УДК 591.5:681.3

© 1993 г. А.Ф. АЛИМОВ, АЛ. ЛОБАНОВ, О.Н. ПУГАЧЕВ

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РЕЛЯЦИОННОГО И СЕТЕВОГО ПОДХОДОВ К СОЗДАНИЮ БАНКОВ ДАННЫХ ПО СИСТЕМАТИКЕ, ЭКОЛОГИИ И ГЕОГРАФИЧЕСКОМУ РАСПРОСТРАНЕНИЮ ЖИВОТНЫХ

Рассматриваются вопросы планирования создания банков и баз данных по систематике, экологии и географическому распространению животных. Проводится сравнительный анализ сетевого и реляционного подходов к наиболее полному отражению иерархической классификации. Предлагается рассматривать вид в качестве единицы хранения информации, а систему иерархических отношений единиц хранения — в качестве структурной основы.

Биологическое разнообразие представляет собой наиболее информативный показатель состояния экосистем, оно отражает сложность их структуры и тесным образом связано с их функциональными характеристиками. Это позволило сформулировать структурно-функциональный подход к изучению природных сообществ (Алимов, ,1982), который в последнее время все более активно используется в экологических исследованиях. Отсюда становится очевидным необходимость получения и активного использования информации о многообразии видов.

К настоящему времени сложились необходимые предпосылки для широкого внедрения в биологические фундаментальные исследования компьютерных методов обработки данных, позволяющих надежно хранить, эффективно анализировать и легко передавать разнообразную информацию о многочисленных видах животных (Скарлато и др., 1989). Среди таких предпосылок наиболее значимо то, что в различных институтах и учреждениях биологического профиля накоплен огромный объем информации и появилась вычислительная техника с параметрами, адекватными этому объему. Наконец, среди властных структур постепенно начинается проявляться понимание значимости биологической информации для сохранения окружающей среды и оценки ее состояния. Однако пока такое понимание чаще всего лишь декларируется и не получает необходимой финансовой и материальной поддержки.

В Зоологическом институте РАН накоплена разнообразная информация об огромном числе видов животных, которая хранится в виде коллекционных материалов, так и в виде знаний специалистов-зоологов по отдельным группам животных в объеме мировой фауны. Необходимость перехода на компьютерные методы хранения и анализа этой информации становилась все более очевидной уже давно. В настоящее время Компьютерный совет института разработал программу компьютеризации зоологических исследований, включающую несколько этапов. Первый этап предусматривает создание отдельных баз и банков данных[1], что требует разработки их структур, ввода больших массивов данных и разработки специфических программных средств; второй — создание более сложных комплексных информационно-поисковых систем; третий — создание экспертных систем для оценки состояния биоценозов и влияния антропогенных факторов. При этом учитывалось, что первый этап — самый длительный, требующий значительных капитальных вложений. Разработка зоологических

[1]Под базой данных мы понимаем один специализированный файл данных, а под банком данных — совокупность нескольких взаимосвязанных баз данных и прикладных программ, предназначенных для работы с ними.

баз и банков данных ведется на основе следующих положений: 1) единицей хранения информации является вид; 2) основой зоологических банков данных должна быть система иерархических отношений единиц хранения, т.е. видов (формализованное представление существующих классификаций животных); 3) программное обеспечение должно создавать гибкие возможности постоянного внесения изменений в эту систему иерархии.

