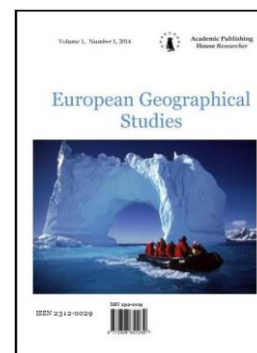


Copyright © 2015 by Academic Publishing House *Researcher*



Published in the Russian Federation
European Geographical Studies
Has been issued since 2014.
ISSN: 2312-0029
E-ISSN: 2413-7197
Vol. 8, Is. 4, pp. 156-167, 2015

DOI: 10.13187/egs.2015.8.156
www.ejournal9.com



UDC 004.9:550.84

Principles of Paleontological and Biological Databases

¹Yuriy V. Agarkov
²Andrey Yu. Agarkov

¹ Sochi Scientific Research Centre, Russian Academy of Sciences, Russian Federation
senior researcher, associate Professor
354000, Sochi, Teatralnaya street, 8a, office 6
E-mail: agarkov2011@yandex.ru

² Paleontological Institute, Russian Academy of Sciences, Russian Federation
PhD student, Laboratory of micropaleontology
346888, Rostov region, Bataysk, Artemovskaya, 55
E-mail: aagarkov77@yandex.ru

Abstract

In this paper we consider the structure of the universal synonymical and alternative-synonymical informational system. The purpose of the system is to store any information in the field of natural sciences. We describe main and associated storage objects, types of the tables for input/output and data structure of the basic dictionary of species, auxiliary dictionaries, bibliographies and blocks of temporal-spatial spread.

Keywords: informational system, biodiversity, radiolarians, diatoms, paleoecology, monitoring of the environment.

Введение

Накопление огромного количества информации по систематике и распространности больших по видовому разнообразию групп современной и ископаемой фауны и флоры приводит к значительным затруднениям в ее использовании, частому повторному открытию видов, создающему дополнительный информационный шум. Для хранения и обработки различных данных по биоразнообразию было создано несколько различных вариантов отечественных и зарубежных информационных систем (ИС) и баз данных [1-3].

Одни из них предназначены для решения конкретных узких задач, другие решают комплексные проблемы информатизации исследований, но ограничены временными или пространственными рамками. В последние годы начата разработка универсальных баз данных по современным и ископаемым организмам [4].

Основным их недостатком, делающим их применение невозможным для хранения разнородной биологической и палеонтологической информации, является отсутствие универсальности структуры при морфологическом, экологическом и других видах анализа в комплексе крупных таксономических групп (класс, тип). Также важна возможность

создания активной, быстро редактируемой синонимии и альтернативной синонимии без отрыва от иных сопутствующих сведений. Жесткая привязка к валидным видам всей атрибутивной информации ведет к отрыву от реальных экземпляров и при изменении синонимии и систематики приводит к потере или искажению данных, требует перепроверки первоисточников с огромными трудозатратами, создает информационный шум и дает неверную информацию о содержании валидных видов, по сути, делает такую информацию практически бессмысленной.

Для устранения всех перечисленных и многих других проблем авторами разработана структура хранения практически всех данных, используемых в естественных науках, и позволяющая ввод информации из первоисточников, не интерпретируя сведения, отмеченные в них. Основой для нее послужила предложенная Ю.В. Агарковым в 1993 году и в дальнейшем постоянно совершенствующаяся многоцелевая информационная система (ИС) по радиоляриям (RADBASE), а затем и другим организмам (BIOBASE) [5-7].

Описание базы данных

База данных представляет собой комплексную геоинформационную систему, которая содержит и постоянно обновляет все сведения по современным и ископаемым организмам, как авторского, так и обширного опубликованного материала. Она состоит из нескольких взаимосвязанных блоков: собственно биологической базы, библиографической, базы кодировки геологического пространства и картографической основы. Каждый из этих блоков представляет собой отдельную информационную систему, которая может быть использована и для создания прочих баз данных (рис. 1).

