ДЗЗ повышает качество представления результатов научных исследований, уровень согласованности создаваемых данных, что, в свою очередь, облегчает использование этих данных при проведении междисциплинарных исследований. Все это обусловило необходимость внедрения в Байкальском регионе современных форм коллективного использования информационно-вычислительных и коммуникационных ресурсов. Тех-

нологической основой построенной инфраструктуры стала интегрированная информационно-вычислительная сеть (ИИВС) Иркутского научно-образовательного комплекса (ИРНОК). Основная цель создания ИИВС – информационно-телекоммуникационная поддержка деятельности научно-образовательного сообщества Байкальского региона, заключающаяся в концентрации ресурсов путем объединения корпоративных сетей институтов, вузов и специализированных ресурсных центров с учетом использования современных технологий.

Началом создания инфраструктуры ИИВС считается 1994 год, когда в ИДСТУ СО РАН был разработан системный проект и начата его реализация. Целью проекта являлось объединение корпоративных информационно-вычислительных ресурсов научных и образовательных учреждений Байкальского региона высокоскоростными каналами связи, а также поддержка и развитие доступа в Интернет. Для реализации проекта были привлечены финансовые средства Министерства науки и технологий РФ, РФФИ, СО РАН, внебюджетные средства ИДСТУ СО РАН, институтов ИНЦ СО РАН и ряда государственных вузов, впоследствии вошедших в ИРНОК.

Как транспортная инфраструктура, ИИВС, построенная на основе многомодовых волоконно-оптических линий связи была введена в эксплуатацию в том же 1994 году и включала две выделенные опорные точки – в ИДСТУ СО РАН и Институте солнечно-земной физики (ИСЗФ) СО РАН. Сеть, на момент своего запуска, имела один внешний канал пропускной способностью 128 Кбит/с с точкой доступа на узле ОАО “Ростелеком”, а выход в глобальную сеть осуществляется по инфраструктуре RBNet. По мере роста требований пользователей Сети, внешний канал связи поэтапно расширялся.

Следующий этап развития ИИВС ИРНОК связан с началом в 2002 году работ по проекту «Создание интегрированной инфор-

мационно-вычислительной сети (с высокоскоростными каналами связи) научных и образовательных учреждений Байкальского региона (Республика Бурятия, Иркутская и Читинская области)» в рамках Федеральной целевой программы (ФЦП) Интеграция науки и высшего образования России на 2002-2006 годы. Головным исполнителем проекта выступает ИДСТУ СО РАН, а соисполнителями – Бурятский научный центр (БНЦ) СО РАН, Бурятский государственный университет, Иркутский государственный технический университет (ИрГТУ) и Читинский государственный технический университет. В рамках исполнения этого проекта, при организационной и финансовой поддержке Целевой программы «Информационно-телекоммуникационные ресурсы СО РАН», в 2002 г. Сеть меняет внешнего оператора. Им становится Институт вычислительных технологий (ИВТ) СО РАН. Канал, пропускной способностью 2 Мбит/с, связавший ИИВС ИРНОК с Сетью передачи данных (СПД) СО РАН, предоставляется Компанией «Транстелеком», выигравшей тендер РАН. При выполнении проекта прокладываются новые ВОЛС между сетями организаций-участников проекта, а также до точек присутствия сети ЗАО Компания «Транстелеком» в городах Иркутск и Улан-Удэ.

В период с 2002 по 2006 гг. ИИВС ИРНОК развивается по следующим направлениям: в ИДСТУ СО РАН сдаются в эксплуатацию узел связи и Суперкомпьютерный центр (СКЦ) ИНЦ СО РАН; открывается Центр управления Сетью, полностью реорганизуется оптический кросс узла, сегменты Сети расширяются до 100 Мбит/с, производится модернизация ядра Сети; расширяются внешние каналы доступа, которые достигают в 2005 году суммарной пропускной способности в 14 Мбит/с.; улучшается управляемость сети, повышается уровень информационной безопасности, появляются новые возможности для мониторинга и анализа информационных потоков; проводится перенумерация Сети, приводящая ее в полное соответствие требованиям СПД

СО РАН; все «академические» почтовые домены переходятся под управление системных администраторов службы поддержки и развития ИИВС.

Построенная по корпоративному принципу, Сеть передачи данных (СПД) СО РАН с примкнувшей к ней в ИИВС, становятся способны обеспечить телефонную, а так же видеоконференц-связь между организациями-участниками. В 2006 году вводится в эксплуатацию первая очередь корпоративной телефонной сети (КТС) ИНЦ, осуществляется ее присоединение к КТС СО РАН (Новосибирск), открываются первые междугородние телефонные направления - Новосибирск, Якутск, Тюмень. С 2007 года эксплуатируется система видеоконференц-связи.

