Электронные коллекции Зоологического института по морским животным и метаданные*

©И.С. Смирнов, О.Н. Пугачев, А.Л. Лобанов, А.Ф. Алимов, Е.П. Воронина

Зоологический институт РАН
smiris@zin.ru

Аннотация

Одной из важных проблем, стоящих перед современным обществом является сохранение биологического разнообразия. ИПС способствуют решению фундаментальных проблем зоологии.
Зоологический институт РАН разрабатывает различные проекты по биологическому разнообразию России, сопредельных территорий и полярных регионов.
Одним из условий создания современных информационно-поисковых систем, упрощения и ускорения поиска необходимой информации по биоразнообразию в Интернете является интеграция и стандартизация баз данных и метаданных. В настоящее время в зоологических ИПС используются следующие стандарты для ввода, описания и представления данных: Darwin Core, RDF, Dublin Core Metadata Elements и др.

1. Информационно-поисковые системы по биоразнообразию

Одной из важных проблем современности, стоящих перед человечеством, является сохранение биоразнообразия Земли. Решить эту проблему невозможно без изучения видового состава и классификации живых организмов, что является предметом таксономии и систематики, основанных на исследовании, в первую очередь, биологических коллекций, собранных на протяжении многих лет [6]. Информационно-поисковые системы облегчают работу ученых, имеющих дело с полевыми и каталожными данными биологических коллекций, зачастую носящими фрагментарный характер. Зарубежный опыт создания баз данных довольно обширен и насчитывает уже десятки лет, поскольку внедрение ЭВМ за границей началось раньше и шло интенсивнее, чем в России. На сегодняшний день уже значительная часть крупных естественноисторических учреждений мира (Natural History Museum, London; Museum National d'Histoire Naturelle, Paris; California Academy of Sciences, San Francisco; National Museum of Natural History, Washington; National Science Museum, Tokyo) имеет свои сайты, на которых выставлены пока достаточно разрозненные электронные каталоги или базы данных по коллекциям различных групп животных [3,4]. На сайте Американской антарктической программы (USAP) появился электронный каталог беспозвоночных животных http://www.nmnh.si.edu/iz/usap/usapdb.html, а крупным Интернет-проектом, объединяющим самые разные аспекты ихтиологической науки, в том числе и коллекции рыб различных музеев мира, стал проект FishBase: http://fishbase.com.
К сожалению, отсутствие универсального подхода у зарубежных и отечественных авторов, наличие множества оригинальных компьютерных методик и систем управления базами данных (СУБД) на основе различных модификаций компьютеров, а также существование ряда проблем (например, невозможность ввода кириллических и других неанглоязычных символов), не позволяет широко применять уже спроектированные модели банков данных и информационно-поисковых систем. Опыт создания баз данных в Зоологическом институте РАН позволил разработать, в частности, информационно-поисковую систему «ОКЕАН», включающую станционную базу данных о местах сбора и поимки морских беспозвоночных и рыб. В сочетании с таксономическим классификатором, содержащим сведения о составе фауны определенного региона и коллекционной базой данных (сведения о месте и способе хранения собранного материала), станционная база данных позволяет проводить поиск информации по многочисленным и разнообразным запросам. Система «ОКЕАН» обеспечивает возможность ввода данных и выполнение различных запросов с использованием как латиницы, так и кириллицы [1, 5].

2. Интеграция и стандартизация

Растущее число доступных в Интернете ресурсов заставляет пользователя в поисках

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Ярославль, Россия, 2007.

необходимой информации посетить каждый сайт в отдельности. Один из путей упрощения и ускорения этого поиска - интеграция и стандартизация баз данных и метаданных.

Интеграция и стандартизация становятся все более актуальными задачами и проблематикой работы многочисленных международных организаций [3,4]. Такой организацией является Рабочая Группа по Таксономическим базам данных (Taxonomic Databases Working Group - TDWG), первая встреча которой состоялась в 1985 г. в Женеве, а очередная ежегодная встреча, спустя двадцать лет, была проведена в 2005 г. в Зоологическом институте РАН, в Санкт-Петербурге [2] (http://www.zin.ru/conferences/tdwg/index.html).

Своей основной задачей TDWG считает разработку и согласование стандартов обмена данными между отдельными базами (http ://www.nhm. ac.uk/hosted_sites/tdwg/). Как показал опыт работы группы, многие ранние стандарты успешно используются, некоторые исчезли, а некоторые находятся до сих пор в процессе стабилизации. С другой стороны, среди большого числа сложных, многоуровневых стандартов - ABCD (Access to Biological Collection Data, http://www.bgbm.org/TDWG/CODATA/Schema/), Darwin Core (http://speciesanalyst..net/docs/dwc/), HISPID (http://plantnet.rbgsyd.nsw.gov.au/HISCOM/HISPID/HISPID3/hispidright.html), Linnean Core (http://wiki.cs.umb.edu/twiki/bin/view/Ants/EntomologyMarkupDiscussion), многие сосредоточены на обмене данными, а не на действительной записи данных для конкретных экземпляров.

