РАСПРЕДЕЛЕННАЯ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКАЯ СРЕДА ДЛЯ ИССЛЕДОВАНИЙ ЭКОЛОГИЧЕСКИХ СИСТЕМ*

А. М. Федотов, В. Б. Барахнин, А. Е. Гуськов, Ю. И. Молородов
Институт вычислительных технологий СО РАН, Новосибирск, Россия
e-mail: fedotov@ict.nsc.ru, bar@ict.nsc.ru,
arcon@ict.nsc.ru, yumo@ict.nsc.ru

Some aspects for establishment of a virtual environment to share scientific results with other researchers are considered. This environment will allow a final user to work in thematic portal, functioning as a data management system that supports environmental studies in Siberia.

Введение

Информационные технологии оказывают огромное влияние на все области человеческой деятельности, связанные с накоплением и обработкой информации. За относительно небольшое время существования информационно-коммуникационных технологий накоплен очень большой объем разнообразных данных, представленных исключительно в электронной форме. Возникают задачи обеспечения доступа (в том числе и удаленного) пользователей к разнородным информационным ресурсам, защиты авторских прав на документы, систематизации большого объема разнородных типов документов.

Интеграция информационных ресурсов в единую информационную среду и организация доступа к вычислительным ресурсам — это одни из важнейших направлений развития современных информационных технологий. Решение проблем создания и интеграции информационных ресурсов и продуктов должно стать необходимым условием развития многих стран, в том числе и России. Стремительное развитие глобальных информационных и вычислительных сетей ведет к изменению фундаментальных парадигм обработки данных, которое можно охарактеризовать как переход к поддержке и развитию распределенных информационно-вычислительных ресурсов [1]. Технологии использования распределенных информационно-вычислительных ресурсов получают все больший приоритет в информационном обществе. При этом наблюдаются переход к исключительно распреде-

*Работа выполнена при частичной финансовой поддержке Российского фонда фундаментальных исследований (гранты № 06-07-89060, № 06-07-89038), президентской программы "Ведущие научные школы РФ" (грант № НШ-9886.2006.9) и интеграционных проектов СО РАН.

ленной схеме создания, поддержания, хранения ресурсов[1] и стремление к виртуальному единству посредством предоставления свободного доступа к любым ресурсам сети через ограниченное число точек доступа. Постулируется принцип формирования в ресурсах сети единого, математически однородного поля компьютерной информации, которое способно стать универсальным и машинонезависимым носителем данных, унифицированных программ и глобально распределенных вычислительных процессов [2].

Необходимо разработать механизмы, обеспечивающие как функционирование общей информационно-аналитической рабочей среды, так и доступ к научным ресурсам и их сохранность, чтобы решать задачи информационной поддержки научных исследований. Эти вопросы приобретают особую важность в области изучения экологических систем, когда различные группы исследователей (вследствие особенностей решаемых проблем, а также природы вопросов, связанных с окружающей средой), разделенные географически, должны осуществлять совместную работу, обмен данными и знаниями и координировать свои действия с целью оптимизации использования информационно-вычислительных ресурсов, сервисов и приложений.

Тесное кооперирование информационных технологий и наук о Земле способствует пониманию как глобальных, так и региональных природных процессов, формирующих природную среду. Здесь информационные технологии играют определяющую роль в разработке базовой инфраструктуры исследований, позволяя ученым сконцентрироваться на их задачах, и обеспечивают среду публикаций научных результатов в Интернет. Следует отметить, что в настоящее время сбор данных об окружающей среде и корректное использование полученных таким образом наборов данных приобретают все большее значение. Данные об окружающей среде часто незаменимы; они всегда уникальны (хотя бы, например, по времени их получения). Их сбор часто обходится очень дорого. По этим причинам огромную важность имеет задача извлечения максимальной пользы от данных, полученных в результате каждого исследования.

Эти задачи и легли в основу разработанного в СО РАН проекта по созданию интегрированной системы предоставления фактографической информации и предоставлению в объединенное пользование вычислительного и поддерживающего оборудования и программного обеспечения, необходимого для эффективного доступа к данным — виртуальной среды работы с ресурсами. Цель проекта — выбор и отработка технологий, предназначенных для интеграции распределенных баз данных наблюдений, мониторинга, анализа и моделирования состояния экосистем, чтобы изучать и прогнозировать природные, социальные и экономические последствия, вызванные как естественными, так и антропогенными изменениями в экосистемах, и разрабатывать рекомендации и подходы к оптимальному управлению функционированием экосистем.

В рамках проекта предполагается создание пилотной модели виртуальной среды (распределенной базы данных и метаданных). Она может быть использована как основа для создания распределенной информационной виртуальной среды с целью обмена данными (наукоемкими ресурсами) для научных исследований с единой точкой доступа (порталом). При этом информационный портал будет выступать как:

- центр аккумуляции информационных ресурсов, обеспечивающий их централизованное хранение и оперативную обработку;
- центр регистрации распределенных информационных ресурсов;

[1]Эффективная эксплуатация информационных ресурсов возможна только в том случае, когда они постоянно поддерживаются авторами, т. е. на основе технологий использования распределенных информационно-вычислительных ресурсов, которые получили название GRID-технологий.