В 2008-2009 годах проводится глубокая модернизация всей кабельной инфраструктуры ИИВС. Выработавшие свой ресурс многомодовые оптические линии связи выводятся из штатной эксплуатации и заменяются на более производительный и удобный одномод. С учетом новых требований значительно изменяется топология сети, на смену традиционной звездообразной структуре приходит более надежная кольцевая. В ядре сети устанавливаются новое магистральное оборудование, уровень доступа поэтапно переводится на пропускную способность в 1Гбит/сек. В 2009 году происходит реструктуризация внешних подключений Сети, существенно расширяются каналы доступа в Интернет и в Сеть передачи данных СО РАН, организуется новый 2М-канал связи до Байкальского музея СО РАН в п. Листвянка (60 км. от Иркутска). В этом же году вводится в эксплуатацию специализированный ресурсный центр хранения данных на базе LSI Logic Engenio 3994.

С начала 2005 года в состав инфраструктуры ИИВС входит суперкомпьютерный центр (СКЦ) коллективного пользования. Развитие материально-технической базы СКЦ осуществляется в рамках программы «Суперкомпьютер СО РАН». Первым вычис-

лительным ресурсом центра являлся кластер МВС-1000 с пиковой производительностью 170 GFlops, созданный при участии ИПМ им. Келдыша РАН и ФГУП «Квант» (Москва). В 2008 г. второй вычислительный кластер – BLACKFORD вошел в список TOP50 самых мощных компьютеров СНГ (сборка ИДСТУ СО РАН – 20 вычислительных узлов платформа Intel S5000PAL, интерконнект GigaEthernet, 40 CPU QC Intel Xeon E5345, 160 процессорных ядер, производительность – 1,493 TFlops (peak), 0,924 TFlops (Linpack)). В 2012 году российской компанией «Т-Платформы» реализован проект создания гибридного вычислительного кластера (рис 1.), названного именем известного ученого, директора-организатора института, академика В.М.Матросова. В качестве вычислительных узлов кластера использована универсальная платформа T-Blade V-Class V205S с процессорами AMD Opteron 6276 «Interlagos». Пиковая производительность системы составляет 33,7 Тфлопс.

Рис. 1. Вычислительный комплекс «Академик В.М. Матросов».

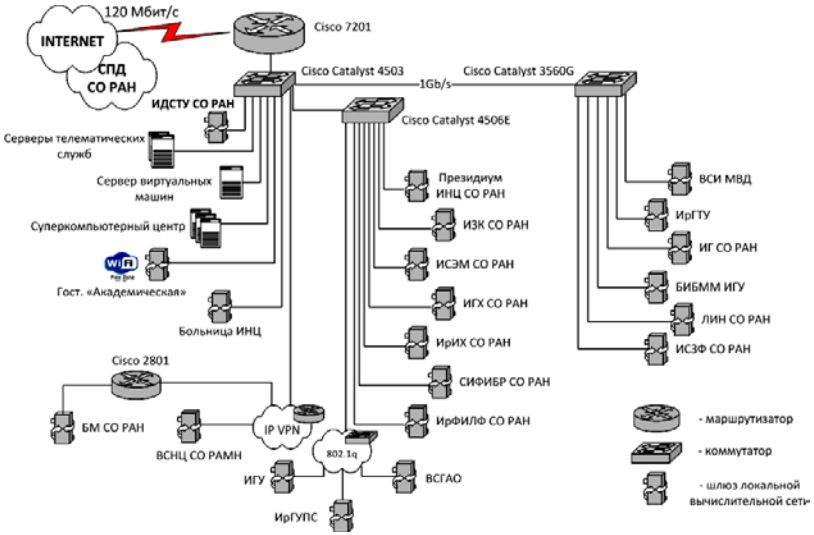


Рис. 2. Схема сети ИИВС.

В 2013 году ресурсная компонента ИИВС расширяется путем включения в нее блэйд-системы HP BLc7000, основным назначением которой становится поддержка «облачных» вычислений на базе VMWare vSphere. Комплекс виртуализации на аппаратном уровне интегрирован в существующую инфраструктуру хранения данных, что позволяет Институтам вести разработку ПО с использованием технологий BigData.

На текущий момент ИИВС объединяет (Рис. 2) высокоскоростными (до 10Gbit/s) каналами локальные вычислительные сети (ЛВС): Президиума ИНЦ СО РАН; Института динамики систем и теории управления СО РАН; Института земной коры СО РАН; Института систем энергетики СО РАН; Института химии СО РАН; Сибирского института физиологии и биохимии растений

СО РАН; Иркутского филиала института лазерной физики СО РАН; Института геохимии СО РАН; Института географии СО РАН; Института солнечно-земной физики СО РАН; Лимнологического института СО РАН; Байкальского музея ИНЦ СО РАН; Восточно-Сибирского института МВД; Иркутского государственного технического университета; Иркутского государственного университета путей сообщения; Восточно-Сибирской государственной академии образования; Иркутского государственного университета; Восточно-Сибирского научного центра СО РАМН, средне- образовательные школы Академгородка, Больница ИНЦ СО РАН – всего 13 научных организаций, 5 крупнейших государственных ВУЗов г. Иркутска, а также социально-значимые объекты Иркутского Академгородка.

