Разработка палеонтологической информационной системы "PaleoData" на базе ИСИР для Палеонтологического института РАН♣

© Кузнецов А.М., Голубев В.К., Ляльченко Р.В., Удова Е.С.

Палеонтологический институт РАН
akuzn@paleo.ru, vg@paleo.ru

Аннотация

Работа посвящена вопросам создания единого хранилища палеонтологической информации, которое может быть использовано в исследованиях по стратиграфии, палеобиогеографии, систематики, статистической палеобиологии. В статье кратко рассмотрены основные аспекты разработки информационной системы для Палеонтологического института РАН на базе ИСИР. Описана логическая модель палеонтологических данных на основе составленной онтологии предметной области. Проведена интеграция модели данных палеонтологической системы с моделью данных ИСИР.

1. Описание предметной области

Палеонтология - наука об органическом мире геологического прошлого, о его становлении и развитии, об условиях его существования. В основе всех палеонтологических выводов и реконструкций лежат сведения о конкретных ископаемых остатках вымерших организмов, представленных как отдельными экземплярами, так и целыми коллекциями. Именно палеонтологические коллекции являются предметной основой для систематики и номенклатуры ископаемых организмов, а также для установления геологического возраста отложений, стратиграфической корреляции геологических разрезов, построения опорных стратиграфических схем. Особое значение для поддержания стабильности номенклатуры и для исследований в области систематики имеет надежное хранение типовых и оригинальных (монографических) коллекций.

Труды 7ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2005, Ярославль, Россия, 2005.

Не менее важным, чем сохранность коллекций, является их информационное сопровождение, поскольку сведения о конкретных коллекционных экземплярах, будучи первичными, являются наиболее объективными в палеонтологии и исторической геологии. Однако данная информация, в большинстве случаев, оказывается рассеянной среди огромного количества публикаций (причем опубликованные сведения о сохранности и местах хранения экземпляров часто не соответствуют реальности в настоящее время) или содержится только на этикетках, в описях коллекций и полевых дневниках (то есть практически недоступна подавляющему большинству исследователей).

В настоящее время в некоторых музеях и институтах России до сих пор используются информационно-поисковые системы (картотеки) на бумажных носителях. Основные недостатки работы с такими системами:

- для добавления новых данных требуются большие затраты труда;
- неизбежно накапливается большое количество ошибок;
- не обеспечивается надежность перекрестных ссылок (например, если экземпляр описан в нескольких публикациях и т.п.);
- записи в этих системах сортируются только в определенном порядке, что сильно затрудняет или делает практически невозможным пересортировку или поиск нужных для конкретной задачи записей;
- такие картотеки существуют в единственном экземпляре и могут быть частично или целиком утрачены;
- доступ к информации имеет только узкий круг людей, непосредственно работающих с этими данными.

Все это сильно снижает эффективность использования палеонтологических данных [1].

В Палеонтологическом институте РАН в настоящее время зарегистрировано более пяти тысяч коллекций ископаемых остатков различного геологического возраста - от докембрия до современности, происходящие как с территории

бывшего СССР, так и из различных местонахождений всех частей света. В ПИН РАН только централизованно хранятся коллекции, описанные в более чем 400 монографиях и статьях сотрудников института и специалистов из других организаций. Ежегодно принимаются на хранение коллекции к новым публикациям. Ведется постоянная работа по инвентаризации старых коллекций к работам, опубликованным начиная с конца XIX века. Кроме того, значительное количество монографических коллекций хранится в лабораториях института [2]. Но, к сожалению, большая часть сведений по указанным коллекциям практически недоступна мировому научному сообществу.

Развитие современных информационных технологий дает возможность научному сообществу решить проблемы хранения, организации быстрого доступа и обработки больших массивов информации по коллекциям (в том числе и палеонтологическим) путем создания электронных баз данных и информационных систем.