Биологический блок является универсальным и позволяет хранить информацию по любой группе фауны и флоры, не зависимо от их морфологической характеристики, возрастного диапазона существования и разнообразия экологических обстановок существования. В настоящее время с целью комплексного анализа биологических данных он включает сведения по современным и ископаемым микроорганизмам (радиоляриям, диатомовым водорослям, силикофлагеллатам, эбриидеям, актиномицидам, кокколитофоридам, фораминиферам) и многоклеточным (губки, кораллы, моллюски, насекомые, высшие животные и растения). Всего на данный момент содержатся сведения более чем по 130 000 видам вышеперечисленных групп организмов, собранные со всех континентов и океанов отобранные авторами в различных экспедициях и из более чем 20 000 опубликованных и рукописных работ.

Главной отличительной особенностью разработанной ИС является основной объект хранения информации. Нами использован не вид или подвидовой таксон, как принято во всех известных аналогичных разработках, а совокупность реально существующих экземпляров, одинаково называемых (в биномиальной ботанической, зоологической или палеонтологической номенклатуре) и главное одинаково понимаемых определенной группой биологов.

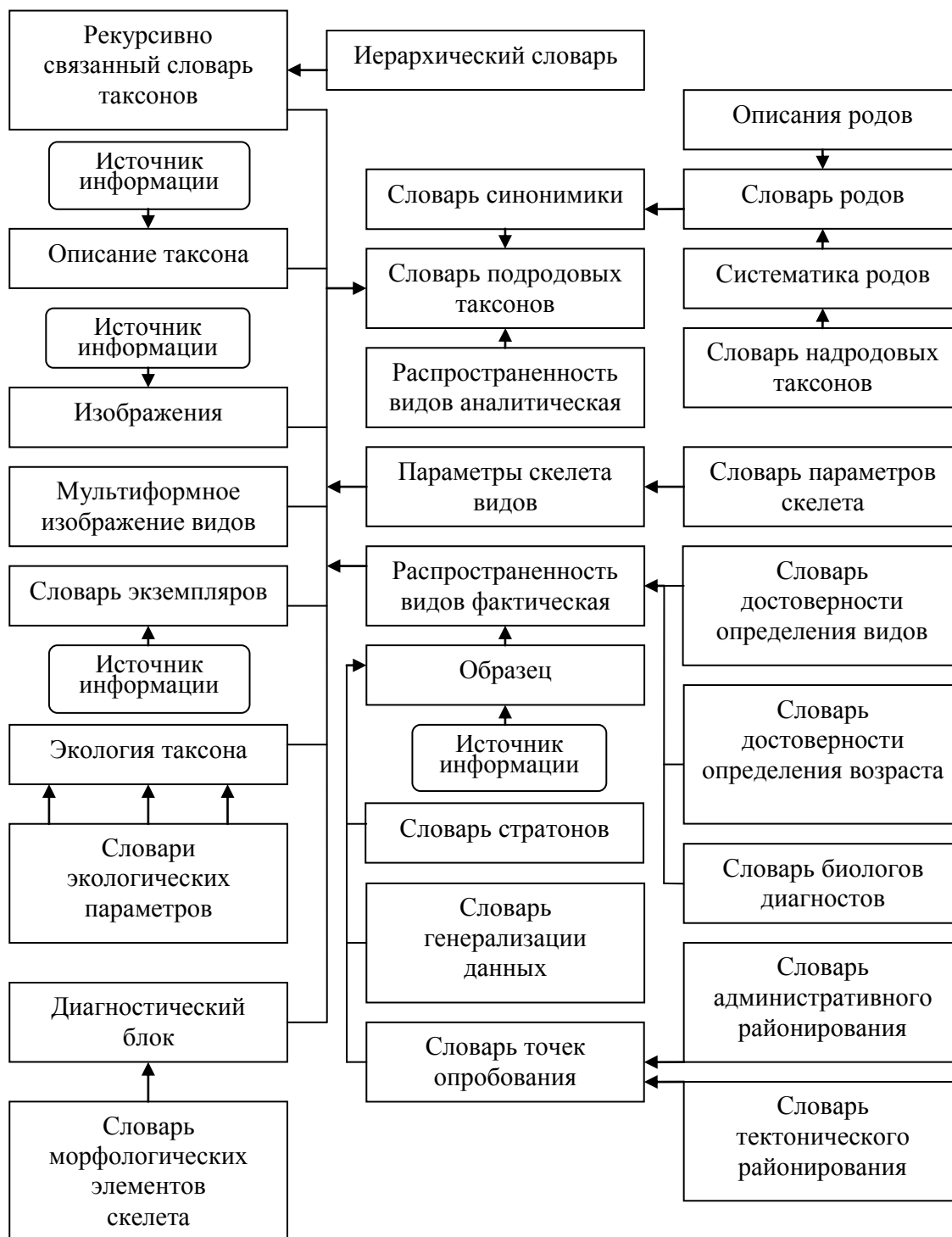


Рис. 1. Упрощенная модель данных биологического блока

Эти экземпляры являются чувственно конкретными и не зависят от исследователя, могут храниться в коллекциях и иметь различные описательные характеристики. Поскольку для общения с людьми и описания мы используем слова, которые при систематиках и классификациях превращаются в термины (абстрактно конкретные), то и в ИС мы вынуждены применять их, заменяя ими реальные объекты. Однако этот набор терминов на латинском языке мы не можем приводить к нормальным формам (разделять на родовое и видовое название, фамилию автора первоописания и др.) и должны понимать только как

целое. Если же в литературе встречается абсолютно одинаковый набор терминов, но отражающий другие морфологические объекты, то мы должны воспринимать его, как самостоятельный таксон. В упрощенном понимании, объектом ИС является терминологический синоним.