Впервые задача полного отражения иерархической классификации на основе реляционного подхода была решена в названном институте путем разработки оригинального классификатора и специального программного обеспечения (Лобанов, 1986; Лобанов, Сергеев, 1986). В этом классификаторе название каждого таксона имеет два кода — цифровой и буквенный. Цифровые коды являются изменяемой частью классификатора и отражают представление о принятой системе группы животных (или о нескольких альтернативных системах). Все валидные (действительные) названия имеют разные цифровые коды. Они должны быть присвоены так, чтобы упорядочение таксонов по этим кодам давало систематический список таксонов. С точки зрения специалиста, по теории баз данных такой порядок цифровых кодов соответствует левосписковому описанию (Глушков, 1982) дерева классификации таксонов. Синонимы и непригодные названия имеют одинаковые цифровые коды с соответствующими им валидными названиями. В базы данных (БД) цифровые коды не вносятся. Буквенный мнемонический код (акроним) служит для сжатия информации при хранении на машинных носителях и для сокращения объема вводимой в ЭВМ информации при создании БД и формирования запросов. Акроним применяется во всех БД, использующих классификатор вместо полного названия. Акроним образуется обычно из первых букв латинского названия таксона и присваивается конкретному таксону раз и навсегда. Он не изменяется даже при сведении названия в синонимы или при переводе его в непригодные. Это правило устраняет необходимость каких-либо манипуляций с содержимым многочисленных БД при изменениях в номенклатуре и систематике.

Важное преимущество такого классификатора перед известными ранее заключается в том, что возможность представления иерархии таксонов с любой детальностью, требующейся зоологу-систематику, реализована в нем полно и последовательно. При этом предельно облегчена процедура внесения изменений в иерархию и обеспечено экономное представление нескольких альтернативных систем таксонов. Специально разработанные программные средства обеспечивают выдачу из БД правильных ответов на запросы с учетом последних данных по синонимии и классификации.

В процессе работы над созданием коллекционных БД пришло понимание того, что большие объемы информации и все возрастающая сложность зоологических банков в ряду коллекционные — фаунистические эколого-фаунистические могут превысить возможность широко распространенных реляционных систем управления базами данных (СУБД). Это обстоятельство заставило обратить внимание на другой класс СУБД — постреляционные сетевые. Опыт разработки и эксплуатации банков данных и поисковых систем в одном институте позволяет нам провести сравнительный анализ реляционного и сетевого подходов.

Существуют, как известно, три модели СУБД — реляционная, сетевая и иерархическая. В реляционной модели база данных представляет собой один файл, адекватный одной плоской таблице (со связями 1:1). В сетевой постреляционной модели она представляет собой совокупность объектов, связанных с разными типами связей от 1:*n* и 1:1, до связей типа *m*:*n* и рекурсивных, что позволяет поддерживать и иерархию их отношений. В иерархической системе моделируется граф в виде дерева и содержатся только связи 1:*n*. История применения и соперничества этих типов СУБД довольно сложна, но сводится в основном к соревнованию двух первых моделей. Именно им удавалось в отдельные периоды и для некоторых классов ЭВМ удерживать пальму первенства. Мы не ставим задачу освещения здесь этой истории, обратим внимание только на три момента.

Пик популярности сетевой модели КОДАСИЛ на больших ЭВМ типа IBM 360/370 пришелся на период их преобладания на рынке (в СССР это машины ЕС серий Ряд-1 и Ряд-2). Затем наступил период повального увлечения реляционными СУБД на персональных компьютерах. При этом во главу угла ставился дружественный интерфейс для рядового пользователя, внешняя же память компьютера ограничивалась единицами или несколькими десятками мегабайт. Наконец, в современный период становится обычным наличие жесткого диска объемом в 100—600 мегабайт, а быстро входящие в употребление оптические диски поднимают эти значения до тысяч мегабайт, т.е. до гигабайт. Пользователи персональных ЭВМ теперь часто создают БД таких объемов, при которых уже невозможно повысить производительность системы при помощи индексирования файлов и тщательной декомпозиции отношений. И хотя реляционные СУБД еще занимают львиную долю рынка, профессиональные программисты все чаще обращаются к сетевой модели, которая в складывающихся условиях может занять довольно обширную нишу для своего применения.

