Концепция виртуального биологического музея и GRID: от склада данных к распределенному ресурсу

© Тирас Х.П., Ильясов Э.Ф., Соболев С.И.*, Воеводин Вл.В.*

Институт теоретической и экспериментальной биофизики РАН,
Пущино, Московская обл., 142290, tiras@iteb.ru,
*НИВЦ МГУ, Москва, 119992, Москва, Ленинские горы, МГУ,
НИВЦ, voevodin@parallel.ru

Аннотация

Представлена концепция Виртуального биологического музея (ВБМ, http://www.iteb.ru/biomuseum) как сетевого распределенного ресурса, создаваемого в соответствие с методологией и технологиями GRID. В рамках данной концепции ВБМ предстает как распределенная коллекция качественных электронных изображений живых («плоских») биологических объектов сделанных по единому стандарту. Обеспечивается одновременный доступ ко многим локальным коллекциям для анализа изображений с помощью оригинальных пакетов программ, посылаемых на сервер музея пользователями сети. ВБМ проектируется как единая точка доступа, поиска и обработки данных.

Возможности GRID обеспечивают решение ряда фундаментальных и прикладных биологических задач (систематика, экологический мониторинг).

1. Виртуальные биологические коллекции: введение в работу с живыми объектами

Проблема создания виртуальных биологических имиджевых коллекций имеет короткую историю, однако, логика формирования и развития концепции таких музеев весьма поучительны на фоне стремительного развития сетевых технологий [10-13].

Первоначальная модель виртуального биологического музея (ВБМ), которая рассчитывалась на возможности «традиционного» интернета, видоизменяется с появлением GRIDa - сети нового поколения, обеспечивающей новые возможности и открывающей новые перспективы для виртуальных биологических ресурсов [2, 13].

Структура электронной коллекции (музея) должна обеспечить выполнение трех основных задач любого музея: ***создание*** образца, его ***сохранение (и копирование)*** и ***описание (анализ).*** На начальной стадии виртуальный музей рассматривался скорее как электронная версия традиционного музея, смысл которого был в сборе как можно большего числа образцов «в одной точке». Принципиальным отличием компьютерного музея, от обычного, была его большая компактность, технологичность хранения и воспроизведения образцов и коллекционирование изображений ***живых*** биологических объектов [1, 10, 12].

2. «Традиционные» электронные биологические информационные ресурсы

Привычные биологические информационные ресурсы - базы данных первичных структур нуклеиновых кислот и белков и разнообразные информационные базы данных по биоразнообразию – характеризуются большим количеством отдельных файлов относительно небольшого объема. Однако, по существу, это базы данных, содержащие текстовую информацию.

В их число входят ***универсальные БД***, которые содержат сведения по номенклатуре, систематике, типовым материалам, географическому распространению и симбиотическим связям организмов. К ним относятся системы TROPICOS (Миссурийский Ботанический сад), ZOOINT (Зоологический институт РАН), ALICE (ILDIS: международный проект по бобовым Intemational Legume Database and Information Service).

БД по номенклатуре используют подходы отражения в БД альтернативных таксономических классификаций. В ЗИН РАН разработан классификатор ZOOCOD из 46 уровней иерархии, в котором все используемые в систематике названия таксонов получают двух- или трех буквенный код, сохраняющийся при изменениях в номенклатуре [5-7]. Наиболее известны: БД STATUS на CD-ROM, включающая все названия покрытосеменных, зарегистрированные Index Kewensis, а также БД по сводке С.К. Черепанова "Сосудистые растения СССР" [3], созданная в системе TROPICOS.

Труды 5ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003, Санкт-Петербург, Россия, 2003.

Дескриптивные БД создаются не только с целью составления унифицированных описаний, но и ключей для определения и составления классификаций объектов и предполагаемых схем их филогении. Наиболее известны системы DELTA (Австралия, [15]) и Linnaeus-II (ETI, Голландия, [18]), ориентированная на создание электронных монографий на CD-ROM.