Задачей разработки структуры, при таком понимании объекта, является с помощью кодировок, не имеющих семантической нагрузки, т.е. не имеющих смысла, а лишь заменяющих термины для устранения возможных текстовых ошибок при вводе оператором, объединить все синонимы. На первом уровне соединяются все синонимы подвидового уровня, а далее – и на уровне видового таксона. При этом обязательно указывается по кодировке специального словаря тип записи (базиномим, синоним, тавтономим) и уровень иерархии от подформы до вида. Все остальные сведения являются атрибутами базового термина и отражают те или иные свойства экземпляров на различном уровне иерархии и обобщения данных.

В целом палеонтологический (биологический) блок представляет серии реляционных таблиц для каждой крупной таксономической единицы (класс, тип) и связанных системами ключевых полей. Вследствие универсальности таблиц хранения информации и гибкости операционной обработки данных система позволяет проводить анализ как для каждого класса или отдела, так и для всех таксонов высокого ранга вместе. Основными в описываемом блоке являются следующие таблицы:

1. Словарь подродовых таксонов
2. Систематика подродовых таксонов
3. Описание подродовых таксонов
4. Изображение подродовых таксонов
5. Распространенность подродовых таксонов
6. Морфологические признаки подродовых таксонов
7. Словарь морфологических признаков подродовых таксонов
8. Параметры скелета подродовых таксонов
9. Экология подродовых таксонов
10. Возрастное распространение подродовых таксонов
11. Словарь родов
12. Описание родов
13. Систематика родовых таксонов
14. Словарь надродовых таксонов
15. Систематика надродовых таксонов
16. Описание надродовых таксонов
17. Таблица синтеза данных для анализа распространения

Взаимосвязь базовых и части вспомогательных таблиц, которых насчитывается более ста, приведена в модели данных (рис. 1), а общий список таблиц, используемых для хранения исходных данных и их принадлежность к различным блокам приведена в таблице 1.

Базовой в ИС является словарь видов и подвидовых таксонов. Его структура отработана на более чем 20 000 таксонов радиолярий, а в целом таксономический словарь по всем группам насчитывает более 130 000 валидных видов и их синонимиики.