Каковы возможности и перспективы обеих моделей в создании банков зоологических данных? Лучшим ответом было бы практическое сравнение реляционной и сетевой моделей, проведенное на основе тестирования нескольких банков данных разного объема и структур, реализованных параллельно в рамках того и другого подхода. Пока у нас такой возможности нет. Поэтому наш анализ носит несколько умозрительный характер и основывается на рассмотрении кажущихся нам наиболее важными отдельных аспектов создания и эксплуатации зоологических БД. Реляционная модель рассматривается главным образом на примерах СУБД семейства d BASE (d BASE III +, d BASE IV, FoxBASE +, FoxPro, Clipper), а также систем R:base и Paradox. Сетевая модель в нашем анализе представлена системой MDBS.

Важным преимуществом реляционных СУБД является доступность для биологов наиболее популярных программ этого типа, которые позволяют создать базу данных и начать ее наполнение через 15—20 мин после знакомства с системой, а затем эффективно выполнять многие виды работ с одной или несколькими связанными БД практически без программирования, средствами только пользовательского интерфейса. Поскольку сетевые СУБД такими свойствами не обладают, то сказанное выше еще много лет будет оставаться решающим фактором при выборе СУБД в пользу реляционной модели в тех случаях, когда мы имеем дело с простыми по структуре банками данных сравнительно небольшого объема. Как справедливо отметил один из компьютерных обозревателей (Когаловский, 1990), подавляющее преобладание на рынке реляционных СУБД делает выбор модели системы соответственно характеру предметной области практически предрешенным уже на первых этапах — это почти всегда реляционная система. Не говоря уже о числе установленных систем, а только перебирая названия, мы можем убедиться — реляционных систем многие десятки, а среди сетевых на персональных ЭВМ пока фигурируют только db-Vista и MDBS.

Трудоемкость создания действующего банка данных тоже пока остается фактором, определяющим выбор реляционной модели. Семейство d BASE сейчас не имеет себе равных по числу редакторов экранных форм и форм отчетов, генераторов меню и прикладных программ, систем графического отображения и анализе информации, а также средств высокого уровня для разработки приложений, которые позволяют создавать сложные прикладные системы практически без программирования. Немалую роль имеет обеспеченность литературой. По семейству d BASE на русском языке вышло уже не менее восьми книг. Ни одна другая система не имеет такого числа опубликованных на русском языке руководств. Это несомненно играет роль в том, что в каждом биологическом институте, где имеются персональные компьютеры, есть один или несколько биологов, способных создать внешне вполне профессионально сделанную реляционную систему для небольших объемов данных.

Но в области количественных показателей реляционные системы теряют свое преимущество. Объем файлов БД за счет хранения в этих системах «пустого места» в виде незаполненных полей фиксированной длины несравненно больше, чем в сетевых.

Это становится первой преградой, когда один файл по объему превышает 1 мегабайт и уже не помещается на дискету. Если же учесть, что для эффективной реляционной системы необходимы индексные файлы, а для систем среднего и большого объема необходимо прибегать к кодированию повторяющихся длинных терминов и, следовательно иметь словари (классификаторы), то становится бесспорным, что сетевые системы, в которых нет и дублирования информации, в этом отношении экономичнее реляционных. Этот недостаток реляционных СУБД малозаметен только при незначительных объемах данных: предел для удобной работы в реляционной системе — 2—10 мегабайт.

- После создания и отладки БД любой пользователь хочет, чтобы его запросы выполнялись быстро. И здесь опять положение реляционных систем очень устойчиво. Использование индексов спасает только, если поиск затрагивает одну, две или максимум три взаимосвязанных БД. При увеличении их количества длительность поиска нарастает лавинообразно и начинает не удовлетворять даже искушенного пользователя. Преимущество сетевой модели в этом аспекте бесспорно, так как сопоставление информации происходит в ней не за счет склейки таблиц по общим полям, а посредством физических связей. К недостаткам реляционных систем относится и то, что они обычно не берут на себя определение целесообразности использования индексов для повышения эффективности доступа к данным (в результате в некоторых случаях подключение индекса может, наоборот, замедлить доступ к данным).