ИИВС работает на базе кольцевой оптоволоконной структуры (рис. 3). Основными технологиями передачи данных в сети на канальном уровне являются GigabitEthernet и 10G-Ethernet. Некоторые участники сети, ввиду своей удаленности от опорных узлов подключены через высокоскоростные каналы связи, арендуемые у коммерческих провайдеров. Подключение таких организаций осуществляется с использованием технологий VLAN или IP VPN, что позволяет терминировать их транспортные сети непосредственно на внутреннем маршрутизаторе ИИВС не нарушая логической схемы Сети. Использование в ИИВС кольцевых топологий с применением протокола RSTP обеспечивает устойчивую работу сети. ИИВС, являясь частью корпоративной сети передачи данных СО РАН, обеспечивает абонентам полную связность со всеми российскими и зарубежными глобальными сетями. В настоящий момент пропускная способность внешних каналов связи ИИВС составляет 120 Mbit/s. В целях осуществления гибкой динамической маршрутизации по протоколу BGP сети ИРНОК объединены в автономную систему с номером AS8506. Магистральное оборудование сети целиком представлено про-

дукцией компании Cisco Systems. На опорных коммутаторах сети функционируют системы защиты от подделки mac-адресов, arp-спуфинга, включения в сеть посторонних пользователей, системы контроля за всеми видами пакетных штормов, что позволяет избежать излишней нагрузки на маршрутизатор в случае реализации этих видов атак. В сети активно используется стандарт IEEE 802.1q, позволяя строить под некоторые виды задач виртуальные сети достаточно сложной конфигурации.

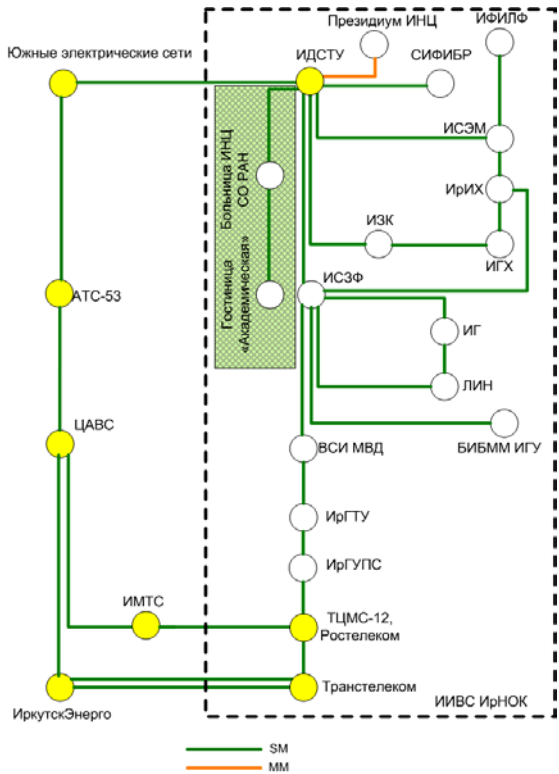


Рис 3. Оптоволоконная инфраструктура ИИВС.

Для обеспечения функционирования центрального узла связи ИИВС ИРНОК и Суперкомпьютерного центра создана инженерная инфраструктура, обеспечивающая необходимый климатический режим и резервное электропитание от аккумуляторных батарей ИБП.

Для надежного и доступного хранения данных большого объема в ИИВС используется централизованная система хранения данных (СХД) на базе оборудования LSI Logic Engenio 3994 класса

SAN (SAN – Storage Area Network(сеть хранения данных)) объемом 64 Тб. В 2015 году планируется дальнейшее расширение ресурса единой сети хранения данных путем включения в нее оборудования EMC VNX5200 стартовой емкостью 128Тб. Наличие собственной общей оптоволоконной сети связи между пользователями ИИВС позволяет использовать дисковые ресурсы СХД как с применением прикладных протоколов SMB, FTP, iSCSI и FCoE, так и аппаратно – путем оптоволоконного подключения дисковых адаптеров серверов Институтов к SAN-коммутаторам сети хранения данных. Особая роль СХД отводится в работах по созданию на базе ИДСТУ СО РАН Ресурсного центра для хранения данных дистанционного зондирования Земли высокого разрешения. Существенной функцией СХД также является ее использование для функционирования дисков виртульных машин в рамках концепции единого Центра обработки данных.