Синонимика организована путем взаимодействия идентификаторов: кода названия таксона (синонима), кода синонимической принадлежности (валидный вид, вид в открытой номенклатуре, базиномим, синоним, тавтономим и др.), синонимиики подвидового таксона (подвид, сорт, форма, подформа), кода синонимиики вида и кода типа записи, отражающего ранг таксона и его уровень (таблица 2).

Название таксонов в ИС приведено к первой нормальной форме и приводится в полном соответствии с кодексами номенклатуры конкретного царства и исключительно в авторском написании. Для создания активной синонимиики словарь таксонов рекурсивно связан сам с собой, для альтернативной синонимиики создано взаимодействие с иной таблицей (Table002) с аналогичной кодировкой, но другими режимами кодов типа записи.

Таблица 1

Основные таблицы палеонтологического блока

1	Table001	Словарь видов	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура для всех групп организмов
2	Table002	Систематика видов	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура для всех групп организмов
3	Table003	Описание видов	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура для всех групп организмов
4	Table004	Изображение видов	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура для всех групп организмов
5	Table005	Распространенность видов	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура для всех групп организмов
6	Table006	Диагностика видов	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура хранения первичных данных
7	Table007	Параметры скелетов вида	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура для всех групп организмов
8	Table008	Экология видов	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура хранения первичных данных
9	Table009	Возрастное распространение видов	Палеонтологическая	Подродовая	Операционная	Расчетная таблица по анализу таблиц 5, 304, 311
10	Table010	Словарь родов	Палеонтологическая	Родовая	Постоянная	Одинаковая структура для всех групп организмов
11	Table011	Описание родов	Палеонтологическая	Родовая	Постоянная	Одинаковая структура для всех групп организмов
12	Table012	Словарь надродовых таксонов	Палеонтологическая	Надродовая	Постоянная	Одинаковая структура для всех групп организмов

Продолжение таблицы 1.

№	Имя таблицы	Название таблицы	Класс	Группа	Тип	Примечания
13	Table013	Систематика надродовых таксонов	Палеонтологическая	Надродовая	Постоянная	Одинаковая структура для всех групп организмов
14	Table014	Описания надродовых таксонов	Палеонтологическая	Надродовая	Постоянная	Одинаковая структура для всех групп организмов
15	Table015	Примечания к подродовым таксонам	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура для всех групп организмов
16	Table016	Видовое богатство	Палеонтологическая	Подродовая, родовая, надродовая	Операционная	Одинаковая структура для всех групп организмов
17	Table017	Экологические параметры	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура для всех групп организмов
18	Table018	Систематика рода	Палеонтологическая	Родовая	Постоянная	Одинаковая структура для всех групп организмов
19	Table019	Экземпляры	Палеонтологическая	Подродовая	Постоянная	Одинаковая структура для всех групп организмов
20	Table020	Описания	Палеонтологическая	Подродовая, родовая	Постоянная	Единая для всех групп организмов
21	Table021	Изображения	Палеонтологическая	Подродовая	Постоянная	Единая для всех групп организмов

Продолжение таблицы 1

№	Имя таблицы	Название таблицы	Класс	Группа	Тип	Примечания
22	Table060	Определитель видов	Палеонтологическая	Видовая	Постоянная	Одинаковая структура для всех групп организмов
23	Table121	Количественные расчеты появления, исчезновения и численности видов по векам	Палеонтологическая	Видовая	Операционная	Одинаковая структура для всех групп организмов
24	Table122	Расчеты видов по эпохам	Палеонтологическая	Видовая	Операционная	Одинаковая структура для всех групп организмов
25	Table201	Источник информации	Библиографическая		Постоянная	Единая для всех баз данных
26	Table202	Библиографическое описание	Библиографическая		Постоянная	Единая для всех баз данных
27	Table301	Точки опробования	Географическая, геологическая	Районирования	Постоянная	Единая для всех баз данных
28	Table304	Стратоны	Геологическая	Стратиграфическая	Постоянная	Единая для всех баз данных
29	Table311	Образец	Геологическая, Биологическая	Стратиграфическая	Постоянная	Единая для всех баз данных
30	Table401	Идентификатор вида измерения	Палеонтологическая Биологическая	Видовая	Постоянная	Единая для всех баз данных