2. Интегрированная система информационных ресурсов (ИСИР)

Основной задачей Интегрированной Системы Информационных Ресурсов (ИСИР) является объединение информационных ресурсов различного типа в единую систему, обеспечивающую их поддержку и предоставление пользователям Интернет. Единая информационная среда формируется из разнородных и распределенных источников информации, называемых репозиториями, содержащих ресурсы в реляционных и объектных базах данных, LDAP - каталогах, XML и RDF хранилищах и т.п. ИСИР предоставляет ряд служб по поддержке репозиториев, например, репликацию и обмен данными, индексирование и поиск ресурсов, технологию построения веб-порталов для доступа к данным и манипулирования ими [3]. В основе концепции построения ИСИР лежит современная технология открытых систем, предполагающая использование единой метаинформации, описывающей ресурсы различного типа.

Модель данных ИСИР построена на концепции взаимосвязанных типизированных ресурсов [4, 6] таким образом, что она в заданной предметной области инкорпорирует достаточно большой класс электронных коллекций. Она содержит основные сущности и их атрибуты из предметной области, а также основные связи между ними. Модель данных содержит ресурсы следующих типов: *Организационная единица* (уточняющаяся сущностями *Организация* или *Подразделение*), *Персона*, *Проект*, *Публикация* [5]. Технологии ИСИР предоставляют пользователю навигацию между различными видами ресурсов и единый механизм доступа ко всем ресурсам через стандартный браузер.

3. Информационная система Палеонтологического института РАН

Разработка информационной системы Палеонтологического института РАН на базе ИСИР позволит решить важнейшие задачи, связанные с информацией о палеонтологических данных:

- создание и постоянное пополнение массива данных по палеонтологии;
- организацию быстрого доступа к ним;
- предоставление требуемой информации в удобном для дальнейшей обработки виде;
- предоставление исследователю первичной палеонтологической информации без какой-либо специальной ее обработки.

Палеонтологическая информационная система будет содержать все данные о коллекционных экземплярах, включая местонахождение, геологический возраст, родовые и видовые названия, под которыми экземпляр описан в литературе, группа (таксон высокого ранга), статус экземпляра (категория типа, оригинал или "дублет"), полную библиографическую информацию о публикациях, в которых описан экземпляр, ссылки на страницы и изображения, место хранения в институте (музее).

Доступ через Интернет к информации о палеонтологических коллекциях позволит широкому кругу специалистов, студентов и аспирантов активно использовать как сами коллекции, так и разнообразную информацию о них в современных исследованиях или в процессе обучения палеонтологии и исторической геологии.

Кроме того, система будет использоваться для информационного обеспечения хранения и инвентаризации коллекций, контроля за движением экземпляров внутри музея (экспозиция, техническая и научная обработка) и за его пределами (временная и постоянная передача в другие организации для изучения, выставки и т.д.), автоматизации рутинных операций, подготовки к публикации печатных каталогов.

4. Объектная модель данных информационной системы

Все информационные объекты базовой схемы ИСИР делятся по следующим типам:

Ресурсы - логические единицы хранения, условные единицы информации, которые могут существовать вне зависимости от других информационных объектов, представляют первичный интерес для пользователей системы.

Структуры - зависимые объекты, существующие только в контексте логических единиц хранения (ресурсов).

Термы - объединяющиеся в "таксономии" - централизованные словари, классификаторы, тезаурусы ИСИР.

Схема данных палеонтологической системы получается путем расширения базовой схемы ИСИР

новыми классами и отношениями. Кроме того, анализ модели данных ИСИР показал, что необходимо ввести новый тип информационных объектов - Объектные отношения. К этому типу относятся объекты, которые рассматриваются как отношения с атрибутами.

4.1. Ресурсы палеонтологической системы

Базовая модель данных ИСИР содержит ресурсы следующих типов:

- *Организационная единица* (Unit), уточняющаяся сущностями *Организация* (Organization) или *Подразделение* (Department).
- *Персона* (Person).
- *Проект* (Project).
- *Публикация* (Publication).

При разработке палеонтологической системы были дополнительно выделены следующие ресурсы: *Образец* (Specimen), *Особь* (Individuum), *Хранимый Объект* (Storing Object), *Место Хранения* (Storage Position), *Таксон* (Taxon), *Стратон* (Straton), *Разрез* (Section), *Местонахождение* (Site), *Коллекционный массив* (Storing Array), *Зарегистрированные коллекции* (Collection Registration).