В настоящее время стандартом ввода данных считается Darwin Core 2 (http://darwincore.calacademy.org/), представляющий собой набор определений элементов данных, разработанный для поддержки совместимости и интеграции первичных данных по биоразнообразию, предложенный как проект в 2004 г., и в настоящее время развивающийся как стандарт TDWG.

3. Таксономические стандарты

Несмотря на бурный рост информационных технологий биологические и, в частности, зоологические исследования медленно поддаются стандартизации и компьютеризации в силу большой сложности систематических и номенклатурных отношений. При организации фаунистических и экологических банков данных и ИПС встают две серьезные проблемы: а) ведение записей по систематике и таксономии организмов, особенно учет синонимических названий, и б) представление географических данных.

К базам данных, отсылающих к таксономическим системам, относятся экологические списки видового состава, описания отдельных экземпляров или видов, данные для единиц хранения биологических коллекций. Наличие, обмен и понимание таксономической информации имеют исключительную важность для систематиков, экологов, биологов и законодателей. Такая информация поддерживается несколькими глобальными и локальными таксономическими сервисами баз данных и оперирует валидными (действительными или правильными) научными названиями видов, составляющими таксономическую концепцию. Базы данных обычно отражают одну таксономическую точку зрения, стремясь соотнести ее с синонимическими названиями. Работа систематиков направлена на улучшение классификации и определений организмов, установление рамок таксонов путем создания описаний, ключей, списков и т.д. Основная проблема заключается в том, что названия таксонов являются частью общей (принципиальной) схемы, а пользователи предпочитают оперировать отдельными названиями независимо от общей схемы. В этой ситуации становится очевидной необходимость развития абстрактной модели таксономической концепции, делающей различные модели доступными провайдерам данных. Схема концепции таксона TCS (Taxon Concept Schema - http://tdwg.napier.ac.uk/index.php?pagename=HomePage) разработана с тем, чтобы удовлетворять эту потребность. Модель в виде XML документа предложена как стандарт, обеспечивающий обмен данных без искажения, и облегчающий поиск между разными моделями данных. В 2004 г. Дж. Купером предложена схема Linnean Core (http://bdei.cs.umb.edu/twiki/bin/view/UBIF/LinneanCoreChoiceOfName) для более точного определения проблемы научного названия. Впоследствии схема была доработана, элемент Name заменен элементом Taxon Name, а названия таксонов повышены до элементов высшего уровня, определяющих валидные взаимоотношения, которые можно установить между именами. Схема TCS способствует представлению таксономических систем согласно опубликованным таксономическим классификациям, ревизиям и т.д. Вместе с тем, задача корректного использования названий таксонов в принципиальных схемах метаданных остается не решенной

Во многих случаях с одним названием может быть связано более одного вида, поскольку типовой экземпляр данного названия подпадает под определение нескольких разных авторов, кроме того, могут существовать разные взгляды на рамки конкретного вида. Для точного использования таксона необходимо указывать научное название и его автора. В базах данных это достигается использованием элемента «According to», соответствующего латинским выражениям sensu или sec. (secundum - согласно). К сожалению, большинство данных не содержит этот элемент, обычно пользователи употребляют названия видов без уточнения автора. При небольшом числе используемых названий концепции номинальных таксонов могут быть удовлетворительны, но с увеличением объема данных и уровней

взаимодействия необходимо корректное обозначение рамок таксонов. Таким образом, концепция таксона включает в себя его название и описание. При ссылке на таксон необходимо использовать его название плюс «According to», что дает точную адресацию его рамок.

4. Представление географических данных

Представление географической информации в базах данных требует обеспечения ее точности, актуальности и периодического обновления, и может быть в виде точки, линии или полигона. Для ее ввода привлекаются оцифрованные бумажные карты, данные наземных съемок, GPS-измерения, данные дистанционного зондирования, а также данные, полученные фотограмметрическими методами. Источниками географических данных являются полевые описания места сбора материала, относящиеся к точечному типу географических объектов, поскольку описывают участок фиксированной площади от нескольких десятков квадратных сантиметров до нескольких сотен квадратных метров [7]. Для полевого описания места используются следующие формы географической привязки:

1. Указание географической широты и долготы - универсальный метод, соответствующий высокой степени точности и распространен в современных полевых записях. В морских гидробиологических коллекциях широта и долгота места сбора материала (станции) указаны на основании судовых данных (ship log-based).