Структура таблицы "Словарь видов" (Table001)

№	Название поля	Тип поля	Число символов	Ключевое поле	Описание
1	F008105 Идентификатор вида	N		Да	Идентификатор объединяет все синонимы вида
2	F008001 Идентификатор записи уровня таксона	N		Да	Устанавливает тип записи в таблице (уровень иерархии таксона и уровень валидности). Берется из таблицы «Hierarchy»
3	F008202 Идентификатор синонимов	N		Да	Под этим кодом объединяются все синонимы выделяемого таксона
4	F008201 Идентификатор названия	N		Да	Содержит уникальный код, присваиваемый группе экземпляров выделенного таксона и имеющих данное синонимическое название в ряде работ
5	F001007 Род	A	25		Содержит название рода
6	F001006 Подрод	A	18		Содержит название подрода
7	F008203 Знаки открытой номенклатуры	A	25		Поле содержит элементы названия таксона обычно расположенные между названием рода и вида
8	F001005 Вид	A	28		Содержит название вида
9	F008204 Автор вида и тип подвидового таксона	A	21		Используется для любых необходимых вставок между названием вида и подвидового таксона
10	F001004 Подвид	A	38		Содержит название подвидовых таксонов (подвид, сорт, форма, подформа)
11	F002003 Автор	A	23		Сокращенное имя автора таксона
12	F002002 Полный автор	A	76		Автор, впервые давший данное название таксона
13	F008002 Источник информации	N			Код работы, в которой впервые появилось данное название таксона и в данном смысле. Берется из таблицы Table202.
14	F001107 Устаревший род	A	55		Старое название рода при переносе вида из одного рода в другой
15	F009001- F009012 Метка 1-12	A Или N	5		Поле для различных меток

Авторские описания и изображение таксонов (Table003, Table004) содержат все необходимые ссылки на страницы в первоисточнике и номера палеонтологических таблиц и рисунков. Просмотр осуществляется как по одиночным изображениям, так и в мультимедийном формате с возможностью выбора количества одновременных показов (рис. 2).

Морфологические параметры являются самими сложными при создании единой структуры хранения. Эту проблему можно решить путем создания таблицы, содержащей код таксона, код признака и параметр, а затем необходимо преобразовать с помощью подпрограмм в пересчетный массив, учитывающий особые признаки для каждой крупной таксономической группы.