Перевод учета коллекций в целом на компьютерную базу пока не представляется возможным в связи с огромным объемом накопленных материалов. Более реально создание БД типовых материалов [9]. Многие БД типовых коллекций доступны через Интернет: БД типовых ботанических коллекций Нидерландов (L, WAQ), различных коллекций (палеонтологических, по млекопитающим, типовых образцов растений) Стокгольмского Музея Естественной Истории [16, 19], гербарных коллекций Гарвардского университета, Калифорнийского университета и ряда других американских гербариев, Ботанического музея в Копенгагене и т.д. Из отечественных гербариев только каталог типов MW (гербарий МГУ) доступен через сервер FLORIN. Создается компьютерный каталог типов Гербария БИН РАН [8] в среде TROPICOS, гербария ВИР [14], гербария БГУ (Минск) [4].

В отличие от них, виртуальные имиджевые коллекции характерны большим объемом отдельных файлов, поскольку первичным объектом такой коллекции является электронное изображение (имидж) биологического объекта (растения, животного, микроорганизма, клетки или других структур). Такие высококачественные имиджи биологических объектов имеют объем от десятков мегабайт, что приводит к необходимости оперировать с гигантскими массивами данных. Следует отметить, что до настоящего времени не имеется аналогов Пущинского ВБМ, который изначально проектировался как собрание изображений живых биологических объектов с числом отдельных образцов достаточным для корректного статистического анализа [10-11].

3. Пущинский ВБМ – первый опыт создания виртуальных коллекций.

Надо сказать, что возможности сетевых технологий в первоначальном варианте ВБМ не были задействованы в полной мере. В частности, предполагалось, что различные локальные коллекции будут обмениваться своими ресурсами по мере решения конкретных задач. Это создавало проблему для работы с полноценными изображениями биологических объектов в tiff-формате. Действительно, если обеспечивать обмен такими данными, то неизбежно возникают вопросы с трафиком больших и очень больших массивов данных.

Например, минимальный объем директории с образцами листьев одного вида (липы обыкновенной, ***Tilia cordata***) из одной географической точки – по 50 листьев с 10 деревьев (tiff-формат) Пущинского виртуального биологического музея (пилотная версия веб-сайта музея http://www.iteb.ru/biomuseum) составляет порядка 10 Гб. Поэтому для решения текущих задач первоначально планировалось использование «сжатых» jpg-файлов. Оставалась проблема обмена «тяжелыми» tiff-файлами, которую предполагалось решать через обмен CD-room [11-12]. В первоначальной концепции ВБМ никак не была прописана идея распределения ресурсов, создания сетевого музея не предполагающего концентрацию всей информации в одном месте.

В настоящее время создана пилотная версия Пущинского виртуального биологического музея (ВБМ). Формируется коллекция листьев типичных деревьев средней полосы России: липы, клена и дуба, имеющие разную форму листовой пластинки.

Сканирование проводится с оптическим разрешением 600 dpi, которое принято в качестве эталона. Опыт показал, что для оптимальной цветопередачи, надежности и быстродействия более подходят сканеры AGFA. Коллекция виртуального музея создается на сканере AGFA SnapScan e50.

Первые коллекции ВБМ призваны решать как фундаментальные научные задачи (классификация и систематика растений на базе количественных признаков), так и прикладные научные задачи (разработка новых подходов к экологическому мониторингу).

В 2001 г. были отсканированы листья растений в г. Москва (район Ботанического сада МГУ на Воробьевых горах) и городах Троицк, Пущино, а также в районе Тульских засек (Тульская область, Щекинский район). Эти территории определены как относительно "экологически чистые". В 2002 году получен материал из г.г. Климовска, Серпухова, Тулы, а также из района Варшавского шоссе (г. Москва), которые определены как "экологически грязные" места Москвы и Подмосковья.