Одним из основных ресурсов системы является *Образец* (Specimen) - информационный объект, предоставляющий сведения об экземпляре палеонтологической коллекции. Каждый образец характеризуется номером коллекции, номером экземпляра, номером элемента (части экземпляра) и многими другими свойствами. Экземпляр может принадлежать какой-нибудь особи, являться хранимым объектом. Кроме того, образец имеет многочисленные связи с ресурсами ИСИР *Персона* и *Публикация*.

Ресурс *Особь* (Individuum) - информационный объект предоставляет сведения о палеонтологической особи, то есть о конкретном вымершем организме, чьи ископаемые остатки представлены в коллекции. Одна особь может быть представлена в коллекции несколькими образцами. Важным в описании особи является сведения о ее местонахождении (географические, литологические, стратиграфические и др. данные), максимальный типовой статус (голотип, паратип и т.п.).

Ресурс *Хранимый Объект* (Storing Object) - информационный объект, предоставляющий сведения о хранимом объекте. Хранимым объектом может быть как один образец, так и совокупность из нескольких образцов. Каждый хранимый объект имеет место хранения (постоянное или временное). Временное место хранения возникает, например, когда объект увозят на выставку. В этом случае еще устанавливается и цена (стоимость) хранимого объекта. Также имеется организация, ответственная за хранение объекта.

Ресурс *Место Хранения* (Storage Position) - объекты этого типа предоставляют сведения о месте хранения образцов или коллекций ископаемых организмов. Местом хранения может выступать как целая организация (например, Палеонтологический институт РАН) или ее подразделения, так и любые объекты, используемые для хранения (витрина, шкаф, сейф и т.п.). Место хранения объекта может иметь иерархическую структуру (например, могут быть указаны организация, номер комнаты, номер шкафа и номер лотка, в котором хранится образец). Объекты, относящиеся к этому типу ресурсов, характеризуются названием, типом хранения, датой создания и ликвидации, причиной ликвидации.

Ресурс *Таксон* (Taxon) - информационный объект предоставляет номенклатурные сведения о биологических таксонах. Биологический таксон - это группа связанных той или иной степенью общности свойств и признаков организмов, имеющая собственное название и определенный таксономической ранг, который отражает иерархическое положение таксона в системе органического мира Земли. В палеонтологической информационной системе таксон участвует в связях с такими ресурсами, как *Персона, Публикация, Образец*.

Ресурс *Разрез* (Section) - данный ресурс предоставляет сведения о геологических разрезах: географическое и структурно-тектоническое положение, литологическое описание.

Ресурс *Стратон* (Straton) - информационный объект, предоставляющий номенклатурные сведения по стратиграфическим подразделениям, или стратонам. Стратон – совокупность горных пород определенного геологического возраста, имеющая собственное название и определенный таксономический ранг.

Ресурс *Местонахождение* (Site) - информационный объект, предоставляющий данные о местонахождениях ископаемых остатков. Местонахождение – непрерывная (с точки зрения исследователя) совокупность горных пород, содержащих ископаемые остатки вымерших организмов определенной таксономической группы. Ресурс включает следующие сведения: собственное название или персональный номер местонахождения, географическое и стратиграфическое положение, дата открытия и др. Важными являются связи с другими ресурсами: *Персоной, Публикацией, Разрезом, Стратоном*.

Ресурс *Зарегистрированные коллекции* (Collection Registration) - информационный объект, предоставляющий сведения о коллекциях, зарегистрированных в Палеонтологическом институте РАН.

Ресурс *Коллекционный Массив* (Storing Array) - информационный объект, предоставляющий описание коллекционных массивов, хранящихся в Палеонтологическом институте РАН

4.2 Структуры палеонтологической системы

Базовым классом для всех видов структурных зависимых объектов в схеме ИСИР является kernel:Structure. В схеме данных

палеонтологической системы было выделено два структурных класса: *Палеонтологическая Дата* (PaleoDate) и *Цена* (Price).