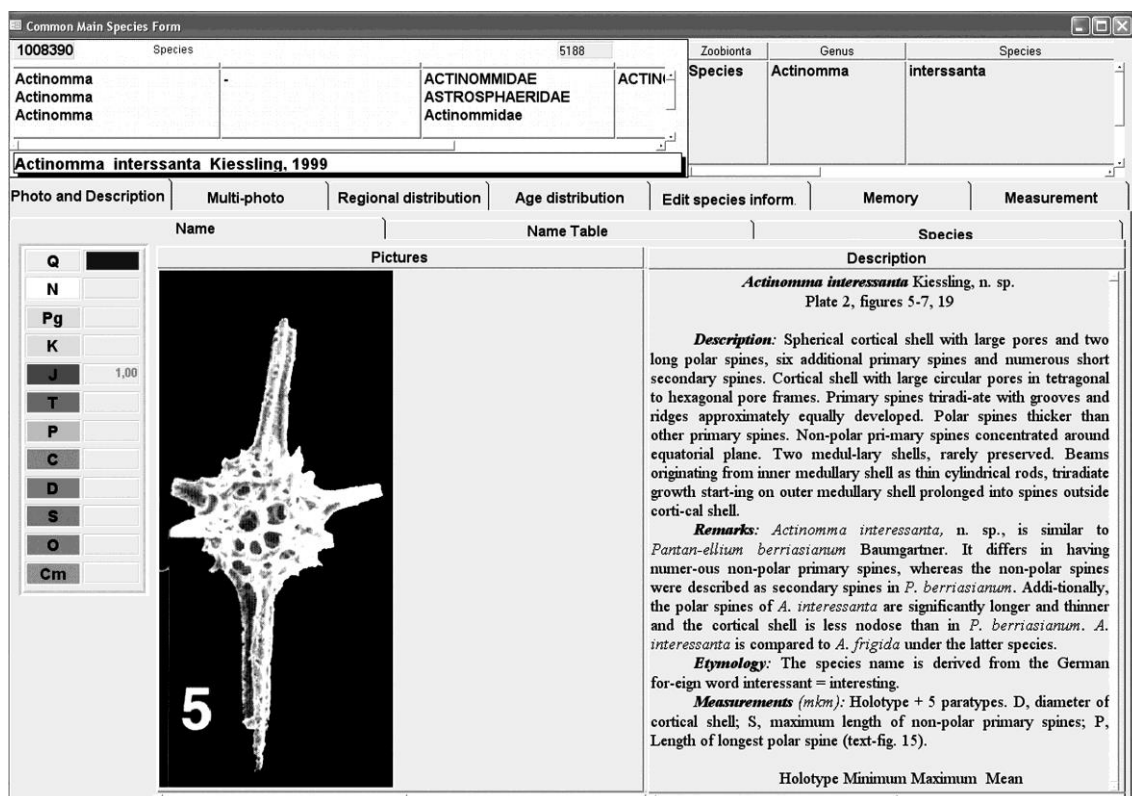


Рис. 2. Основная форма. Страница Edit species information

Блок пространственно-возрастной распространенности позволяет хранить сведения о таксонах в любой удобной пользователю системе привязки к местности (административная, тектоническая и др.), включая географические координаты, позволяющие в будущем связывать ИС с электронными картами. Организация информации построена по принципу таксон-образец, где для образца приведена вся исчерпывающая информация, включая уровни достоверности определения вида, возраста, географического положения точки.

Возрастная кодировка производится в различных вариантах (ярусы, свиты, зоны) и не ограничивает пользователя (возможна кодировка возрастной распространенности группы по зонам других ископаемых групп, многократное альтернативное кодирование и др.). Стратиграфический словарь включает более 5500 стратиграфических терминов. При учете абсолютной геохронологии используется международная шкала [8-10], но предусмотрены и другие датировки. Исходная информация хранится в авторских стратонах, что позволяет в случаях изменения понимания его объема или переноса оперативно менять ранговые возрастные объемы. С этой целью создан целый ряд программ перевода из одной шкалы в другую. Например, из зональной или региональной в стандартную и наоборот.

Для увеличения полноты информации из первоисточника необходимо использовать словари достоверности или точности данных. В ИС существуют сведения об авторах

определения видов, поскольку автор первоисточника и автор диагностики иногда не одно лицо. Важно указывать сомнения автора о точности диагностики, полноте экземпляров или тот факт, что сведения основаны на следах жизнедеятельности организмов, спорах, вегетативных клетках, переработанном или перенесенном материале и др. Учтена и дифференциация данных по уровню обобщения исходных сведений, например, были ли использованы конкретные образцы или приводится информация по группе образцов, а то и по разрезу или скважине в целом.

Количественные данные должны учитывать методики подсчетов и давать возможность пересчитывать информацию под один из стандартов.

Особую информацию представляет литологический блок, характеризующий геологическую среду отбора (образец), а также блок погодных условий ложа при изучении живых объектов.