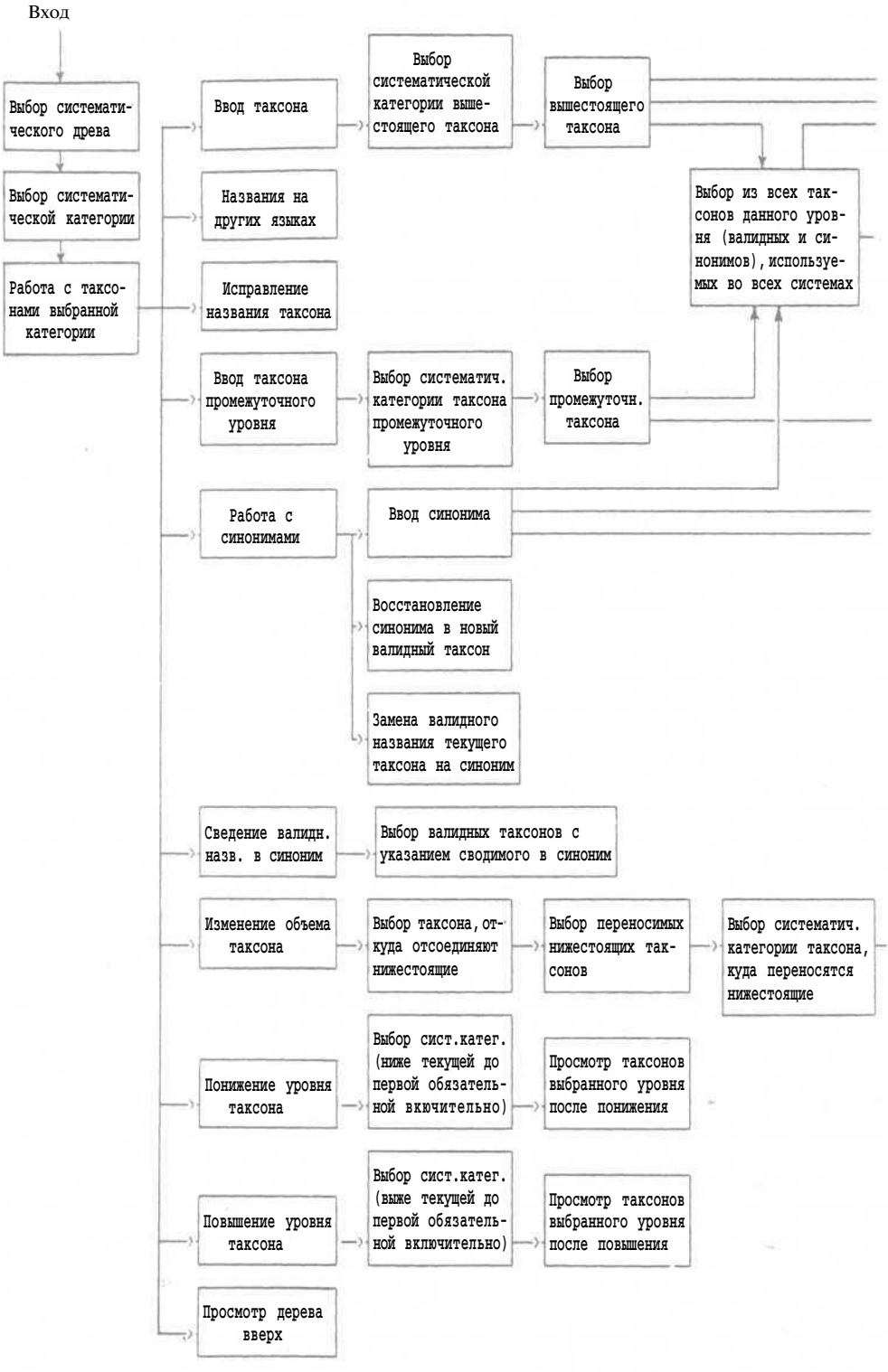
Немаловажным аспектом при сравнении указанных моделей представляется возможность обмена данными с другими системами. В этом отношении более удачной оказывается реляционная система, так как она свободно может обмениваться данными с другими реляционными системами и принимать данные от сетевой системы. Однако передача данных из реляционной системы в сетевую может быть затруднена. Важное значение имеет возможность обмена данными между удаленными и автономными пользователями одного банка данных. В реляционной системе это может быть сведено к передаче одного файла, в сетевой эта возможность очень проблематична.