Палеонтологическая Дата (PaleoDate) - объект принимает значение, представляющее некоторое временное событие. Обычный формат представления даты, который используется для хранения значений типа дата в базе данных, в данном случае не подходит. Это связано с тем, что в древность ископаемых остатков вымерших организмов может составлять от первых тысяч до первых миллиардов лет. Очень часто дата задается не конкретным значением, а интервалом значений (например, 250-260 млн лет назад).

Объект этого класса имеет следующую структуру:

- Дата.
- Дата начала временного интервала.
- Дата окончания временного интервала.
- Период (например, до н.э. или н.э.).
- Тип временного интервала: год, век, тысячелетие и т.п.
- Комментарий.

Цена (Price) - зависимый объект, который существует только в контексте ресурса *Хранимый Объект* (Storing Object). Предоставляет сведения и историю об оценочной стоимости *Хранимого Объекта*. Имеет простую структуру:

- Стоимость объекта.
- Тип валюты стоимости.
- Тип стоимости (закупочная, страховая, оценочная).

4.3 Словари палеонтологической системы

В палеонтологии все исследуемые объекты, а также их свойства подлежат определенной классификации. Поэтому названия тех признаков (категорий), по которым происходит та или иная классификация, были вынесены в отдельные словари. На данный момент в палеонтологической системе выделено 14 словарей: *Группа* (Group), *Характер Хранимого Объекта* (ObjectNature), *Категория Открытой Номенклатуры* (OpenNomenclature), *Краткое Описание Образца* (ShortDescription), *Статус Образца* (SpecimenStatus), *Тип Места Хранения* (StorageType), *Ранг Стратона* (StratonType), *Ранг Таксона* (TaxonRank), *Характер Отношений Таксонов* (TaxonRelationType).

4.4 Объектные отношения палеонтологической системы

Особенностью организации палеонтологической информационной системы является не только сложность самих объектов, но и отношений между ними. Взаимодействие некоторых ресурсов друг с другом характеризуется рядом дополнительных свойств. В основном, это следующие свойства:

- дата установления (описания) отношения;
- персона, установившая отношение;
- публикация, в которой описано данное отношение.

Объекты, обладающие такими атрибутными отношениям, были выделены в отдельный класс - *Объектные Отношения* (ObjectsRelation).

Класс *Стратиграфия Разреза* (SectionStratigraphy) - информационный объект, предоставляющий сведения разных авторов по стратиграфическому расчленению геологических разрезов. Этот объект представляет собой отношение между следующими ресурсами: *Разрез, Местонахождение, Стратон, Персона, Публикация*.

Класс *Палеонтология Разреза* (SectionPaleontology) - этот информационный объект предоставляет сведения разных авторов по палеонтологической характеристике геологических разрезов. Он устанавливает отношение между ресурсами: *Разрез, Местонахождение, Таксон, Персона, Публикация*.

Класс *Корреляция Разрезов* (SectionCorrelation) - информационный объект, предоставляющий сведения разных авторов по геологической корреляции разрезов. Является отношением между двумя *Разрезами*, *Публикацией* и *Персоной* (автором корреляции разрезов).

Класс *Корреляция Стратонов* (StratonCorrelation) - информационный объект, предоставляющий сведения разных авторов по стратиграфическому взаимоотношению стратонов. Является отношением между двумя *Стратонам, Публикацией* и *Персоной* (автором корреляции стратонов).

Класс *Описание Образца* (SpecimenReference) - информационный объект, предоставляющий опубликованные и неопубликованные палеонтологические сведения об экземпляре. Устанавливает связь между ресурсами: *Образец, Таксон, Публикация, Персона*.

Класс *Система Таксонов* (TaxonSystem) - информационный объект, предоставляющий сведения разных авторов об отношениях между таксонами, в том числе и систематических. Является отношением между двумя *Таксонами, Публикацией* и *Персоной*.

Класс *Агенты Публикации* (PublicationAgents) - информационный объект, устанавливающий связь между ресурсами информационной системы и *Публикациями*.