Экологический блок существенно различается по набору анализируемых факторов в зависимости от уровня лабильности организмов различных таксонов высокого ранга к среде обитания. Наиболее широко представлены сведения по диатомовым водорослям, распространенным во всех биотопах и имеющим широкий спектр организмов по солености, температуре и др. Главными экологическими параметрами, учитываемыми при внесении в ИС являются: среда обитания (водные, почвенные, аэральные и др.); подвижность (планктон, бентос с детализацией на 3 уровня, определяющий объекты обрастания, субстрат); соленость; галобность; температурный режим; региональные флористические и зоологические провинции (широты, области); отношение к течениям (реофильные, лимнофильные); апвеллинговые формы; приливные формы; способность (нутриентность); Ph-среды; вегетационный период. Данный блок применяется при восстановлении палеоэкологических обстановок и мониторинга окружающей среды.

Библиографический блок в палеонтологическом анализе в основном используется для передачи данных об источниках информации, однако он может работать как самостоятельная библиографическая информационная система, позволяющая включать значительное количество работ. Данный блок помимо стандартного описания источника информации на языке оригинала позволяет хранить реферат источника, а иногда и сам полнотекстовый источник со всеми имеющимися в нем иллюстрациями. Поисковая система данного блока может позволить находить публикации и рукописные источники практически по любому параметру (трехуровневая тематика публикации, автор и год издания, вид издания, язык, описываемые в работе классы организмов, геологический возраст, четырехуровневая рубрикация географического охвата работы, петрографическая характеристика объектов и многое другое).

Предложенная структура хранения данных позволяет осуществлять быстрый поиск информации по любому виду и его синонимам, просматривать описания и изображения экземпляров, знакомиться с особенностями распространения видов, строить схемы филогенетических связей, биоразнообразия и послойные карты биогеографического районирования, определять геологический возраст горных пород.

Примечания:

1. Riedel W.R. IDENTIFY; a Prolog program to help identify fossils // Computers and Geosciences. 1989. Vol. 15 (5). P. 809-823.
2. Riedel W.R., Tway L.E. Intelligent from and back ends geological database // Proceedings of the 4th International Conference on Geoscience Information. Geological Survey of Canada, Open File. 1994. 2315. P. 400-408.
3. Radiolarian catalogue and systematics of Middle Jurassic to Early Cretaceous Tethyan genera and species / Peter O. Baumgartner, Luis O'Dogherty, Spela Gorican, Ruth Dumitrica-Jud, Paulian Dumitrica, Alain Pillevuit, Elspeth Urquhart, Atsushi Matsuoka, Taniel Danelian, Annachiara Bartolini, Elisabeth S. Carter, Patrick De Wever, Norio Kito, Marta Marcucci, Torsten Steiger // Middle Jurassic to Lower Cretaceous Radiolaria of Tethys: Occurrences, Systematics, Biochronology. Mémoires de Géologie. Lausanne, 1995. Vol. 23. P. 37-685.
4. Catalogue of life: <http://www.catalogue-oflife.org> (дата обращения: 01.10.2015)
5. Агарков Ю. В. Информационная система по радиоляриям // Тез. Докл. Десятого семинара по радиоляриям. М.: Ин-т литосферы РАН, 1996. С. 6.

6. Агарков Ю. В. Новые данные о видовом разнообразии радиолярий мезозоя // Матер. 11 Семинара по радиоляриям (19-24 июня 2000). - СПб-М, 2000. С. 9-10.
7. Agarkov Yu. V. Universal Biological Information System "Biobase" // Proceedings of the Sixth International Conference Environmental Micropaleontology, Microbiology and Meiobenthology, Moscow, 2011. P. 18-21.
8. Афанасьев С.Л. Изотопная геохронологическая шкала венда – фанерозоя // Геол. и геофиз. 1993. 34, № 3. С. 3–9.
9. Афанасьев С.Л. Наноциклитная геохронологическая шкала венда – фанерозоя // Бюл. Моск. о-ва испыт. природы. Отд. геол. 1993. Т. 68, № 2. С. 24–30.
10. The Geologic Time Scale 2012 / ed. F.M.. Gradstein, J.G. Ogg, M. D. Schmitz, G.M. Ogg. Amsterdam, London, New York, Oxford. Tokyo. 2012. 1160 p.