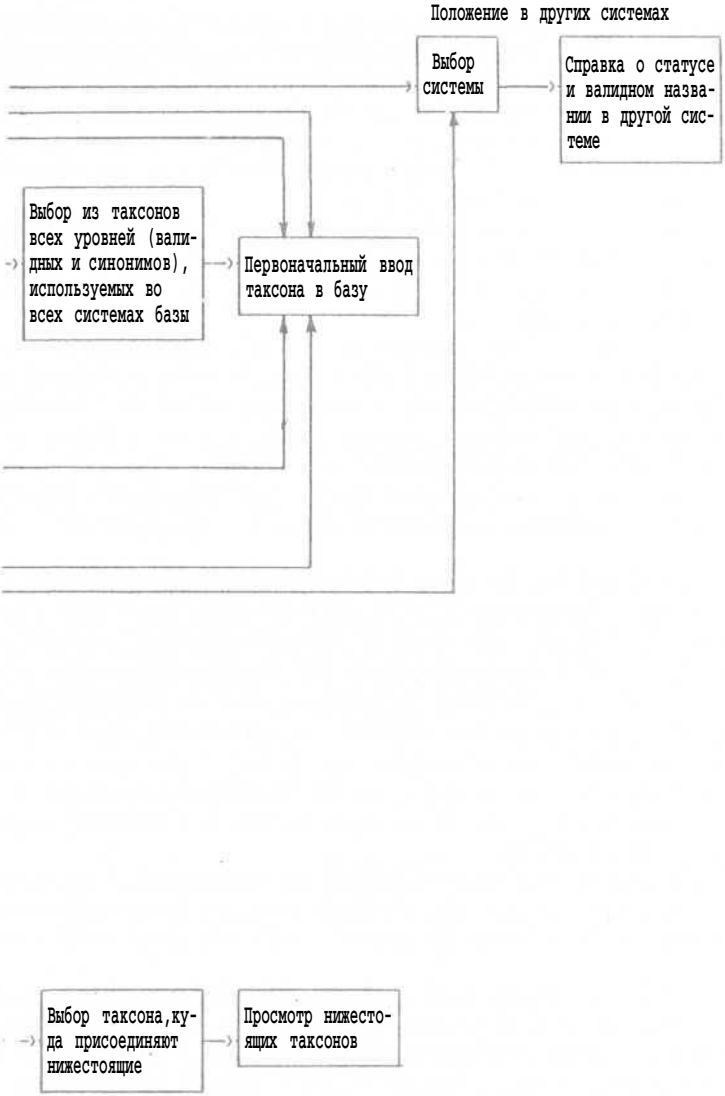
Изменение структуры БД (реструктуризация) легче осуществляется в реляционных системах. В сетевых системах для этого требуется значительно большее количество ресурсов.

Весьма существенно, что графическое представление данных и различные формы анализа, реализованные в специализированных системах (не в СУБД), обычно рассчитаны на прямое использование файлов реляционных СУБД без конвертирования. В сетевых системах для этого требуется создание промежуточного файла.