4.5 Расширение модели данных ИСИР

Создание классов объектной схемы данных палеонтологической системы повлекло расширение классов базовой схемы ИСИР. Для интеграции модели данных палеонтологической системы с моделью данных ИСИР, были введены новые роли ресурсов ИСИР - *Персона* и *Организационная единица*. В соответствии с новыми ролями свойства этих ресурсов расширились.

Ресурсу Организационная единица присвоены следующие роли:

- Хранение объекта.
- Ответственность за объект.
- Ресурсу *Персона* присвоены следующие роли:
- Автор определения.
- Автор коллекции.
- Автор описания.
- Куратор.
- Автор нового видового сочетания.
- Первооткрыватель местонахождения.
- Автор стратона.
- Объект описания.

5. Пользовательский интерфейс палеонтологической системы

Доступ пользователей к информации, хранящейся в палеонтологической системе, осуществляется через веб-интерфейс. В системе имеется функция авторизации пользователей (поддерживается ядром ИСИР). Для авторизации в системе необходимо ввести логин и пароль. Неавторизованному пользователю могут быть доступны только поиск и просмотр некоторой информации. Часть информации может быть скрыта от неавторизованного пользователя даже на просмотр (например, данные о местонахождениях ископаемых организмов). Авторизованные пользователи могут иметь права администратора, которые позволяют управлять всеми ресурсами, или относиться к какой-нибудь группе пользователей, обладающей правами на управление частью ресурсов. Ядро ИСИР предоставляет сервис аутентификации, который осуществляет процесс идентификации пользователя при обращении к закрытым ресурсам системы. Процедура авторизации, проверяя полномочия пользователя, разрешает или отказывает ему в выполнении определенной операции, например, в доступе к информации. Кроме того, ядро ИСИР обеспечивает поддержку аудита изменений ресурсов, что немаловажно при работе с палеонтологическими данными. Всегда можно узнать дату создания и модификации объекта, а также пользователей, создавших и изменивших его.

6. Формы ввода, вывода и поиска данных

Для каждого типа ресурсов палеонтологической системы были разработаны просмотровые страницы. Это XSP-страницы, формирующие XML-представление выборки необходимой объектной информации из хранилища, и шаблоны XSLT-трансформации, которые визуализируют полученные XML-данные. Так как классы базовой схемы ИСИР были расширены, то потребовалась доработка уже существующих шаблонов ресурсов ИСИР.

Просмотровые страницы отображают всю доступную для данного пользователя информацию об объекте, но не содержат интерактивных элементов. Информация на таких страницах представляется в виде:

- текста для неинтегрированных ресурсов;
- гиперссылки на интегрированный ресурс.

Информационная система позволяет не только просматривать данные о ресурсах, но и редактировать информацию о них, создавать новые объекты или удалять существующие. Такое управление информацией осуществляется через интерактивные страницы системы. Интерактивные страницы, созданные с использованием технологий JSP + JXPath + FormBuilder, представляют собой формы ввода информации. Данные могут вводиться в виде текста или представлять собой ресурс, с которым нужно интегрировать редактируемый объект. Проверка прав доступа пользователя к формам и синтаксическая проверка введенных данных осуществляются автоматически (это предоставляется сервисом "Formbuilder").

На данный момент в палеонтологической системе разработаны страницы ввода информации для ресурсов: *Таксон, Стратон, Образец, Особь, Разрез*.

В палеонтологическую информационную систему уже перенесено значительное количество информации (Образцов - 30089, Таксонов - 7918, Стратонов - 16138, Разрезов - 4754, Публикаций - 1510, Персон - 1438). Естественно, для удобства работы с этой информацией необходим поиск по атрибутам ресурсов.

Страницы с поисковыми формами и краткими результатами поиска также создаются с использованием технологий JSP + JXPath + FormBuilder. Как правило, страница поиска содержит следующие элементы:

- поля для ввода запроса (текстовые или справочники);
- кнопки: "Очистить" - сброс введенных данных, "Поиск" - запуск поиска.