- точка доступа к распределенным информационным ресурсам;
- центр администрирования информационных ресурсов;
- центр сбора и обработки статистики использования информационных ресурсов.

1. Модель виртуальной среды

Поиск является наиболее востребованной функцией в информационном обществе. В настоящее время разработано большое число всевозможных инструментов, реализующих поиск в разнообразных условиях и с различными критериями. Отметим, однако, что фундаментальная с точки зрения поиска проблема в распределенных системах — отсутствие стандартных механизмов классификации, каталогизации и систематизации ресурсов не позволяет осуществлять поиск достаточно эффективно [3]. Один из возможных вариантов — создание корпоративной сети, ресурсы которой удовлетворяют определенным правилам, допускающим их автоматизированную аналитическую обработку.

Другая проблема — это использование ресурса. Идеальной является ситуация, когда формат содержимого ресурса совпадает с форматом, который использует исследователь, — например, когда формат задается стандартом представления предметных данных определенной направленности. Такими стандартами, например, являются: SEGY — геофизические данные, FITS — астрономические данные, SDF — химические данные. Однако если формат ресурса не согласован с требуемым форматом, исследователю придется создавать или искать средства преобразования форматов. В общем же случае решить эту задачу пока не представляется возможным. Тем не менее одна из целей работы — предложить механизм конвертирования (преобразования) ресурсов из одного формата в другой, причем достаточно общий, чтобы быть применимым к определенному классу ресурсов (например, для экологических систем).

Следующий логический шаг — автоматизация функции получения данных из внешних источников. Несмотря на сервисный характер, эта функция содержит несколько компонентов, требующих отдельной технологической проработки. Так, необходимо определить механизмы обнаружения фактов обновления или появления новых ресурсов, определить процедуру получения ресурсов с учетом требуемого формата.

Заметим, что предлагаемые решения поставленной задачи жестко ориентированы на конкретную предметную область и соответствующие ей модели и схемы данных [4]. Другой класс альтернативных решений имеет характерного представителя — Microsoft SharePoint. Данный продукт предоставляет средства обмена документами и ориентирован на корпоративный электронный документооборот. Здесь основной единицей манипуляции является файл, причем его содержимое системой никак не анализируется. SharePoint имеет средства поддержки Web Services — технологии обмена слабоструктурированными данными в формате XML, но не предоставляет дополнительных средств, позволяющих реализовать полноценный автоматизированный обмен разнородными данными.

Прежде чем перейти к постановке задачи, необходимо уточнить объект исследования. К результатам научных исследований в первую очередь относятся:

— *данные*, полученные в результате научных исследований, физических или натурных экспериментов и представленные в виде электронных коллекций, документов[2], изображений, многомерных массивов и т. д;

[2]К этому типу данных следует отнести и различные научные публикации как в печатном, так и в электронном виде, а также персональные данные об участниках корпоративной работы.

— *модели*, описывающие принципиальные компоненты, специфику и ограничения предметной области с необходимой степенью детализации. Модели являются своего рода "каркасом" предметной области, которому соответствуют данные и в рамках которого функционируют алгоритмы;

— *алгоритмы*, разработанные для решения определенного класса задач в соответствии с определенной моделью, т. е. входом и выходом алгоритма являются данные, соответствующие этой модели, а его функционирование происходит в рамках ограничений модели с учетом ее специфики.

Первоочередным объектом на текущий момент являются *данные* по причине больших возможностей для их формализации и аналитической обработки. Данные могут быть представлены в виде электронных коллекций, изображений, многомерных массивов и пр., в дальнейшем целостная и самодостаточная единица данных, обладающая уникальным идентификатором, будет называться ресурсом. Выделяются следующие базовые категории данных (ресурсов):

• *табличные данные* — ресурсы, описывающие многомерные массивы однородных элементов;

• *бинарные данные* — ресурсы, содержание которых представляет собой двоичный код и для просмотра которых используется специализированное программное обеспечение (например, изображения, звуко- и видеозаписи);

• *слабоструктурированные данные* — ресурсы, содержание которых представляет собой упорядоченную последовательность элементов (структуру) с априори заданной семантикой (форматом). Заметим, что табличным и бинарным данным может сопутствовать структурное описание, определяющее их специфику (метаданные).

Каждая категория расслаивается на типы ресурсов, причем каждый тип может содержать несколько подтипов ресурсов. Например, категория "Бинарные данные" может содержать тип "Изображения", который, в свою очередь, содержит подтип "Фотографии".

Отметим, что между предложенными базовыми категориями ресурсов нет четких границ. Так, бинарные ресурсы в ряде случаев можно рассматривать как табличные данные (например, растровые изображения) и как слабоструктурированные данные (например, векторные изображения). Кроме того, любые ресурсы должны сопровождаться дополнительными сведениями для их систематизации и классификации (метаданными), которые следует относить к слабоструктурированным данным. Согласно этому и другим соображениям первоочередной интерес для исследования представляют именно слабоструктурированные данные.