Таким образом, проведенное сравнение показывает, что обе модели имеют свои достоинства и недостатки. Различная степень сложности и разный объем информации в банках данных предопределяют, на наш взгляд, место реляционных СУБД с дружественным интерфейсом и легким доступом к программным средствам создания банков данных (dDBASE, FoxBASE, Clipper и пр.) и сетевых СУБД (MDBS и др.), поддерживающих при необходимости сложную информационную иерархическую структуру с возможностью быстрой обработки больших объемов информации. Поскольку информационные системы будут создаваться на разных уровнях, преимущества реляционных СУБД станут очевидны при подготовке и верификации блоков информации для баз и банков данных более высокого уровня. Уже на уровне коллекционных баз данных по крупных таксонам живых организмов, не говоря уже об уровне информационно-поисковой системы института и выше, становятся очевидными преимущества сетевых СУБД. На уровне отдельных специалистов, иногда лабораторий, использование разработок в сетевой модели потребует значительных капитальных вложений. Однако решение задач экологического мониторинга на основе банков данных, на наш взгляд, потребует применения именно сетевой модели СУБД.

В результате в Зоологическом институте РАН были избраны два пути развития. Был создан ряд банков данных в рамках реляционной модели с помощью специально организованных классификаторов и использующих его возможности программ на языке dBASE. Эти банки данных, созданные без каких-либо дополнительных затрат


Вход
Выбор систематического древа
Выбор систематической категории
Работа с таксонами выбранной категории
Ввод таксона
Названия на других языках
Исправление названия таксона
Ввод таксона промежуточного уровня
Работа с синонимами
Сведение валидн. назв. в синоним
Изменение объема таксона
Понижение уровня таксона
Повышение уровня таксона
Просмотр дерева вверх
Выбор систематической категории вышестоящего таксона
Выбор систематич. категории таксона промежуточного уровня
Ввод синонима
Восстановление синонима в новый валидный таксон
Замена валидного названия текущего таксона на синоним
Выбор валидных таксонов с указанием сводимого в синоним
Выбор таксона, откуда отсоединяют нижестоящие
Выбор сист.катег. (ниже текущей до первой обязательной вкючительно)
Выбор сист.катег. (выже текущей до первой обязательной включительно)
Выбор вышестоящего таксона
Выбор промежуточн. таксона
Выбор переносимых нижестоящих таксонов
Просмотр таксонов выбранного уровня после понижения
Просмотр таксонов выбранного уровня после повышения
Выбор из всех таксонов данного уровня (валидных и синонимов), используемых во всех системах
Выбор систематич. категории таксона, куда переносятся нижестоящие



Технология работы с разрабатываемым программным средством «Система»

силами сотрудников института, наглядно демонстрируют описанные выше преимущества реляционных СУБД.

Параллельно впервые проведены работы по решению задачи отражения сложной иерархии изменяющейся системы организмов в рамках сетевой модели. Технология работы с такой программой представлена на рисунке. Она обеспечивает возможность ведения на одном компьютере не более десяти вариантов систематического древа. Каждый пользователь имеет право вести (вводить, модифицировать) только свои системы, системы же других пользователей он может только просматривать. В программе запрещено дублирование одного и того же названия в различных системах в пределах одной базы данных. В ней предусмотрена возможность удобного выбора как валидных названий таксонов, так и синонимов от всех таксонов в других системах, ведущихся в одной БД. При этом модификация названия таксона в одной системе не повлечет за собой изменений в других.

Верхние уровни систематического древа (от класса и выше), вариантов которых может быть несколько (2—3), более рационально вести централизованно на уровне института. В каждой лаборатории специалисты по отдельным группам животных будут вести локальные базы данных по их систематике. При этом на одном компьютере можно вести несколько систематических баз данных.

Предусматривается строгое отслеживание семи уровней систематического древа: царство, тип, класс, отряд, семейство, род, вид.

При первоначальном вводе данных невозможно будет завести, к примеру отряд, не связав его с классом и т.д. Предусматривается возможность использования таксонов 38 систематических категорий (от царства до вариетета). Важно, что этот список одинаков для всех лабораторий.

Обеспечивается автоматизация основных операций, выполняемых систематиком: 1) сведение названия таксона в синонимы; 2) восстановление из синонима валидного названия таксона, причем восстановление *nomina nuda* запрещено; 3) описание нового таксона; 4) изменение статуса (повышение или понижение уровня) таксона; 5) перенос таксона, т.е. разделение без изменения статуса, изменение объема путем переноса таксонов нижестоящего уровня из уменьшаемого таксона в новый, объединение без изменения статуса, разделение с изменением статуса, объединение с изменением статуса.