Эти географические точки попарно распределяются по двум параметрам: удаленность от Москвы, и роли местных (локальных) факторов экологической нагрузки. Таким парами являются, в Москве (МГУ и Варшавское шоссе), далее Климовск-Троицк, Пущино-Серпухов, Тула-Тульские засеки.

Существенной особенностью прижизненного имиджа биологического объекта является сосредоточение в нем, в скрытом виде, полной информации о данном объекте. Эта информация выявляется в ходе анализа данного имиджа, поэтому разработка программного обеспечения для анализа изображений является критически важным аспектом работы ВБМ. Поэтому одновременно с созданием коллекций изображений, разрабатываются специализированные программы для анализа изображений листьев: программы Leaf (разработчик Седельников З.В) и LeAna (разработчик Деев А.А). Анализ изображений второго основного объекта ВБМ – плоских червей (планарий) проводится с помощью пакета Plana4.0 (разработчик Деев А.А).

Одним из направлений фундаментальной работы ВБМ является определение "пространства разре-

шенных форм" для исследуемых образцов. Липа распространена по всей Европе, поэтому для последующего анализа внутривидового биоразнообразия листьев липы целесообразно получить образцы из других, в том числе краевых зон ее ареала. В 2002 г. начато формирование коллекции листьев липы бассейна р. Волга: собран и отсканирован материал из г.г. Ярославля и Нижнего Новгорода. В 2003 г. начато формирование коллекций западного ареала распространения липы обыкновенной, в частности, из г. Белграда (Сербия и Черногория).

Помимо листьев растений, создаются коллекции изображений плоских червей – планарий бассейна р. Оки. Разработанный временный стандарт создания изображений планарий включает фотосъемку на пленочную камеру NiconF60 с последующим сканированием с помощью слайд-сканера Minolta ScanDual II с разрешением 2820 dpi. Общий объем коллекции изображений планарий ***Jirardia tigrina*** и ***Ijimia tenuis*** составляет более 400 Мб (каждое изображение планарии составляет порядка 4 - 5 Мб).

В настоящее время полный объем коллекции ВБМ составляет более 300 Гб дискового пространства, в том числе 140 Гб включают более 2000 файлов с изображениями листьев липы из «чистых» и «грязных» точек Московской области.

4. Компьютерная биология и биологические информационно-вычислительные ресурсы

Наряду с развитием концепции виртуальных биологических коллекций развивались представления о самой науке, компьютерной биологии, в рамках которой и создавались сами коллекции [13]. Компьютерная биология (vital informatics) рассматривается как новая наука, на границе биологии и информатики. В США это направление науки названо биоинженерией (biomedical engineering) [17]. Здесь акцент сделан на степень (глубину) неинвазивного проникновения в биологический объект: от видеомикроскопии (имиджи поверхности биообъекта) до различных вариантов компьютерной магниторезонансной томографии MRT (от 0.5 до 1.5 мм под поверхность биообъекта в случае fMRT).

Как любая наука, компьютерная биология имеет свои специфические объекты и методы исследования. Она «начинается» с момента ***создания*** электронного изображения ***живого*** биологического объекта, которое составляет предмет данной науки, и «заканчивается» ***анализом*** этого изображения (получением новой информации об объекте), который является специфическим методом этой науки. В ходе анализа изображений биологического объекта создается ***новый тип биологических ресурсов - информационный***, отличительными особенностями которого являются его практическая неисчерпаемость и неистощимость [13].

Говоря о GRID-подобных системах в биологии, необходимо отметить, что для анализа биологической (и, в особенности, медицинской) информации уже накоплен немалый опыт использования распределенных информационно-вычислительных мощностей, родственных GRID-системам. Стоит отметить такие наиболее известные проекты, как:

- Lifemapper (http://www.lifemapper.org/) - проект распределённых вычислений, по составлению мощного электронного атласа биологического разнообразия Земли;
- Distributed Folding (http://www.distributedfolding.org/) - проверка новых алгоритмов моделирования белков;
- Folding@home (http://folding.stanford.edu/) - исследования сворачивания белков с целью поиска лекарственных соединений;
- Find-a-drug.com (http://www.find-a-drug.com/) - поиск лекарств от рака, СПИДа и других болезней;
- MD@home (http://www.md-at-home.ru/) - изучение свойств олигопептидов.