2. Постановка задачи

Приведем необходимые определения для понимания предмета задачи [5].

Наукоемкий ресурс является идентифицируемой электронной сущностью и содержит опубликованные данные, имеющие научную ценность. Каждый наукоемкий ресурс принадлежит одному из источников.

Источник представляет собой внешнюю информационную систему (базу данных, каталог, хранилище и т. п.), которая содержит наукоемкие ресурсы.

Объекты виртуальной среды являются центральной сущностью виртуальной среды. Объекты содержат сведения о сущностях предметной области. Каждому объекту может

соответствовать несколько ресурсов из различных источников[3]. Поскольку ресурсы из разных источников могут содержать противоречивые данные об одном объекте, ресурс следует считать опубликованной *версией* объекта, а несколько версий одного объекта — *смежными ресурсами*. Объект должен относиться к определенной категории.

Категории объектов виртуальной среды используются для определения формата соответствующих ресурсов, методов их обработки и отображения. Существуют три базовых категории, каждая из которых может содержать несколько подкатегорий: элементы коллекции (документы), массивы данных, медиа-объекты.

Каталоги содержат все объекты и ресурсы, доступные в виртуальной среде.

Модель виртуальной среды для обмена результатами научных исследований должна обеспечивать следующую ***функциональность***.

Публикация наукоемких ресурсов пользователями должна включать процедуры регистрации, именования, аннотирования и определения формата [2, 6]. Аннотирование состоит из указания описательных метаданных ресурса для целей его систематизации и каталогизации. Определение формата заключается в указании способа извлечения содержимого ресурса для его последующей аналитической обработки или изменения формата. Отметим, что все перечисленные действия целесообразно выполнять не для каждого ресурса в отдельности, а группой для всех ресурсов одной категории одного источника.

Аналитическая обработка ресурсов должна включать автоматизированные функции аннотирования, определения смежных ресурсов и определения релевантных объектов. Автоматизированная функция аннотирования осуществляет выборку метаданных ресурса и записывает их в соответствующий каталог виртуальной среды. Функция определения смежных ресурсов согласно определенным критериям осуществляет поиск среди ресурсов тех, которые соответствуют идентичным объектам. Функция определения релевантных объектов осуществляет поиск объектов, которые логически связаны друг с другом, например, персона является автором публикации или сотрудником организации.

Доступ к опубликованным ресурсам должен включать в себя функции отображения каталогов объектов и связанных с ними ресурсов, поиск объектов и ресурсов по каталогу. Особый интерес представляет функция конвертирования ресурсов, которая используется при запросе пользователем ресурса в указанном формате.

Для *автоматизированного функционирования среды* необходимы функция мониторинга ресурсов и актуализации их метаописаний, функция уведомления пользователей о появлении новых ресурсов и обновлении существующих, функция диспетчеризации. Функция мониторинга выполняет проверку появления новых ресурсов в источнике, а также проверку состояния ресурса, в частности, определяет его доступность и факт обновления. Если обновление ресурса имело место, то функция актуализации модифицирует сведения о ресурсе, хранимые в виртуальной среде. При этом некоторым пользователям может быть отослано уведомление о появлении новых ресурсов или обновлении существующих. Взаимодействие этих функций, а также отправку новых или обновленных ресурсов заинтересованным пользователям обеспечивает функция диспетчеризации.

Виртуальная среда должна обладать следующими характеристиками:

— *распределенность ресурсов.* Ресурсы могут располагаться на географически удаленных серверах;

[3]Наиболее характерным объектом, которому могут соответствовать несколько ресурсов из различных источников, являются персональные данные. Персональные данные — неотъемлемая коллекция виртуальной среды, они необходимы как минимум для регламентации прав и разграничения доступа к ресурсам.

— *программная разнородность ресурсов.* Ресурсы имеют различную программную природу, т. е. они могут храниться под управлением различных СУБД и формироваться с помощью разных алгоритмов и их реализаций на различных языках программирования;

— *несогласованные схемы данных и форматы ресурсов.* Ресурсы могут иметь различные форматы и описывать данные в несогласованных между собой схемах;

— *расширяемая модель данных среды.* Необходимо обеспечить возможность периодического внесения новых поддерживаемых схем данных, описывающих наукоемкие ресурсы;

— *интероперабельность среды и открытость используемых стандартов.* Основной функцией среды является взаимодействие с внешними системами;

— *адаптируемость к требованиям пользователей.* Возможность предоставлять пользователю ресурсы в соответствии с его требованиями.

3. Схема функционирования

Рассмотрение требований и характеристик виртуальной среды указывает прежде всего на необходимость разработки технологического решения, которое позволит совместить разнородные модели и схемы данных и взаимодействовать с источниками по единой унифицированной схеме. На текущий момент известны три технологии подобной унификации:

— протокол Z39.50;

— протокол X.500 (LDAP);

— CORBA.

