Итоги и перспективы разработки информационной системы по биоразнообразию животных России (ZOODIV – BIODIV)♣

© И.С. Смирнов, О.Н. Пугачев, А.Г. Кирейчук, М.Б. Дианов, А.Л. Лобанов, Р.Г. Халиков, А.А. Голиков, В.А. Кривохатский

Учреждение Российской академии наук Зоологический институт РАН, г. Санкт-Петербург
smiris@zin.ru

Аннотация

Более шести лет в Зоологическом институте РАН разрабатывается информационно-поисковая система по биологическому разнообразию животных России. Основу системы составляет таксономический классификатор (специализированная база данных), отражающий систематическое положение любого из таксонов, включенных в систему. Традиционный подход с применением СУБД FoxPro претерпевает изменения, и система в последнее время переводится на более продвинутую платформу MS SQL Server.

Кроме чисто поискового значения и накопления разнообразной информации, разрабатываемая система может использоваться в качестве простейшего определителя животных.

1 Введение

В России в начале 2000-х годов делались активные шаги для интеграции данных по видовому биологическому разнообразию в глобальной сети интернет. Уже более шести лет на веб-портале Зоологического института (ЗИН) РАН размещена информационно-поисковая система (ИПС) «Биоразнообразие России» – БИОДИВ (BIODIV), начальное развитие которой поддерживалось с 2002 по 2004 годы Федеральной целевой научно-технической программой «Исследования и разработки по приоритетным направлениям развития науки и техники» (государственный контракт № 43.073.11.2510). За год страницы ИПС БИОДИВ посещает около 1 миллиона пользователей [14, 12].

Разработка таксономических баз данных (БД) и информационных систем (ИС) началась в ЗИН РАН в 1986 г. [7]. Под таксоном в работе понимается «группа в классификации, состоящая из дискретных объектов, объединяемых на основании общих свойств и признаков» [18], под макротаксоном – группа выше ранга семейства или отряда. Ранги таксонов – это универсальные уровни иерархии, имеющие собственные названия. Всего их используется в биологии более 40. В зоологии обязательными являются: вид, род, семейство, отряд, класс и тип [18].

Любая информация по биоразнообразию в любой ее составляющей части одинаково важна и должна быть одинаково доступна при обращении к ней потенциального пользователя. Под биологическим разнообразием понимают видовое богатство в том или ином сообществе или биогеографическом регионе [19]. Самый простой и логичный способ организации информации – это таксономический классификатор, отражающий систематическое положение любого из включенных таксонов, который состоит из частных классификаторов, связанных между собой. Все классификации – авторские; у разных специалистов они отличаются между собой. Под классификацией понимается процесс группировки объектов исследования или наблюдения в соответствии с их общими признаками. В результате разработанной классификации создаётся классифицированная система (часто называемая так же, как и процесс, – классификацией). Таксономия – теория классификации и систематизации сложноорганизованных областей действительности, имеющих обычно иерархическое строение (органический мир, объекты географии, геологии, языкознания, этнографии и т. п.) [18]. Кроме демонстрации консенсусных вариантов, система BIODIV предусматривает возможность размещения параллельных альтернативных классификаций.

Биологические базы данных опираются на таксономические таблицы, в которых должны полно отражаться сложные многоуровневые иерархические схемы классификации таксонов. Отражение иерархий в плоских таблицах реляционных баз данных является нетривиальной задачей. Несколько способов решения этой проблемы известны в теории информационных систем [17, 15]. В ЗИН РАН большинство этих способов было разработано независимо в ходе многолетних работ по созданию зоо-

Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

логических баз данных [7]. Все эти разработки отражены в стандарте ZOOCOD, многократно описанном в ряде публикаций [4 – 6, 16, 8].

2 Информационно-поисковые системы БИОДИВ и ЗООДИВ

Основной классификатор проекта «БИОДИВ» включает 5 таксономических царств, объединяющих все живые организмы. Это: бактерии, грибы, растения, протисты и животные. Раздел «Бактерии» связан непосредственно с базой данных Всероссийской Коллекции Культур Микроорганизмов (ВКМ).

Для 4 других царств живых организмов специально для проекта БИОДИВ были разработаны и выставлены в интернет следующие макроклассификации:

1) грибы – 8 макротаксонов и ссылки с них на сайты Ботанического института (БИН) РАН и ВКМ;

2) растения – 15 макротаксонов и ссылки с них на сайты БИН РАН;

3) протисты – 29 типов и собственные разделы сайта для каждого из них на сайтах БИН РАН и ЗИН РАН (оригинальная классификация С.А. Карпова (С.-Петербургский государственный университет) и его коллег);

4) животные – совместно со специалистами ЗИН РАН В.В. Малаховым (Московский государственный университет) разработана новая оригинальная классификация животных – 227 таксонов, включающая 35 типов и 150 классов. Для каждого типа создан собственный иллюстрированный раздел (некоторые разделы доведены полностью или частично до видов).