Одна из последних модернизаций ИИВС – развертывание единого масштабируемого программно-аппаратного комплекса, обладающего высокой производительностью и отказоустойчивостью за счет использования «облачных» технологий, позволяющих равномерно распределять нагрузки на физическую инфраструктуру (вычислительные, дисковые и сетевые ресурсы). Про-

граммно-аппаратный комплекс имеет в своем составе два специализированных SAN-коммутатора, позволяющих объединять системы хранения данных от различных производителей (дисковые ресурсы) в общую Сеть хранения данных с использованием высокоскоростных (на текущий момент – 8Gbit/s) оптоволоконных каналов связи.

На основе нового программно-аппаратного комплекса HP BLc7000 с применением решений платформы виртуализации VMWare vCenter развернуто облако IaaS (рис.4).

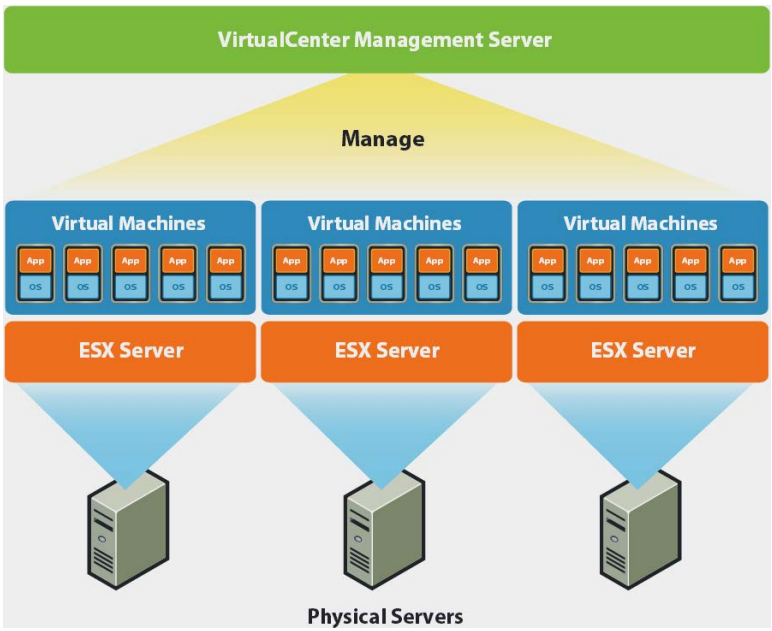


Рис. 4. Структура облака IaaS.

В облако перенесены ряд ключевых сервисов, обеспечивающих работу компонент ИИВС таких как:

- сервис IP-телефонии;
- сервис корпоративного LDAP-каталога, разрабатываемого в рамках Интеграционного проекта СО РАН РАН №73;
- Сервис электронных библиотек на базе платформы IRBIS64;
- Региональный узел системы ZooSPACE;
- Контроллеры WiFi-зон ИНЦ СО РАН и Гостиницы «Академическая»;
- Изолированный гостевой шлюз в Интернет (DNS, NAT, СУБД, контроль доступа) для пользователей «свободных» сетевых WiFi- и проводных Интернет-зон ИНЦ СО РАН и Гостиницы «Академическая»;
- Сервисы каталогов ActiveDirectory для обслуживания регионального домена в рамках проекта «Корпоративное облако СО РАН», а также для нужд VMWare vCenter Server;
- WEB-серверы, обслуживающие сайты организаций, входящих в ИИВС, а также специализированные сервисы, такие как, например, прямые видеотрансляции со стационара на о. Долгий (оз. Байкал) ;

В последние годы ведется активная работа с использованием системы видеоконференц-связи. Проводятся научные семинары с участием представителей нескольких городов, совместные заседания, обсуждения проектов, организуются дистанционные чтения лекций, виртуальные экскурсии по Байкальскому музею, прямые трансляции выступлений с конференций в сеть Интернет. На площадях ИДСТУ СО РАН функционирует специализированный видеоконференц-зал (рис. 5).

Рис. 5. Малый видеоконференц-зал.

Всё это в комплексе формирует Информационно-телекоммуникационную инфраструктуру Иркутского научно-образовательного комплекса, которая успешно выполняет свои основные функции:

- информационно-вычислительного и телекоммуникационного обеспечения проведения междисциплинарных фундаментальных и прикладных научных исследований, в том числе по перспективам развития территорий Иркутской области, а также поддержки образовательной деятельности;
- выход в глобальные сети и корпоративные научно-образовательные сети;
- концентрации, с целью коллективного использования и развития, дорогостоящих телекоммуникационных, информационных, вычислительных, дисковых и других ресурсов

научных и образовательных учреждений Иркутской области;

- эффективной эксплуатации и развития корпоративных систем хранения, обработки, данных и доступа к информационно-вычислительным и мультимедийным научно-образовательным ресурсам учреждений, подведомственных ФАНО и государственных ВУЗов города Иркутска;
- поддержки баз научных тематических данных и знаний, необходимых для подготовки, анализа и принятия решений в задачах социально-эколого-экономического развития территорий.
- поддержки удаленного доступа к открытым научно-образовательным ресурсам для населения и бизнеса, а также органов государственной власти и местного самоуправления Иркутской области;
- поддержки выполнения и внедрения региональных научно-исследовательских и опытно-конструкторских проектов, заданных органами государственной власти и местного самоуправления Иркутской области по перспективному развитию территорий области;
- внедрения функций «электронной библиотеки»: обеспечение возможности коллективного использования электронной литературы, реферативных журналов и т.д.