Эти проекты основаны, прежде всего, на добровольном участии пользователей персональных компьютеров по всему миру. Участие в таком проекте предполагает установку на машине пользователя небольшой клиентской программы, которая во время простаивания компьютера через сети Интернет получает от центрального сервера системы порцию данных, обрабатывает ее и возвращает на сервер результат. Подобные системы потенциально обладают огромной вычислительной мощностью, однако они не предназначены для распределенного хранения и доступа к большим объемам данных.

5. ВБМ и GRID-технологии: новые возможности и перспективы

Развитие идеи GRIDa, как сети нового поколения, на первый взгляд, создало новую ситуацию, в том числе, и для виртуальных биологических коллекций. Появилась возможность для обмена полноценными изображениями биологических объектов, что обеспечивают решение задачи оперативного обмена гигантскими массивами имиджевой информацией [13].

Однако подобное, экстенсивное, развитие проекта не могло удовлетворять реальным задачам ВБМ. Двухлетний опыт работы Пущинского ВБМ показал, что такие коллекции с самого начала работы требуют значительного дискового пространства, что делает практически невозможным концентрацию в одном месте всех создаваемых образцов.

С появлением и развитием идеи и технологий GRID пришло оптимальное решение: в новой редакции ВБМ рассматривается как сетевой, распределенный ресурс. Каждый участник проекта создает свой сегмент по единым технологическим правилам (стандартам), а программные ресурсы GRIDa призваны обеспечить их эффективное взаимодействие.

Доступ к GRID в обозримом будущем, очевидно, будет ограничен кругом профессиональных пользователей. Однако, пользователей биологических ресурсов заведомо больше: помимо биологов-

профессионалов к ним потенциально относятся массы школьников и студентов, для которых ВБМ станет повседневным образовательным ресурсом. Более того, силами тех же студентов, предполагается решать проблему развития локальных коллекций ВБМ. Следовательно, надо думать о гибридном варианте музея: «тяжелые» (tiff-файлы) в GRIDe для профессионалов, и «легкие» (jpg-файлы) – для массового пользователя на базе обычных возможностей Интернета [12].

Распределенная структура музея обеспечивает новые возможности для ВБМ. К их числу относится возможность быстрого, в том числе одновременного, доступа ко многим локальным коллекциям для анализа изображений с помощью оригинальных пакетов программ, посылаемых на сервер музея пользователями сети.

При переходе от традиционного, к распределенному музею меняются базовые принципы выполнения его основных функций. Например, если раньше запрос к музею типа «выдай мне для обработки все образцы данного вида» приводил лишь к локальному поиску и работе с локально расположенными данными, то теперь искомая выборка может формироваться из географически распределенных архивов. Хранящиеся в разных архивах данные могут быть продублированы, поэтому появляется задача оптимизации формирования выборки. Данные, представленные в различных архивах, могут отличаться по спектральным параметрам использованных сканеров, следовательно, встает задача оптимизации и приведения к общему знаменателю исходных изображений.

Новая структура музея позволит выполнить новые, доселе недоступные функции, выполняемые ***одновременно*** по всем архивам музея. Скажем, поиск среди образцов по фрагменту изображения можно запустить одновременно на всех серверах распределенного виртуального музея, передав туда исходный фрагмент. Эта возможность будет по достоинству оценена систематиками и палеобиологами, которые зачастую обладают фрагментами того или иного образца и перед которыми стоит задача «воссоздания» исходного образца для последующего его описания и идентификации.