Для развития только одной ветви классификатора, посвященной царству животных (фауна), в ЗИН РАН задействовано более 50 отдельных баз данных, выполненных в формате ZOOCOD, реализованном в СУБД FoxPro [8]. Доведенные до семейств и далее до видов, модельные таксоны (наиболее полно даны классификации пауков, насекомых, рыб, птиц и млекопитающих) имеют разные назначение, объемы информации, географические границы (Мир, СНГ, Россия) и иллюстративное сопровождение. Под модельными таксонами понимаются группы животных, представляющие наибольший экономический и экологический интерес, на которых отрабатываются вопросы, связанные с быстродействием системы, способами отображения разнообразной информации и т. п. В дальнейшем предусматривается разработка системы фильтров, позволяющих манипулировать с базами данных таким образом, чтобы отсортировывать демонстрируемые таксоны по определенным заданиям или наборам признаков. Пользователь, таким образом, сможет отсортировать, например, из классификации мировой фауны список таксонов фауны России или другого региона.

Для некоторых таксонов от основных классификаторов BIODIV сделаны ссылки на классификаторы других информационных систем. Так, классификации некоторых отрядов насекомых, сопровождаемые коллекционными данными, оригинально располагаются на страницах проекта ZInsecta, но открываются и со страниц ИС «Биоразнообразие России» [3].

В итоге созданы 32 таблицы баз данных, содержащие сведения о 45 тысячах таксонов.

Эта информация доступна на 423 веб-страницах портала ЗИН (типы HTML- и ASP-страниц) [9].

Страницы проиллюстрированы – на них демонстрируется 1700 фотографий и рисунков. Создан специальный фотоальбом, включающий галереи художественных снимков животных и растений разных авторов.

Общий объем сайта BIODIV в 2007 г. составил 61 мегабайт.

2.1 Проблемы разработки ИПС по биоразнообразию

Традиционный подход с применением СУБД FoxPro для разработки ИПС по биоразнообразию в последнее время стал создавать ряд проблем: отсутствие развитых средств интеграции в интернете, трудности многопользовательского доступа, производительности, масштабируемости и расширяемости; архаичная структура таблиц: синонимичные таксоны располагаются строго под валидным таксоном; необходимость использования расчетного поля HIERCOD для удобства отображения многоуровневой иерархии в форме таксономического древа; программирование в среде FoxPro: «проблема одного разработчика», закрытая система; проблемы совместимости баз данных FoxPro с современными версиями управляющих элементов и программных компонентов; проблема кодировок.

Синонимы в биологической таксономии – два или более названия, относящиеся к одному и тому же биологическому таксону. Только один из всех синонимов может быть названием, под которым данный таксон должен быть известен. Этот таксон называется валидным. Обычно это тот синоним, который был обнародован раньше других [18]. Валидный таксон в зоологии и ботанике – действительный, веско обоснованный, реально существующий или общепризнанный таксон, и его название не подлежит сомнению вследствие достаточно точно выполненного первоначального описания. Невалидными (недействительными), например, считаются таксоны, попавшие в синонимию в качестве младших синонимов, т. е. номинальные виды, роды и т. п., описанные позже под другими названиями, но, по существу, заново описывающие уже известные, таксономически обозначенные формы [20].

Совокупность недостатков СУБД FoxPro подвигла обратиться к более совершенному программному продукту – СУБД MS SQL Server, который характеризуют следующие ключевые особенности: клиент-серверная СУБД корпоративного уровня: многопользовательский доступ с разделяемыми правами, высокая производительность, масштабируемость и расширяемость; индустриальный стан-

дарт структуры данных, построения запросов, средств импорта и экспорта данных; поддержка производителя и совместимость с новыми версиями программных компонентов, поддержка юникода; мощные встроенные средства программирования, расчета значений полей и автоматического заполнения полей, проверки и обеспечения целостности данных; средства централизованного управления элементами баз данных, пользователями и правами доступа; развитые средства обеспечения безопасности и резервного копирования.

Ограничение финансирования заставило сосредоточиться на создании ИПС только для животных России и сопредельных территорий. С 2006 года начался проект ЗООДИВ (ZOODIV) («Биоразнообразие животных России»). От проекта BIODIV новый проект ZOODIV наследует не одно царство животных (Animalia), а еще и царство (скорее, даже группу еще более высокого ранга) протистов (Protista). Это вызвано тем, что самые неясные вопросы макроклассификации живых организмов лежат на самом высоком таксономическом уровне и споры о числе царств и о взаимоотношениях протистов с грибами и животными до сих пор не прекращаются, поэтому было решено пока оставить «простейших» в рамках проекта [10]. Разработки Ботанического и других институтов получили возможность самостоятельного развития вне рамок существовавшего проекта BIODIV [9].