В рамках сетевой модели впервые в ЗИН РАН создана и эксплуатируется в ГосНИОРХ информационно-поисковая система «Болезни рыб». Система позволяет выбрать оптимальную стратегию борьбы с паразитарными заболеваниями в хозяйствах в зависимости от возраста и состояния рыб, технологии выращивания, уровня зараженности, температуры и сезона. Программа позволяет накапливать и искать информацию по жизненным циклам паразитов, по клинической картине заболеваний, о существующих лекарственных препаратах и способах их применения. Недавно закончена разработка (также в среде СУБД MDBS) программы «Дневник паразитолога», позволяющая разрабатывать на ее основе систему мониторинга за состоянием паразитарных систем.

Таким образом, разрабатываемые в ЗИН РАН программы позволяют создавать базы и банки зоологических данных, которые учитывают современное состояние программирования и вычислительной техники, а также предполагаемое их развитие в будущем.

Однако в настоящее время работы по созданию банков данных и программирования практически заморожены из-за отсутствия финансирования. Настало (впрочем, частично уже упущено) время серьезно рассмотреть проблемы капитальных вложений в компьютеризацию биологических исследований в нашей стране. Следует при этом учитывать опыт развитых стран. Там уже давно более половины стоимости программно-технического комплекса (иногда и до 60—70%) приходится на программное обеспечение, а не на компьютеры или другие аппаратные средства. Необходимо вкладывать деньги и в создание программ на основе промышленных СУБД, и во ввод информации, привлекая для этого команды профессиональных программистов. Если мы не хотим безнадежно отстать от уровня мировой науки и потерять еще имеющиеся в систематике животных некоторые приоритеты, вопрос о финансировании такого рода работ должен рассматриваться как один из первостепенных.

СПИСОК ЛИТЕРАТУРЫ

Алимов А.Ф. Структурно-функциональный подход к изучению сообществ водных животных // Экология. 1982. № 3. С. 45—51.

Глушков В.М. Основы безбумажной информатики. М.: Наука, 1982. 552 с.

Когаловский М. Технология баз данных // Компьютер. 1990. № 2. С. 10—14.

Лобанов А.Л. Линейно-иерархическая структура баз данных о таксонах животных // Принципы и методы экоинформатики. М, 1986. С, 293—295.

Лобанов А.Л., Сергеев Г.Е. Проект классификатора названий животных и принцип представления информации об распространении в структуре биологических баз данных // Принципы и методы экоинформатики. М., 1986. С. 214—215.

Скарлато О.А., Алимов А.Ф., Лобанов А.Л., Умнов А.А. Машинные банки данных — подход к кадастру животного мира // Всесоюз. совещ. по проблеме кадастра и учета животного мира. Науч.-информ. матер. к совещ. Уфа, 1989. С. 56—64.

Зоологический институт РАН,
С.-Петербург

Поступила в редакцию
30.VII.1991

COMPARISON OF RELATIONAL AND NETWORK APPROACHES TO CREATING DATA BASES IN THE ANIMAL TAXONOMY, ECOLOGY, AND GEOGRAPHICAL DISTRIBUTION

A.Ph. AUMOV, A.L. LOBANOV, O.N. PUGACHEV

Zoological Institute, Russian Academy of Sciences, Universitetskaia nab. 1, St. Petersburg 199034

Some questions of planning of creating data bases and banks in animal taxonomy, ecology, and geographical distribution are considered. Network and relational approaches to the most complete reflection of hierarchical classification are compared analytically. The species is suggested to be taken as a unit of the relevant information storage, while the system of hierarchical relations of such kind of lots is to be used as a structural frame. Data bases and information retreaval systems created and supported in Zoological Institute in St. Petersburg are briefly described.