В последние годы в Институтах Иркутского научного центра наблюдается существенный рост объемов первичных научных данных, включающих видео-информацию, «тяжелые» графические форматы и многомерные числовые массивы. Очевидна необходимость одновременного хранения не только первичного материала, но и результатов его обработки в сочетании с высокими требованиями по надежности, доступности и скорости доступа к данным. Несмотря на большой объём, информационные ре-

сурсы ИРНОК из-за их локализации в институтах не всегда могут использоваться для проведения междисциплинарных фундаментальных и прикладных научных исследований, а также в учебном процессе. Это обуславливает необходимость дальнейшего развития инфраструктуры.

В настоящий момент в рамках ИИВС сконцентрировано большое количество сетевого, вычислительного, климатического и других видов оборудования, совместно используемого научно-образовательным сообществом. Современное состояние и существенная значимость построенной инфраструктуры требуют дальнейшего ее развития в рамках концепции Центра обработки данных (ЦОД), реализующего новый виток развития созданной ИДСТУ СО РАН ИИВС, являющейся основной связующей инфраструктурной компонентой научно-исследовательских организаций региона.

С 2012 года начата работа по созданию на базе ИИВС Центра обработки данных, выполненного в соответствии с мировыми стандартами. Осуществлено комплексное проектирование всех необходимых инженерных систем, включающих:

- Систему размещения оборудования;
- Систему электроснабжения (бесперебойного и гарантированного);
- Климатическую систему (кондиционирование, вентиляция);
- Структурированную кабельную систему;
- Комплексную систему безопасности;
- Комплекс противопожарных систем;
- Автоматизированную систему мониторинга и диспетчеризации.

К настоящему моменту проведена строительная подготовка помещения машинного зала, ведется реконструкция вводно-

распределительной части энергосистемы здания, включающей автономную резервную дизель-генераторную установку Caterpillar Olympian GEP-500, мощностью 500 кВА.

Несмотря на ориентированность ЦОД, в первую очередь, на решение научных и научно-организационных задач, вследствие применения надежных, универсальных, масштабируемых решений его возможности могут быть использованы в любой сфере, где так или иначе используются информационные системы, большие объемы данных, мультимедийные технологии, высокопроизводительные вычисления.

В результате проведенных работ создан серьезный научно-технический задел для создания ведомственного Центра обработки данных и включения его в национальную инфраструктуру сектора исследований и разработок, обеспечивающей проведение многопрофильных, многометодных и междисциплинарных исследований.

ТЕХНОЛОГИИ СОЗДАНИЯ ИНТЕГРИРОВАННЫХ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИХ СИСТЕМ В НАУЧНЫХ ПРОЕКТАХ

Гаченко А.С., Ружников Г.М., Хмельнов А.Е.
Институт динамики систем и теории управления СО РАН
(ИДСТУ СО РАН)
gachenko@icc.ru

Статья посвящена актуальным проблемам разработки информационно-аналитических систем в научной сфере.
Ключевые слова: инфраструктура пространственных данных, базовые пространственные данные, каталоги метаданных, реестры, органы государственной власти, картографические данные.

The article is devoted to current problems of regional component of a spatial data infrastructure development.
Keywords: infrastructure of spatial data, basic spatial data, catalogs of metadata, registers, public authorities, cartographical data.

Цифровые геопространственные данные активно используются в научных исследованиях и решении большого перечня задач управления территориальным развитием.

Специфика формирования геоинформационных ресурсов увеличила актуальность создания проблемно-ориентированных программных средств, интегрирующих универсальные сетевые технологии с ГИС-технологиями, поддерживающих организацию и работу с пространственно-распределенными данными, т.е. создающих интерактивную среду взаимодействия клиентских приложений с ГИС-сервером.

На начальном этапе развития картографические Web-сервера обеспечивали лишь выбор и просмотр заданного набора картинок в форматах GIF, JPEG или другом графическом формате. Первым среди подобных серверов считается National Atlas Information Service (NAIS) of Canada. Следующим этапом развития систем для Web-просмотра картографической информации было подключение функций СУБД для взаимодействия с базами данных (БД), содержащими атрибутивные данные карт. Данный подход предъявил более высокие требования к программно-аппаратному обеспечению сервера, однако, при этом повысилась эффективность его работы за счет структурированного представления картографических данных.