Преимущества, которые просматриваются в рамках проекта ZOODIV, таковы: клиент-серверная информационная система; разделение данных (таблицы в СУБД – серверная часть) и представления данных (пользовательский интерфейс – клиентская часть); использование единого хранилища унифицированных таксономических данных с разделяемым доступом для различных задач и проектов, как внутри сети ЗИН РАН, так и в публичном доступе на веб-портале; импорт имеющихся разрозненных таксономических данных в единый классификатор, предоставление широкому кругу специалистов средств для внесения исправлений и дополнений, создание удобного пользовательского интерфейса для заполнения классификатора новыми данными.

Реализация наработок в рамках проекта ZOODIV привела к возникновению новых проблем и постановке новых задач, диктуемых длительностью разработки структуры системы и алгоритмов обработки данных; ограниченностью ресурсов у разработчиков; специфичностью таксономических данных и методов работы с ними; невозможностью прямого использования готовых технических решений и привлечения разработок сторонних специалистов. Огромные объемы данных обусловили необходимость проведения тестов для оценки производительности и масштабируемости реализуемых алгоритмов; создание открытой системы предопределило использование стандартных средств программирования – MS SQL Server (серверная часть) и универсального веб-интерфейса (клиентская часть – технология ASP; кросс-платформенный скриптовый язык JavaScript, принципиально одинаковый в «серверной» и «клиентской» частях кода ASP-страниц) [13].

2.2 Перспективы разработки ИПС «ЗООДИВ»

Кроме чисто поискового значения, подобные системы могут представлять интерес и с точки зрения их использования в качестве простейших определителей, а затем, после накопления соответствующих иллюстраций, и для создания интерактивных определителей биологических объектов [2]. Эти определители, связанные с остальными базами данных, позволят осуществлять быстрый и достоверный поиск объектов и эффективное использование всех сведений, распределенных в различных базах. В настоящее время отлаживается программное обеспечение определителей и разрабатываются алгоритмы, которые обеспечивали бы оптимальную сортировку признаков для групп с большим числом видов, в том числе и алгоритм обучения, дающий возможность обратного распространения ошибки [1].

В настоящее время общее количество записей в классификаторе проекта ZooDiv составляет 108851 (из них синонимов – 23428). Используемых таксономических рангов – 40, при этом количество таксонов по основным рангам следующее: Phylum (Тип) – 64 (синонимов – 1), Classis (Класс) – 215 (синонимов – 10), Familia (Семейство) – 3609 (синонимов – 396), Genus (Род) – 16238 (синонимов – 3279), Species (Вид) – 79366 (синонимов – 17768) [11].

В дальнейшем предполагается обеспечить весь имеющийся массив данных адекватной поисковой системой и связать его с поисковыми системами интернета, что должно облегчить работу по наполнению главных ветвей классификатора животного мира и проиллюстрировать основные таксоны.

Литература

[1] Вахитов А.Т., Граничин О.Н., Кирейчук А.Г., Лобанов А.Л. Параллельный алгоритм обучения для интерактивного определителя биологических видов// Труды суперкомпьютерной конференции «Научный сервис в сети Интернет: масштабируемость, параллельность, эффективность» (21 – 26 сентября 2009 г., г. Новороссийск). – М.: Изд-во МГУ, 2009. – С. 332-334.

[2] Кирейчук А.Г., Лобанов А.Л., Смирнов И.С., Вахитов А.Т., Воронина Е.П., Пугачев О.Н. Виртуальные коллекции животных и интерактивные определители биологических объектов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Одиннадцатой Всерос. науч. конф. RCDL’2009 (Петрозаводск, Россия, 17 – 21 октября 2009 г.). – Петрозаводск: КарНЦ РАН, 2009. – С. 400-407.

[3] Кривохатский В.А., Лобанов А.Л., Медведев Г.С., Белокобыльский С.А., Дианов М.Б., Смирнов И.С., Халиков Р.Г. Информационная систе-