Существенным достоинством сетевой концепции музея является возможность применения ее организационных принципов для совершенствования работы локального музея (как независимого ни от кого, так и части общего). Например, введение нескольких локальных хранилищ приведет к ускорению поиска, а дублирование данных – к повышению надежности музея и доступности данных. Выполнение интеллектуальной обработки хранящегося материала можно организовать в самых различных вариантах, например:

- программа пользователя формирует запрос к распределенному архиву и не заботится о реальном расположении данных. Доступ унифицирован и для программиста локальные и удаленные данные «видны» одинаково. Реальный просмотр всего распределенного архива – это дело системы;
- в определенных случаях система сможет выбирать сама, когда выгоднее выполнять запрос непосредственно на месте хранения удаленных данных, передавая туда код программы, а в конце забирая результат, а когда выгоднее скачивать входные данные на место централизованной обработки;
- можно создавать порталы специализированной обработки. Предположим, что есть востребованная многими задача. Пусть задача определяется фиксированным набором параметров. Реализуем задачу так, чтобы она вместе с системой сама управляла и поиском, и выборкой данных, обработкой, сбором результатов и т.п. на основе заданных пользователем параметров;
- в ряде случаев, система может задействовать отдельные программные продукты для выполнения определенных заказов по особому протоколу. Часто для выполнения разовой работы требуется особое, дорогое программное обеспечение. Пользователю будет предложен спектр программ, имеющихся на сервере, которые он может не приобретать, а «арендовать» для выполнения данной работы.

В таком случае резко вырастает привлекательность сетевого ресурса для массового пользователя. Более того, ВБМ позволит изменить привычный стиль работы с "тяжелыми" биологическими данными, сделав ее гораздо более эффективной.

Традиционный метод работы с сетевыми ресурсами заключается, как правило, в том, что сначала осуществляется поиск данных различными способами, а затем эти данные (возможно, вместе с программных обеспечением для их обработки) скачиваются пользователем на локальную машину.

ВБМ же сосредотачивает доступ к данным, поиск их и обработку в одном месте, предоставляя, таким образом, единую точку доступа к распределенным биологическим ресурсам.

Сетевой ВБМ впервые обеспечивает оптимальное решение задачи доступа и использования нового вида биологического информационного ресурса. Особенно важным для музейной работы является тот факт, что в ходе работы не повреждается исследуемый объект. Ведь, в отличие от обычной коллекции, любая обработка электронного образца проводится после копирования первичного изображения. В известной степени в компьютерной биологии преодолевается основная проблема экспериментальной науки – принцип дополнительности.

Новое знание об объектах и новые возможности хранения образцов позволяют поставить и решить задачу оптимизации объема коллекций. Так, мере изучения коллекции ВБМ, будет определяться оптимальный объем локальных коллекций листьев разных видов растений. Однако, очевидно, что все локальные коллекции должны иметь солидные (петабайтные) технические ресурсы хранения исход-

ной имиджевой информации для обеспечения надежного (дублированного) режима их хранения.

Такая структура ВБМ представляется оптимальной для решения различных, в том числе прикладных задач. К таковым относится мониторинг состояния окружающей среды, когда по мере пополнения коллекции изображениями листьев из одной географической точки через равные (последовательные) промежутки времени можно будет определить тренд развития ситуации в данном регионе в автоматическом режиме с помощью стандартизованных пакетов программ.

Литература

[1] Алимов А.Ф., Смирнов И.С., Рысс А.Ю., Дианов М.Б., Лобанов А.Л., Голиков А.А. Современные биологические электронные публикации: коллекции, идентификационные системы и базы данных. В «Информационные и телекоммуникационные ресурсы в зоологии и ботанике. СПб, 2001, сс.5-19.

[2] Воеводин В.В., Воеводин Вл.В. Параллельные вычисления. СПб. БХВ-Петербург, 2002. - 608 с.

[3] Гельтман О.Ю. 1997. pcTROPICOS: пять лет применения в гербарии Института ботаники им. Комарова. В «Компьютерные базы данных в ботанических исследованиях». СПб, 1997, сс. 16-18.

[4] Джус М.А., Тихомиров Ю.Н. Компьютерная база данных гербария Белорусского государственного университета В «Компьютерные базы данных в ботанических исследованиях». СПб, 1997, сс. 8-10.