В институтах Иркутского научного центра (ИНЦ) СО РАН научные исследования базируются на уникальных проблемно и предметно-ориентированных базах пространственных данных по ландшафтам и геосистемам, картографированию природы, хозяйства и населения Сибири (Институт географии СО РАН), геологической среде и сейсмическим процессам (Институт земной коры СО РАН), геохимии окружающей среды и осадочных бассейнов (Институт геохимии СО РАН), электроэнергетическим и трубопроводным систем (Институт систем энергетики СО РАН), биоразнообразию фауны и флоры оз. Байкал (Лимнологический институт СО РАН), физиологии растений, молекулярной биологии и экологии растительных организмов (Сибирский институт физиологии и биохимии растений СО РАН), дистанционному зондированию поверхности Земли (Институт солнечно-земной физики СО РАН) и т.д.

В настоящее время актуальны работы по разработке и созданию:

– интеллектных методов и инструментальных средств создания, анализа интегрированных распределённых информационно-аналитических и вычислительных систем с применением ГИС, GRID и Web-технологий;

– единой интегрированной инфраструктуры проблемно и предметно-ориентированных тематических баз пространственных данных институтов ИНЦ СО РАН;

– современных методов и технологий интеграции разноформатных междисциплинарных данных и результатов исследований, базирующихся на пространственных характеристиках и признаках;

– централизованного хранилища цифровой топоосновы и картографической информации с удаленным Web-доступом пользователей;

– системы сервисов геоданных СО РАН;

– новых методов и технологий исследования и обработки пространственных данных, включая разработку логических методов и методов обработки больших массивов данных;

– ГИС-портала институтов ИНЦ СО РАН;

– методов и средств планирования распределенного решения информационно-вычислительных задач;

– специальных баз проблемно-ориентированных и предметно-ориентированных географических данных и знаний для их размещения в Интернет;

– новых моделей и методов, базирующихся на результатах натурных наблюдений и на эмпирических данных, включая создание методов и технологий обработки данных дистанционного зондирования;

– новых методов и технологий анализа и обработки географических данных и знаний в интегрированных системах геоинформационного картографирования.

В рамках комплексного проекта информатизации науки и образования в ИНЦ СО РАН создана Интегрированная информационно-вычислительная сеть Иркутского научно-образовательного комплекса (ИИВС ИрНОК) с пропускной способностью магистрали до 1Gb/s, а также региональный узел доступа к Сети передачи данных (СПД) СО РАН.

В 2012 году российской компанией «Т-Платформы» на базе ИДСТУ СО РАН реализован проект создания «под ключ» гибридного вычислительного кластера, названного именем известного ученого, директора-организатора института, академика В.М. Матросова. В качестве вычислительных узлов кластера использована универсальная платформа T-Blade V-Class V205S с процессорами AMD Opteron 6276 «Interlagos». Пиковая производительность системы составляет 33,7 Тфлопс. В 16-й редакции (от 27.03.2012) списка Топ-50 вычислительный кластер «Академик В.М. Матросов» занял 26-е место.

Суперкомпьютерный центр коллективного пользования, поддерживающий информационно-вычислительное обеспечение фундаментальных и прикладных исследований, проводимых в институтах ИНЦ СО РАН и вузах Байкальского региона.

Таким образом, в ИНЦ СО РАН создана информационно-вычислительная, телекоммуникационная инфраструктура и накоплены уникальные научные проблемно и предметно-ориентированные геопространственные данные, что служит основой формирования Центра поддержки междисциплинарных научных исследований институтов ИНЦ СО РАН.

В целом, отмечается существенный рост потребностей со стороны науки и образования в доступных информационно-вычислительных ресурсах для проведения фундаментальных и прикладных исследований. Это приводит к необходимости интеграции (в том числе, с использованием GRID-технологий) про-

странственно распределенных хранилищ разнородных данных и вычислительных мощностей, организации коллективного доступа к ним.

В связи с этим большой интерес представляют технологии, позволяющие автоматизировать процесс создания комплексных тематических информационных систем.

В настоящее время в Иркутском научном образовательном комплексе (ИРНОК) проводятся работы по формированию и ведению баз данных поддержки междисциплинарных научных исследований Байкальской природной территории на основе материалов, которые собраны в институтах Иркутском научном центре (ИНЦ) СО РАН.

Разработан и наполняется тематическими и пространственными данными **геопортал** по биоразнообразию Байкальской природной территории.

Основной целью создания геопортала является обеспечение общей платформы для совместной работы с распределенными сервисами геообработки и пространственными данными. Эта платформа реализует систему управления доступа к географическим объектам на карте. Разработаны специализированные сервисы, обеспечивающий следующие функции обработки пространственно-распределенных данных:

- хранение геоданных;
- публикация картографических и реляционных данных;
- поиск ГИС-ресурсов по каталогу геопортала.

Наряду с данными сервисами связанными с геообработкой баз данных существует необходимость в хранении и обмене текстовыми документами (заявки, отчеты, интеграционные проекты).