Наиболее приспособлены для решения поставленной задачи протоколы Z39.50 и LDAP. Протокол Z39.50 предлагает следующие возможности. *Абстрагированная модель схемы данных* позволяет осуществлять обмен данными без привязки к конкретной схеме. *Абстрагированная модель поиска* дает возможность осуществлять поиск в разнородных базах данных. При этом в рамках существующих реализаций протокола Z39.50 [7] предусмотрены механизмы преобразования данных из предметных схем в абстрактную схему протокола Z39.50. Следовательно, программное обеспечение на основе протокола Z39.50 позволяет организовать стандартизованный доступ к разнородным распределенным базам данных. Протокол LDAP предлагает широкий набор инструментов для работы с иерархическими данными, каталогами и справочной информацией, которые необходимы прежде всего для внутреннего функционирования виртуальной среды и ведения персональных данных. В целях этого проекта может быть использован корпоративный каталог LDAP Сибирского отделения РАН [8].

Принципиальная схема функционирования виртуальной среды приведена на рис. 1. Виртуальная среда состоит из реестра объектов и ресурсов, основного сервера Z39.50, нескольких функциональных модулей, а также web-интерфейса с публичным и административным разделами для доступа к различным функциям среды. Для каждого источника устанавливается отдельный сервер Z39.50, который осуществляет преобразование данных из схемы источника в абстрактную схему данных.

Каждый модуль виртуальной среды соответствует одному из указанных выше требований и реализует следующие функции.

1. Модули управления структурой каталогов ресурсов:

— создание нового каталога;

— модификация структуры каталога;

— добавление нового источника.

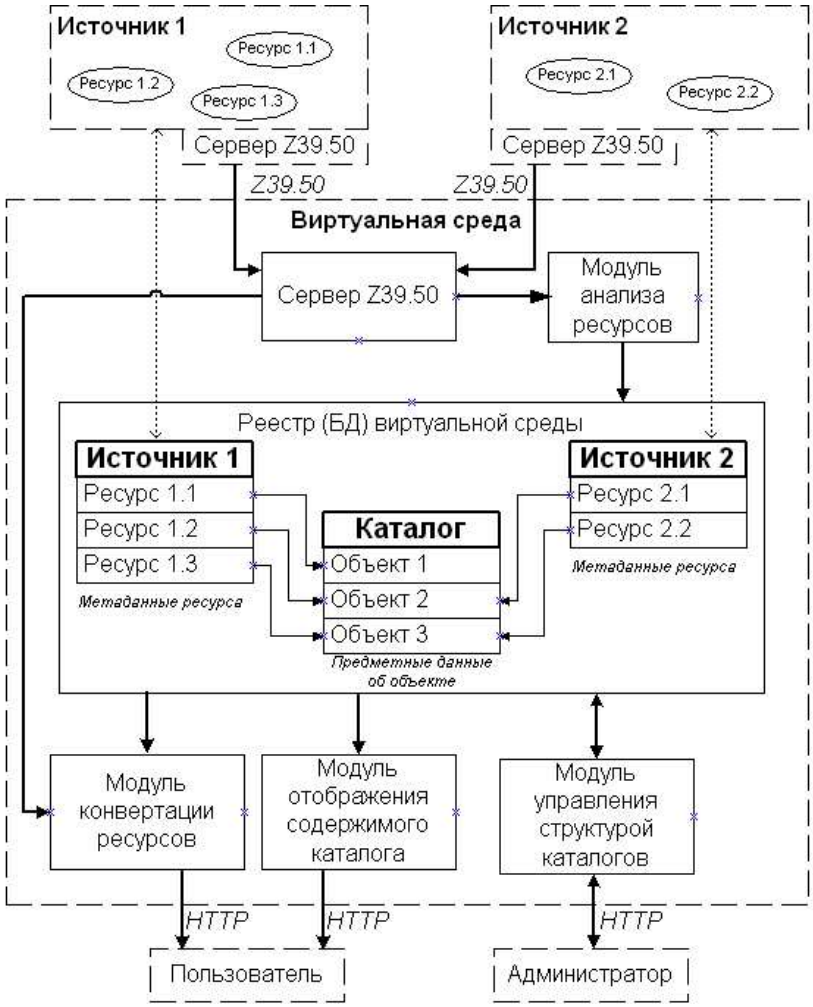

Рис. 1. Принципиальная схема функционирования виртуальной среды.

2. Модули аналитической обработки ресурсов:
— загрузка метаданных ресурса в каталог из источника (аннотирование);
— обновление метаданных ресурса в каталоге из источника;
— поиск смежных ресурсов;
— поиск релевантных объектов.
3. Модули отображения содержимого каталогов:
— поиск и просмотр объектов;
— просмотр метаданных ресурса;
— загрузка ресурса из источника.
4. Модули конвертирования ресурсов.
5. Модули диспетчеризации:
— мониторинг ресурсов;
— актуализация сведений о ресурсах;
— оповещение пользователей;
— автоматизированная отправка конвертированных ресурсов.