[5] Скарлато О.А., Старобогатов Я.И., Лобанов А.Л., Смирнов И.С. 1994. Базы данных в зоологической систематике и информация о высших таксономических группах животных. Зоологический ж., 73 (12), сс. 100-116.

[6] Лобанов А.Л., Зайцев М.В. 1993. Создание компьютерной базы данных на основе классификатора ZOOCOD Труды Зоологического института РАН, 263, сс. 180-198

[7] Лобанов А.Л., Смирнов И.С. 1997. Принципы классификации животных в стандарте ZOOCOD. Труды Зоологического института РАН, 269, сс. 66-75

[8] Никитин В.В., Бородина-Грабовская А.Е., Новоселова М.С. 1997. Создание компьютерного каталога типовых образцов гербария Института ботаники им. Комарова. В «Компьютерные базы данных в ботанических исследованиях». СПб, 1997, сс. 75-77.

[9] Пименов М.Г. Базы данных в таксономии: современное состояние. В «Информационные и телекоммуникационные ресурсы в зоологии и ботанике. СПб, 2001, сс. 9-15.

[10] Тирас Х.П. Виртуальный биологический музей как зеркало компьютерной революции. Химия и жизнь, 2000, № 11-12, сс.24-29.

[11] Тирас Х.П., Ильясов Э.Ф., Жукова Д.В., Петров А.Б. Виртуальный биологический музей – новое поколение интернет-ресурсов. Труды конф. “Научный сервис в сети Интернет”, М., МГУ, 2001, сс.30-32.

[12] Тирас Х.П., Ильясов Э.Ф., Петров А.Б. Виртуальный биологический музей – первый год работы. Труды конф. РЕЛАРН 2002, Н-Новгород, 2002, сс.123-125.

[13] Тирас Х.П., Рождественская З.Е., Ильясов Э.Ф., Петров А.Б., Майоров С.Р. Компьютерная биология – проблемы и перспективы. В «Горизонты биофизики. От теории к практике». Серпухов, 2003, ред. Г.Р.Иваницкий. сс. 62-66.

[14] Чухина, Лунева Н.Н., Лебедева Е.Г. 1997. Базы данных типовых образцов гербария ВИР В «Компьютерные базы данных в ботанических исследованиях». СПб, 1997, сс.101-103.

[15] Dalwitz, M. J., Paine, T. A., Zurcher, E. J. 1996. User's guide to the DELTA system: A general system for processing taxonomic descriptions. Ed 4.06. Canberra.

[16] Osterdahl, F., 1997. MicroRUBIN, a database system for the management of collections at natural history museums. *Proceedings of the Zoological Institute RAS,* 269, pp. 112-118.

[17] NIH\BECON Biomedical Engineering Symposia. 1998. Final Report. NIH. Bethesda.

[18] Schalk, P. H. & W. Los. 1997. The application of interactive multimedia software in taxonomy and biological diversity studies. *Proceedings of the Zoological Institute RAS,* 269, pp.119-129.

[19] Stengard, E. & P. Kenrick. 1997. Development of a collection databases at the Swedish Museum of Natural History. *Proceedings of the Zoological Institute RAS,* 145, p. 150.

Conception of the virtual biological museum and GRID – from the data store to the distributed resource

Tiras Kh. P., Ilyasov F.E., Sobolev S.I., Voevodin Vl.V.

A novel concept of the Virtual biological museum (VBM) in accordance with GRID-technologies as a distributed web-resource is presented.

Pushchino Virtual Biological Museum (VBM) (http://www.iteb.ru/biomuseum) is a collection of electronic images of “flat” biological objects created in accordance with certain technological standards. GRID-technologies provide the possibilities of free access to the deferent collections provided by users using original software. VBM is designed as a site providing the access, search and image data processing. The GRID-technologies provide new possibilities to solve fundamental and applied biological problems (systematic, ecological monitoring).