Вывод результатов поиска осуществляется на страницу результатов. Страница результатов содержит:

- общее количество найденных объектов;
- гиперссылки на просмотровые страницы конкретных объектов.

Результаты поиска выдаются частями (разбиваются на отдельные страницы). Вверху страницы результатов имеется навигатор - список номеров страниц (в виде гиперссылок), которые содержат отдельные части результата поиска. Для перехода к следующей части результата, нужно просто выбрать номер желаемой страницы.

Заключение

В процессе работы были решены следующие задачи:

1. Проведен онтологический анализ палеонтологических данных.

2. Разработана объектная модель палеонтологической компоненты. На ее основе расширена объектная схема репозитория ИСИР.

3. Построено объектно-реляционное отображение объектной модели в модель данных хранилища.

4. Разработаны страницы выборки объектной информации из хранилища. Созданы XSLT-шаблоны, отвечающие за визуализацию полученных XML-данных.

5. Разработаны формы регистрации (редактирования) ресурсов, формы атрибутного поиска ресурсов.

Итогом работы является создание пилотной палеонтологической информационной системы, которая на текущий момент доступна по адресу http://paleodata.paleo.ru. Использование полученных наработок позволяет решать сходные задачи в смежных с палеонтологией науках, таких как археология, биология, геология и т.п.

Литература

[1] Голубев В.К., Лисицын Д.В., Лебедев О.А., Кузнецов А.М., 2001. "Палеодата": база данных по палеонтологии, разрабатываемая в Палеонтологическом институте РАН // Информационные и телекоммуникационные ресурсы в зоологии и ботанике. 2-й Международный симпозиум. СПб.: Зоол. ин-т РАН. С. 112-113.

[2] Кузнецов А.М., Голубев В.К., 2001. Палеонтологическая информационная система "Палеодата": Internet-версия // Геологические, геофизические и геохимические исследования юго-востока Русской плиты: материалы научной межведомственной конференции (Саратов, 2-4 апреля 2001 г.). Саратов: Изд-во СО ЕАГО. С. 51-52

[3] Бездушный А.А., Бездушный А.Н., Нестеренко А.К., Серебряков В.А., Сысоев Т.М. Архитектура RDFS-системы. Практика использования открытых стандартов и технологий Semantic Web в системе ИСИР // Пятая Всероссийская научная конференция: "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - RCDL'2003. http://rcdl2003.spbu.ru/proceedings/J1.pdf

[4] Бездушный А.А., Бездушный А.Н., Нестеренко А.К., Серебряков В.А., Сысоев Т.М. RDFS как основа среды разработки цифровых библиотек и Web-порталов // Электронные библиотеки. 2003. 6. № 3. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part3/BBNSS

[5] Бездушный А.Н., Жижченко А.Б., Кулагин М.В., Серебряков В.А. Интегрированная система информационных ресурсов РАН и технология разработки цифровых библиотек. // Программирование. 2000. 26. № 4. С. 177-185.

[6] Бездушный А.А., Бездушный А.Н., Нестеренко А.К., Серебряков В.А., Сысоев Т.М. Архитектура и технологии RDFS-среды разработки цифровых библиотек и Web-порталов // Электронные библиотеки. 2003. 6. №4. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part4/BNSBS

Development of the paleontological information system “PaleoData” on the ISIR base for the Paleontological institute of the Russian Academy of Science

Kuznetsov A.M., Golubev V.K., Lyalchenko R.V., Udova E.S.

The work is focused on the problems of elaboration of the unified paleontological database, which can be used in the systematical, stratigraphic, paleogeographic studies, or in statistics in paleobiology. The general aspect of development of the informational system for the Paleontological Institute of the Russian Academy of Sciences basing on ISIR are discussed in the paper. These aspects are the following: the description of the logical model of paleontological data on the base of the composed ontology of the subject field, integration of the paleontological data model with data model of ISIR.

♣ Работа выполнена при поддержке Российского Фонда Фундаментальных Исследований (проект № 03-07-90155) и “Фонда содействия отечественной науке”.