В Сибирском отделении РАН в рамках развития этого направления разрабатывается **корпоративный каталог**. В его задачи входит:

- каталогизация информационных ресурсов и сетевых сервисов;
- обеспечение единой инфраструктуры аутентификации и авторизации пользователей информационных систем СО РАН и ДВО РАН;
- обеспечение глобальных политик доступа к информационным ресурсам;
- мониторинг доступности информационных ресурсов и сетевых сервисов;
- создание автоматически актуализируемой распределенной справочной системы СО РАН и ДВО РАН;
- обеспечение работы совместных и частных распределенных информационных систем.

Корпоративный каталог разрабатывается как централизованно-распределенная информационная система с разграничением зон ответственности между центрами и организациями отделений РАН. Корпоративный каталог создается на основе семейства LDAP-серверов (LightweightDirectoryAccessProtocol).

Разрабатываются компоненты инфраструктуры взаимодействия на уровне серверов обмена данными и метаданными. Ведется работа по взаимодействию между серверами на основе LDAP – «облегчённый протокол доступа к каталогам» протокол прикладного уровня для доступа к службе каталогов. LDAP — это относительно простой протокол, использующий TCP/IP и позволяющий производить операции авторизации (bind), поиска (search) и сравнения (compare), а также операции добавления, изменения или удаления записей. Обычно LDAP-сервер принимает входящие соединения на порт 389 по протоколам TCP или UDP. Для LDAP-сеансов, инкапсулированных в SSL (англ. (SSL)SecureSocketsLayer — уровень защищённых сокетов)— криптографический протокол, который обеспечивает установле-

ние безопасного соединения между клиентом и сервером. На основании протокола SSL 3.0 был разработан и принятстандарт RFC, получивший имя TLS. Протокол обеспечивает конфиденциальность обмена данными между клиентом и сервером, использующим TCP/IP.

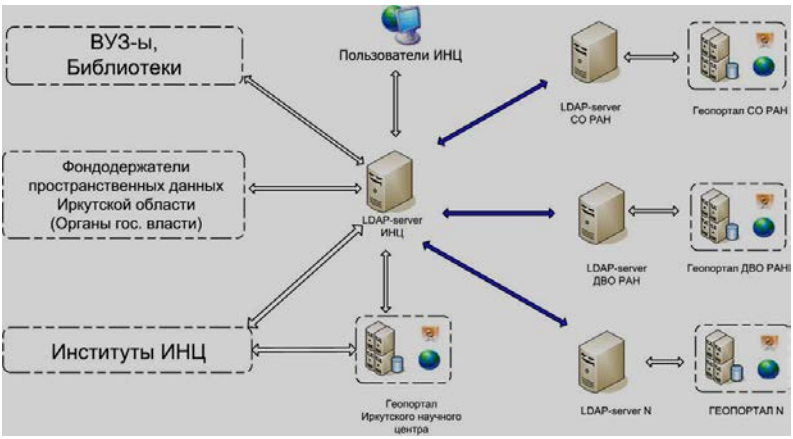


Рис. 1. Структура взаимодействия LDAP-серверов с банками данных фондодержателей Иркутской области.

Каталоги такого рода, как правило, содержат статические и редко изменяемые элементы, так как каталоги изначально оптимизированы для очень быстрого отклика на запросы поиска и чтения данных.

Такие каталоги полностью структурированы. Каждый элемент данных имеет имя, которое одновременно определяет положение элемента в иерархии каталога. Каждый атрибут элемента, как правило, может иметь несколько значений и это является нормальным поведением, в отличие от обычных баз данных.

Каталоги являются очень специфическими системами хранения данных. Их удобно использовать для иерархически скомпонованных объектов. Каталоги могут быть реплицированы между несколькими серверами, для организации удобного доступа и распределения нагрузки. Текстовая информация хорошо подходит для каталогов, так как легко поддается поиску, но данные могут быть представлены и в любой другой форме.

Основная цель разрабатываемого сервиса направлена на взаимодействие между каталогами Дальнего Востока, Иркутска и Новосибирска. Корпоративный каталог отделений РАН создается для поддержки различных сетевых сервисов и приложений.

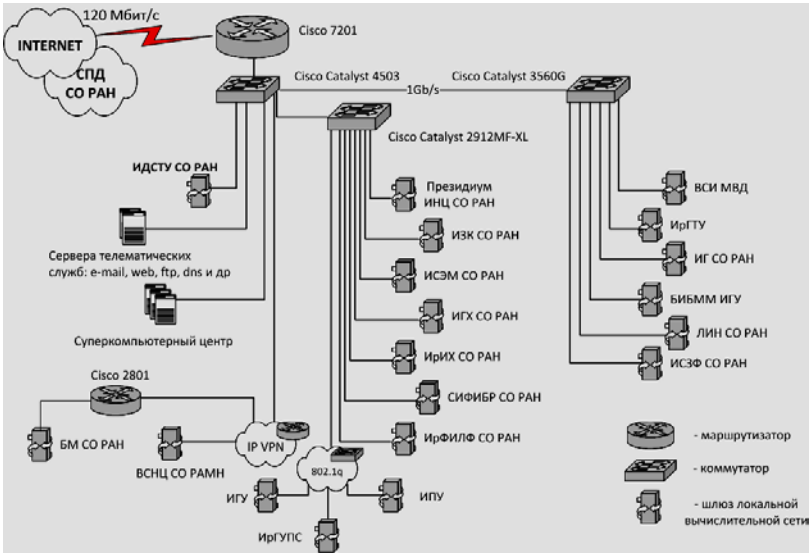


Рис. 2. ИИВС ИРНОК.