Одной из основных определяющих составляющих для функциональности любой информационной системы является ее модель предметной области (словари, логическая схема — основные сущности взаимосвязи). Построение словарей может быть частично авто-

матизировано [9], но в целом — это достаточно большая работа, особенно для задач аналитической обработки результатов. На начальном этапе функционирования виртуальной среды модель предметной области для фиксированных схем данных может быть задана на уровне самих алгоритмов обработки [10, 11].

Для каждого каталога необходимо описание его абстрактной схемы данных. Для определения смежных ресурсов следует производить сравнение сведений о ресурсах в абстрактной схеме данных. Критерием смежности является полное совпадение всех сведений или частичное совпадение для специально указанных полей (так называемый уникальный ключ).

Механизм определения релевантных объектов более сложен, чем нахождение смежных ресурсов. Его идея состоит в том, что фактически ресурс может описывать сведения не только об одном объекте, но и о нескольких. Например, ресурс, описывающий публикацию, содержит также сведения и об авторах. Таким образом, при описании абстрактной схемы для ресурса необходимо определить его "соответствие" не одному, а нескольким объектам виртуальной среды.

Наиболее сложной задачей является разработка моделей и средств конвертации ресурсов в формат данных, запрошенный пользователем. В рамках первой итерации разработки виртуальной среды может быть реализован следующий механизм — в качестве выходного формата представления данных в абстрактной схеме Z39.50 использовать язык XML. Посредством языка XSLT данные из базового XML-формата могут быть преобразованы в любой другой формат XML-семейства. Такое решение, с одной стороны, является нетрудоемким, а с другой — в силу распространенности языка XML потенциально позволяет покрыть большое число вариантов использования.

4. Систематизация и отбор ресурсов

Одной из важнейших задач проекта является задача автоматизации процесса отбора опубликованных в виртуальной среде ресурсов, которые могут представлять интерес для конкретного исследователя или группы совместно работающих исследователей. Для решения этой задачи предлагается алгоритм, основанный на установлении меры сходства между новым ресурсом (документом) и ранее отобранными документами. Рассмотрим данный алгоритм на примере работы с литературными источниками и публикациями. По сравнению с другими ресурсами вследствие своей очень слабой структурированности это наиболее сложный объект виртуальной среды.

У каждого исследователя за годы его работы образуется картотека библиографических описаний статей, книг и т. д., представляющих для него интерес. Основной критерий их отбора — личные интересы ученого. В настоящее время такие картотеки хранятся, как правило, на электронных носителях. Это позволяет создавать интегрированные картотеки путем объединения ресурсов совместно работающих исследователей.

Суть подхода состоит в том, что каждому документу сопоставляется его информационный образ, вычисляемый на основе метаописаний, полученных при аннотировании (каталогизации) документа и его библиографического описания.

Количественная характеристика меры сходства вычисляется по правилам работы с номинальными шкалами. Однако на практике ввиду неполноты библиографических описаний целесообразно осуществлять координатное индексирование документов. Суть его состоит в подсчете количества вхождений в текст (аннотацию, заголовок) документа так

называемых дескрипторов, т. е. лексических единиц информационно-поискового языка (совокупность систематизированных по смыслу дескрипторов вкупе с эксплицитно выраженными смысловыми связями между ними названа *тезаурусом*, точнее, *нормативным тезаурусом*). Важно отметить, что дескрипторами могут быть не только отдельные слова, но и словосочетания.

С использованием дескрипторного словаря предметной области (методика его создания на основе предметного указателя тематической энциклопедии подробно изложена в [9]) автоматически вычисляется координатный индекс документа, т. е. доля вхождений каждого дескриптора в текст. Далее, с учетом информации о том, к какому разделу классификатора относится тот или иной дескриптор (причем вес дескрипторов заглавия больше веса дескрипторов аннотации), проводится автоматическая классификация документа одним или несколькими кодами классификатора.

Заметим, что изложенный алгоритм измерения меры сходства может быть положен в основу некоторой экспертной системы, обладающей определенными продукционными правилами. Так, значения весовых коэффициентов шкал при подсчете меры близости может определяться предполагаемой апостериорной достоверностью данных соответствующей шкалы. Например, полное (или даже "почти полное") совпадение значений атрибута "авторы" документа с проекцией множества документов более весомо в случае, когда количество значений этого атрибута в документе достаточно велико (по сравнению со случаем, когда документ имеет всего одного автора). В такой ситуации мы можем увеличивать значение соответствующего весового коэффициента с одновременным пропорциональным уменьшением других коэффициентов. Используя те или иные квантили на множестве значений мер сходства коллекции новых документов, мы можем ранжировать эти документы по "степени интересности".

5. Информационная среда "Атмосферные аэрозоли Сибири"

Сеть Интернет, предоставляющая постоянный, в режиме реального времени, доступ к различным базам данных, способствует информационному обеспечению научных исследований.