ма по энтомологическим коллекциям в Интернете // Труды Русского энтомологического общества. – 2003. – Т. 74. – С. 59-70.
[4] Лобанов А.Л., Зайцев М.В. Создание компьютерных баз данных по систематике млекопитающих на основе классификатора названий животных ZOOCOD // Вопросы систематики, фаунистики и палеонтологии мелких млекопитающих (Труды Зоологического института РАН). – 1991. – Т. 243. – С. 180-198.
[5] Лобанов А.Л., Смирнов И.С. Принципы построения и использования классификаторов животных в стандарте ZOOCOD // Базы данных и компьютерная графика в зоологических исследованиях (Труды Зоологического института РАН). – 1997. – Т. 269. – С. 66-75.
[6] Лобанов А.Л., Смирнов И.С., Дианов М.Б. ZOOCOD – концепция представления зоологических иерархических классификаций в реляционных базах данных // Информационно-поисковые системы в зоологии и ботанике (Тезисы межд. симпозиума, май 1999). Труды Зоологического института РАН. – 1999. – Т. 278. – С. 66.
[7] Лобанов А.Л., Смирнов И.С. Место и роль информационных технологий в исследованиях Зоологического института РАН // Фундаментальные зоологические исследования: теория и методы (по материалам Международной конференции «Юбилейные чтения, посвященные 170-летию Зоологического института РАН», 23 – 25 октября 2002 г.). – М.-СПб: Товарищество научных изданий КМК, 2004. – С. 283-318.
[8] Лобанов А.Л., Смирнов И.С., Дианов М.Б., Голиков А.А., Халиков Р.Г.. Эволюция стандарта ZOOCOD – концепции отражения зоологических иерархических классификаций в плоских таблицах реляционных баз данных // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всерос. науч. конф. RCDL'2008 (Дубна, Россия, 7 – 11 октября 2008 г.). – Дубна: ОИЯИ, 2008. – С. 326-332.
[9] Портал Зоологического института. Проект BioDiv. – http://www.zin.ru/BioDiv/bd_part.htm.
[10] Портал Зоологического института. Проект ZooDiv. – http://www.zin.ru/ZooDiv/Project.htm.
[11] Портал Зоологического института. Проект ZooDiv, статистика классификатора. – http://www.zin.ru/ZooDiv/animals_stats.asp.
[12] Пугачев О.Н., Алимов А.Ф., Лобанов А.Л., Кривохатский В.А., Смирнов И.С. Первые итоги разработки информационной системы по биоразнообразию России (BIODIV – ZOODIV) // Информационные системы и web-порталы по разнообразию видов и экосистем. Материалы межд. симпозиума, Борок, 28 ноября – 1 декабря 2006 г. – М.: Тов-во науч. изд. КМК, 2006. – С. 170-173.
[13] Пугачев О.Н., Дианов М.Б., Лобанов А.Л., Смирнов И.С., Халиков Р.Г., Голиков А.А. Итоги разработки проекта «Информационная система по биоразнообразию животных России» (ZooDiv) // Отчетная научная сессия по итогам работ 2007 г. Тезисы докладов. 8 – 10 апреля 2008. –Зоологический институт РАН, 2008. – С. 42-44.
[14] Смирнов И.С., Лобанов А.Л., Алимов А.Ф., Пугачев О.Н., Кривохатский В.А.. Информационная система по биологическому разнообразию России // Научный сервис в сети Интернет: Труды Всерос. науч. конф. (22 – 27 сентября 2003 г., г. Новороссийск). – М.: Изд-во МГУ, 2003. – С. 12-14.
[15] Щеваев П.А. Способы хранения иерархических структур в реляционных базах данных // Новые информационные технологии и системы: Труды VI Межд. науч.-техн. конф. Ч. 2. – Пенза: ПГУ, 2004. – С. 226-233.
[16] Lobanov A.L., Ryss A.Yu., Smirnov I.S.. A modern state of the ZOOCOD concept // Information Systems on Biodiversity of Species & Ecosystems. Scientific program & Abstracts. – SPb, 2003. – P. 7-8.
[17] VanTulder G. Storing hierarchical data in a DB. – http://www.sitepoint.com/print/hierarchical-data-database, 2003.
[18] Статья «Таксон» – http://ru.wikipedia.org/wiki/.
[19] Дедю И.И. Экологический энциклопедический словарь. – Кишинёв: Гл. ред. МСЭ, 1990. – 408 с.
[20] Статья «Валидный вид». – http://flores.by.ru/notes.html.

Results and prospects of development of information system on a biodiversity of animals of Russia (ZOODIV – BIODIV)

I.S. Smirnov, O.N. Pugachev, A.G. Kirejchuk, M.B. Dianov, A.L. Lobanov, R.G. Khalikov, A.A. Golikov, V.A. Krivohatsky

More than six years at the Zoological institute of the Russian Academy of Science are developed an information retrieval system (IRS) on a biological diversity of animals of Russia. A basis of IRS is taxonomic classificator (a specialized database), reflecting regular position of any included in system taxa. The traditional approach with application DBMS FoxPro undergoes changes and the system recently is translated on more advanced platform MS SQL Server. Except for cleanly search value and accumulation of the various information, the developed system can be used as the elementary identification system for animals.

♣ Работа выполнена при частичной финансовой поддержке РФФИ (проект 09-04-00789), подпрограммы «Изучение и исследование Антарктики» Федеральной целевой программы «Мировой океан», проекта № 4 «Определение состояния антарктических экосистем, оценка окружающей среды в районе работ Российской Антарктической Экспедиции», и программы «Биоразнообразие»