References:

1. Riedel W.R. IDENTIFY; a Prolog program to help identify fossils // Computers and Geosciences. 1989. Vol. 15 (5). P. 809-823.
2. Riedel W.R., Tway L.E. Intelligent front and back ends geological database // Proceedings of the 4th International Conference on Geoscience Information. Geological Survey of Canada, Open File. 1994. 2315. P. 400-408.
3. Radiolarian catalogue and systematics of Middle Jurassic to Early Cretaceous Tethyan genera and species / Peter O. Baumgartner, Luis O'Dogherty, Spela Gorican, Ruth Dumitrica-Jud, Paulian Dumitrica, Alain Pillecuit, Elspeth Urquhart, Atsushi Matsuoka, Taniel Danelian, Annachiara Bartolini, Elisabeth S. Carter, Patrick De Wever, Norio Kito, Marta Marcucci, Torsten Steiger // Middle Jurassic to Lower Cretaceous Radiolaria of Tethys: Occurrences, Systematics, Biochronology. Mémoires de Géologie. Lausanne, 1995. Vol. 23. P. 37-685.
4. Catalogue of life:<http://www.catalogue-of-life.org> (data obrashcheniya: 01.10.2015)
5. Agarkov Yu. V. Informatsionnaya sistema po radiolyariyam // Tez. Dokl. Desyatogo seminaru po radiolyariyam. M.: In-t litosfery RAN, 1996. S. 6.
6. Agarkov Yu. V. Novye dannye o vidovom raznoobrazii radiolyarii mezozoya // Mater. 11 Seminaru po radiolyariyam (19-24 iyunya 2000). SPb-M, 2000. - S. 9-10.
7. Agarkov Yu. V. Universal Biological Information System "Biobase" // Proceedings of the Sixth International Conference Environmental Micropaleontology, Microbiology and Meiobenthology, Moscow, 2011. P. 18-21.
8. Afanas'ev S.L. Izotopnaya geokhronologicheskaya shkala venda – fanerozoya // Geol. i geofiz. 1993. 34, № 3. S. 3–9.
9. Afanas'ev S.L. Nanotsiklitnaya geokhronologicheskaya shkala venda – fanerozoya // Byul. Mosk. o-va ispyt. prirody. Otd. geol. 1993. T. 68, № 2. С. 24–30.
10. The Geologic Time Scale 2012 / ed. F.M.. Gradstein, J.G. Ogg, M. D. Schmitz, G.M. Ogg. Amsterdam, London, New York, Oxford. Tokyo. 2012. 1160 p.

УДК 004.9:550.84

Принципы построения палеонтологических и биологических баз данных

¹ Юрий Васильевич Агарков

² Андрей Юрьевич Агарков

¹ Сочинский научно-исследовательский центр РАН, Российская Федерация
Кандидат геолого-минералогических наук, доцент, старший научный сотрудник
354000, г. Сочи, ул. Театральная, 8А., к. 6
E-mail: agarkov2011@yandex.ru

² Палеонтологический институт РАН, Москва, Российская Федерация
Аспирант
346888, Ростовская обл., г. Батайск, Артемовская 55
E-mail: aagarkov77@yandex.ru

Аннотация. В работе приведена структура универсальной синонимической и альтернативно-синонимической информационной системы, позволяющей хранить любую информацию в области естествознания. Определены основные и сопутствующие объекты хранения информации. Предложены типы таблиц для ввода-вывода данных и структура базового словаря видов. Охарактеризованы блоки пространственно-временного распространения и библиографии, вспомогательные словари.

Ключевые слова: информационные системы, биоразнообразие, радиолярии, диатомовые водоросли, палеоэкология, мониторинг окружающей среды.