В течение последних 15 лет специалисты институтов Сибирского отделения РАН (Института химической кинетики и горения, Института неорганической химии, Лимнологического института, Института водных и экологических проблем, Кемеровского, Красноярского, Новосибирского и Томского научных центров) проводили в полевых и стационарных условиях регулярные измерения временных характеристик атмосферы. Мониторингом были охвачены территории Западной и Восточной Сибири, Алтайского и Красноярского краев, а также Арктический бассейн России. Этот огромный объем экспериментальных данных неизмеримо расширил и во многом изменил существующие представления о влиянии индустриальных центров на окружающую среду региона. Изучение проблемы глобальных изменений окружающей среды и техногенного влияния на нее индустриальных центров вызвало необходимость комплексного сбора необходимой информации, в том числе и по атмосферным аэрозолям (АА). При этом проводились измерения их массовой концентрации, многоэлементного состава, содержания органического и неорганического углерода, а также регистрировалась пространственно-временная динамика химического состава.

До недавнего времени все эти данные и информация хранились на бумажных носителях: в виде записей в рабочих журналах, на лентах самописцев и фотопленках, лабораторных отчетах и научных публикациях. К сожалению, весь этот ценный эмпирический материал был доступен только его владельцу и, как правило, терялся с уходом хозяина.

Эффективность использования огромного объема эмпирической информации может быть повышена за счет разработки, развития и использования информационных технологий, на основе создания как новых специализированных моделей, ориентированных на решение относительно узких классов задач, так и путем совершенствования технологий организации процесса хранения и обработки данных с использованием существующих методов и моделей.

В этих условиях крайне необходимо выработать и последовательно реализовать единую информационную политику, обеспечивающую развитие и эффективное использование современных информационных технологий и информационных ресурсов. Важной задачей в области повышения эффективности использования региональных информационных ресурсов является формирование единой информационной среды, которая должна включать два основных компонента. Первый обеспечит методологическую базу хранения, обработку и анализ. Это модели систем, процессов и задач предметной области. Второй представляет собой технологическую базу моделирования в виде распределенной интегрированной информационно-вычислительной среды, позволяющей реализовать эти модели и разрешить работу с ними достаточно широкому кругу пользователей [11].

Для решения поставленной задачи необходимо, во-первых, разработать подходы и технологии, обеспечивающие виртуальную интеграцию описаний разнородных информационных ресурсов, расположенных на серверах различных организаций, в единую базу данных на основе открытых международных стандартов, во-вторых, создать систему базовых понятий для описания электронных коллекций атмосферных аэрозолей, в-третьих, построить информационные модели описаний микрофизических и химических характеристик на основе результатов измерения характеристик атмосферных аэрозолей различной природы. Для этого следует провести анализ технологий использования, хранения и обработки данных, построить модели связей биогеохимических циклов с факторами физической среды, обосновать информационную безопасность взаимодействия пользователей: конфиденциальность, целостность и доступность. Современные Интернет-технологии позволяют по-новому организовывать хранилища данных и доступ к ним [12].

Анализ существующего состояния сбора, хранения и обработки результатов измерений характеристик атмосферы показывает, что различными организациями СО РАН создаются собственные информационные компьютерные системы. Они разрабатываются, внедряются и функционируют, как правило, для себя, разобщенно, без координации и согласования с кем-либо структуры, содержания, способов и форматов хранения информации и доступа к ней. Это приводит к дублированию работ, избыточности первичной информации, многократному ее вводу, приводящему к нарушению ее целостности и низкой достоверности, значительному удорожанию разработок и эксплуатации систем. Ведомственная разобщенность затрудняет обмен информацией и доступ к ней, что осложняет информационное обеспечение заинтересованных лиц.

Специалисты Института вычислительных технологий и Института химической кинетики и горения СО РАН разработали информационную модель и структуру метаданных, используя обобщенный подход для формирования и заполнения входных данных, включая их унификацию и связи для создания новой информационно-вычислительной системы, — атлас “Атмосферные аэрозоли Сибири” (`http://web.ict.nsc.ru/aerosol/`).