Разработаны компоненты инфраструктуры взаимодействия на уровне серверов данных и метаданных. Проведена работа по отработке взаимодействия между серверами на основе протоколов LDAP и Z39.50.

ДСТУ СО РАН на основе программно-аппаратной платформы, включающей систему хранения данных (СХД) SAN ReadyStorage 3994 развернут периферийный LDAP –сервер на специально выделенной для этих задач виртуальной машине. На сервере виртуальных машин выделены необходимые ресурсы под задачи связанные с процессами репликации данных между LDAP-серверами СО РАН и ДВО РАН. Существует центральный сервер каталогов, хранящий сводную информацию. Информация на центральном сервере изменяется в результате непосредственного редактирования, а также в результате репликаций с периферийных серверов. Каждый периферийный сервер может хранить все дерево корпоративного каталога или его часть. Основной узел каталога РАН находится на сервере http://elib.sbras.ru:8080/jspui/ и доступен по протоколу LDAP порт 1389 (LDAPS порт 1636). На LDAP-сервере ИДСТУ СО РАН развернут узел ZooSPACE разрабатываемый в ИВТ СО РАН.

LDAP-сервер функционирует в Интегрированной информационно-вычислительной сети (ИИВС) Иркутского регионального научно-образовательного центра, объединяющей в своей инфраструктуре локально-вычислительные сети институтов Иркутского научного центра, государственных высших учебных заведений г. Иркутска, а также учреждений ВСНЦ РАМН. ИИВС ИРНОК работает на базе высокоскоростной оптоволоконной инфраструктуры. Основной технологией передачи данных в сети на канальном уровне является GigabitEthernet. Некоторые участники сети ввиду своей удаленности от опорных узлов подключены через высокоскоростные каналы связи, арендуемые у коммерческих провайде-

ров. Связность с российскими и зарубежными глобальными сетями обеспечивается внешним каналом связи, пропускной способностью 120 Мбит/сек.

Данная телекоммуникационная инфраструктура позволяет оперативно подключать новые научные организации для эффективного информационного обмена при проведении научных исследований.

Например, в 2013 году проведены работы по установке библиотечной каталожной распределенной системы IRBIS64 на выделенный для этой цели сервер, который обеспечивает поддержку библиотечных ресурсов ИНЦ.

Сервер IRBIS64 предназначен для осуществления доступа пользователей Интернет к электронным каталогам собственных библиотек и сторонним библиографическим базам данных распределенной системы автоматизации библиотек IRBIS. Система IRBIS представляет собой типовое интегрированное решение в области автоматизации библиотечных технологий и предназначена для использования в библиотеках любого типа и профиля для использования в качестве одного из основных компонентов библиотечных Интернет-серверов и Интернет-комплексов. Система полностью отвечает международным требованиям, предъявляемым к таким системам, и поддерживает все отечественные библиографические стандарты и форматы.

Базовые операции IRBIS64:

Поиск в произвольной базе данных, имеющей структуру IRBIS64 по неограниченному числу полей, по любым элементам описания и их комбинаций, с применением логики «И», «ИЛИ» и «ФРАЗА ЦЕЛИКОМ», с возможностями определения префиксов и квалификаторов поисковых терминов, грамматической нормализации слов русского языка и применения аппарата усечений.

Уточняющий поиск в результатах предыдущего поиска по условию (последовательный поиск). Сортировка результатов поиска по условиям.

Использование при поиске статических словарей и рубрикаторов, включенных в поисковые формы, с возможностью комбинирования элементов словарей с любыми другими поисковыми предписаниями.

Использование динамических словарей баз данных, с возможностью получения списка терминов словаря и последующего поиска по выбранным терминам; навигация по словарям, включая задание начала сканирования словарей по первым символам, а также в терминах «следующие»-«предыдущие».

Показ найденных записей в стандартных форматах, включая информационный и в виде каталожной карточки.

Данные загружены в региональный информационный узел корпоративного каталога, количество записей в библиотечном каталоге составляет порядка 80 000.

Заключение

В статье изложен подход для организации инфраструктуры и создания интегрированных информационно-аналитических систем, а также описаны области применения систем подобного рода. Изложены оригинальные подходы и методы по их созданию.

СОДЕРЖАНИЕ II ТОМА