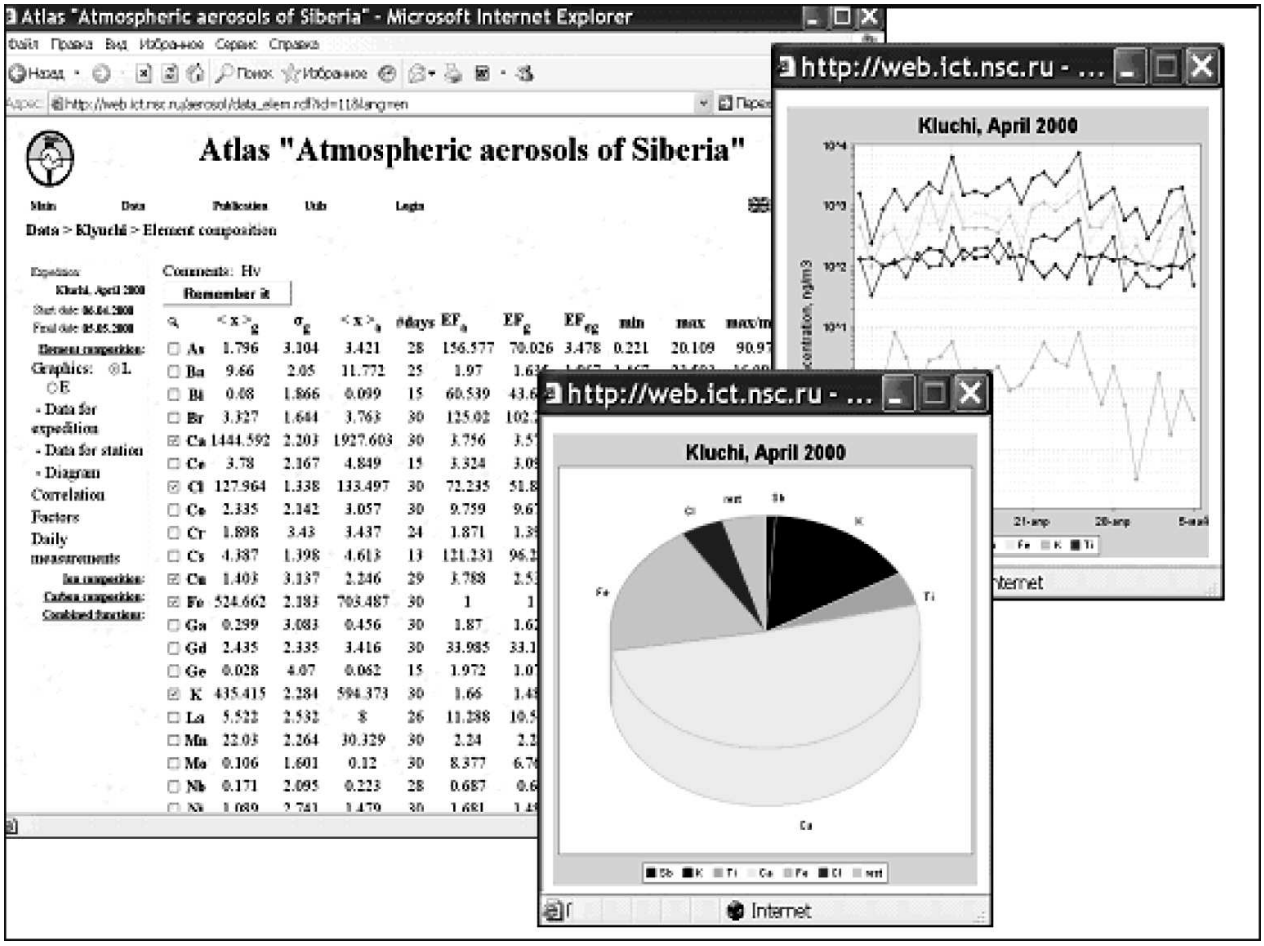

Рис. 2. Атлас “Атмосферные аэрозоли Сибири”.

В этой программной системе аккумулируется весь накопленный до настоящего времени эмпирический материал, обеспечивая решение различных задач в области сбора, обработки и публикации сведений об атмосферных аэрозолях. Важной особенностью созданной информационной системы является коллективная работа с архивами данных.

Атлас решает различные задачи в области сбора, обработки и публикации сведений об атмосферных аэрозолях, такие как:

— добавление, хранение, редактирование исходных данных об атмосферных аэрозолях, включая процедуры авторизации пользователей и ограничение доступа к различным информационным структурам;

— реализация функций математической обработки данных о химическом составе аэрозолей (статистические характеристики, корреляционный и факторный анализ временных рядов и др.);

— разработка web-интерфейса для отображения исходных данных и результатов обработки, включая табличное и графическое представление;

— в web-интерфейсе и на уровне данных реализация поддержки русского и английского языков.

Информационная система предоставляет не только средства хранения данных, визуализацию их зависимостей от времени в табличном и графическом виде, но и возможность обработки данных специализированными алгоритмами для решения различных задач ис-

следования окружающей среды. В ней было выделено несколько классов задач, каждый из которых порождает собственные модели хранения и обработки данных. Для интерпретации полученных данных используются различные статистические методы: корреляционный, факторный, дискриминантный, кластерный и фурье-анализы временных рядов. К настоящему времени в информационно-вычислительной системе реализованы следующие функции:

— подсчет наиболее значимых статистических характеристик покомпонентного химического состава аэрозолей: коэффициентов обогащения, среднего арифметического и геометрического, геометрического отклонений;

— корреляционный анализ, показывающий взаимосвязь химических элементов и их влияние друг на друга;

— факторный анализ временных рядов, основанный на корреляционном анализе и позволяющий выделить ряд факторов, которые обусловливают присутствие тех или иных химических соединений в составе аэрозоля. Результаты обработки и расчетов хранятся в единой базе данных атласа совместно с результатами измерений, с автоматической поддержкой необходимых ссылочных связей и метаданных.

Важными преимуществами предлагаемой технологии создания информационных систем и электронных архивов являются: значительное расширение аудитории, которая может ознакомиться с экспонатами за счет их виртуального присутствия в сети Интернет и на различных носителях (CD); хранение всех описаний и взаимосвязей экспонатов в единой базе данных; сохранение научных результатов для следующего поколения исследователей.

Заключение

В данной работе представлены идейные соображения по созданию виртуальной среды для обмена результатами научных исследований. Очевидно, что рассмотрена лишь небольшая часть технических вопросов. Выделены основные принципы функционирования виртуальной среды, и дальнейшая работа состоит в проработке деталей функционирования и взаимодействия ее частей.

Реализация виртуальной среды позволит конечному пользователю работать в тематическом портале, функционирующем в качестве системы управления и поддержки исследований по окружающей среде Сибири в соответствии со следующими требованиями.

1. **Уровень услуг** — пользователь имеет доступ к одним и тем же услугам независимо от своего местоположения.

2. **Актуальность** — должна предоставляться максимально свежая информация. Обеспечивается децентрализованным управлением: каждый сервер отвечает только за свою локальную часть базы, чтобы обновление данных и сопровождение можно было выполнять немедленно.

3. **Единое пространство имен**, позволяющее представлять информацию как единый логический каталог.

4. **Схемы данных, допускающие локальные расширения.**

5. **Единый протокол доступа.** Приложения, нуждающиеся в ресурсах справочника, должны производить запросы, используя стандартизированный протокол.

Список литературы

[1] Жижимов О.Л., Федотов А.М., Чубаров Л.Б., Шокин Ю.И. Технология создания распределенных информационно-вычислительных ресурсов СО РАН // Тр. Первой Междунар. конф. — САИТ-2005. 12–16 сент. 2005 г., Переславль-Залесский. Т. 2: "Системный анализ и информационные технологии". М., 2005. С. 161–165.

[2] The Grid: Blueprint for a New Computing Infrastructure / Ed. by I. Foster, C. Kesselman. Morgan Kaufmann Pub., San Francisco, CA. 1999.

[3] Федотов А.М., Гуськов А.Е. Информация в Интернете: публикация, поиск, анализ // Информационные технологии в высшем образовании. 2004. № 4. С. 17–35.

[4] Шокин Ю.И., Федотов А.М., Гуськов А.Е. и др. Электронные библиотеки — путь интеграции информационных ресурсов Сибирского отделения РАН // Вест. КазНУ. Спецвыпуск. г. Алматы, Казах. нац. ун-т им. аль-Фараби. 2005. № 2. С. 115–127.

[5] Гуськов А.Е. Модель виртуальной среды для обмена результатами научных исследований // Тр. Междунар. конф. "Вычислительные и информационные технологии в науке, технике и образовании", 20–22 сент. 2006. Павлодар: ПГУ им. Торайгырова, 2006. Т. 2. С. 372–380.

[6] Федотов А.М. Концептуальные подходы к построению распределенных систем // Тр. Междунар. конф. по вычислительной математике МКВМ-2004. Новосибирск: Изд-во ИВМиМГ СО РАН, 2004. С. 132–143.

[7] Жижимов О.Л., Мазов Н.А., Федотов А.М., Шокин Ю.И. Сервер ZooPARK как сервер для построения распределенных информационных систем // Информационные технологии в высшем образовании. Алматы: Изд-во Казах. нац. ун-та им. Аль-Фараби. 2005. Т. 2, № 1. С. 53–67.

[8] Жижимов О.Л., Турпанов А.А., Федотов А.М. Корпоративный каталог СО РАН // Тр. 8-й Всерос. науч. конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" — RCDL-2006, Суздаль, Россия, 2006. С. 226–230.

[9] Барахнин В.Б. Разработка тезауруса предметной области "Математика" // Вычисл. технологии. Т. 8: Региональный вестник Востока. № 3 (19), совместный выпуск. 2003. Ч. 1. С. 111–115.

[10] Гордов Е.П., Ковалев С.П., Молородов Ю.И., Федотов А.М. Web-система управления знаниями об окружающей среде // Вычисл. технологии. Т. 10. Спецвыпуск: Тр. Междунар. конф. и школы молодых ученых "Вычислительные и информационные технологии для наук об окружающей среде". Новосибирск, 13–15 марта 2005 г. Ч. 2. С. 12–19.

[11] Куценогий К.П., Куценогий П.К., Молородов Ю.И., Федотов А.М. Разработка структуры метаданных по атмосферным аэрозолям на основе информационной модели // Вычисл. технологии. 2004. Т. 9. Спецвыпуск: Тр. Междунар. конф. "Вычислительно-информационные технологии для наук об окружающей среде". Ч. 2. С. 25–33.

[12] Фазлиев А.З. Информационные ресурсы и Интернет-технологии для наук об окружающей среде // Там же. Ч. 1. С. 11–20.

Поступила в редакцию 19 октября 2006 г.