2. Описания местности (например, к северо-западу от банки Агульяс) - и наиболее распространенный способ привязки, сопровождающий полевые записи, и сильно варьирующий по степени точности. Такая запись может быть преобразована в точечный объект, но часто это не представляется возможным.

3. Привязка к административным территориальным единицам (город, область, район) и географическим объектам (бассейнам рек, озерам). Этот тип информации соответствует полигональному типу географических объектов и обладает низкой степенью точности локализации, хотя и широко применяется в полевых зоологических исследованиях.

Для перевода в электронную форму записей полевых дневников и их географической привязки при вводе географических данных исследователь вводит документ в соответствующую базу данных, которая, кроме тематической информации, содержит специальный объект (поле), описывающий координаты сбора биологического объекта (широта - latitude, долгота - longitude), а для морских сборов и такие параметры как глубина, тип дна, метод и орудие лова, дату и имя сборщика. Связь биологических (в частности, зоологических) баз данных со стандартной картографической основой при вводе и презентации первичной информации закладывается с помощью использования стандартов ввода данных, таких как Darwin Core. Составление новых и использование уже имеющихся тезаурусов, таких как The Getty Thesaurus of Geographic Names TGN) http://www.getty.edu/research/conducting_research/vocabularies/tgn/) облегчает ввод, поиск и обработку географических данных.

5. Метаданные

Наиболее перспективной и общеупотребительной моделью описания метаданных со стандартным набором элементов является система RDF (Resource Description Framework), созданная международной организацией W3C (World Wide Web Consorcium). Распространенный набор элементов метаданных "Dublin Core Metadata Elements" (http ://dub1incore.org/documents/dces/), разрабатываемый международной группой "The Dublin Core Metadata Initiative" (DCMI http://dublincore.org/), состоит из 15 характеристик, условно разбитых на три группы:

Content элементы, относящиеся к содержанию ресурса	**Intellectual Property** элементы, рассматриваемые с позиции интеллектуальной собственности	**Instantiation** элементы, относящиеся к данному экземпляру ресурса
Title	Creator	Date
Subject	Publisher	Format
Description	Contributor	Identifier
Type	Rights	Language
Source		
Relation		
Coverage		

Для сохранения совместимости с простейшим описанием из 15 элементов и в то же время увеличения детализации описаний разрабатывают дополнительные *квалификаторы* для базовых элементов. Наиболее часто цитируемые предложения: "Dublin Core Qualifiers/Substructure" и "Dublin Core qualifiers". Расширять сам набор элементов можно с использованием уже имеющихся стандартов. Каждый элемент Dublin Core произвольный и повторяемый. В DCMI имеются установленные стандартные пути для детализации элементов, которые должны способствовать использованию схем кодирования и словарей. В Dublin Core нет установленного порядка для представления или использования элементов.

Информационная составляющая Международного проекта «Перепись морского населения» (Census of Marine Life http://www.coml.org/) Океанская биогеографическая информационная система OBIS (http://ww\v.iobis.org/) ориентируется на стандарты Международной системы регистрации данных GCMD (Global Change Master Directory - http://gcmd.nasa.gov/) [3]. Минимальный набор полей **(Metadata terms),** используемый для описания портала следующий: Database name; Citation; Taxonomic coverage; Geographic coverage; Temporal coverage; Habitat coverage; Total distribution records; Total number of taxa; Collection method; Data source; Abstract; Publications from this data; Scientific Contact; Technical contact; Website; Date this form completed (http://www.iobis.org/tech/metadata1/).

Описание данных с использованием международных стандартов позволяет эффективнее разрабатывать локальные информационные ресурсы и интегрировать эти данные первоначально в виде метаданных и затем непосредственно в информационные проекты международных программ, которые получили развитие в последнее время. В качестве иллюстрации ниже приведены метаданные по коллекциям морских беспозвоночных арктического региона, хранящимся в Зоологическом институте РАН, созданные для интероперабельного обмена между участниками международных проектов.

Дата формирования (Report Date): 17марта2007 г.

Название пакета метаданных (Metadata Data Set Name): Коллекция Зоологического Института РАН арктических бентосных беспозвоночных

Общая характеристика (Identification Information)

Ссылки (Citation): Список видов свободноживущих беспозвоночных Евразийских морей и прилежащих глубоководных частей Арктики Сиренко Б.И., ред. 2001. Исследования фауны морей. 51(59). Санкт-Петербург: 1-132.

Описание (Description)

Аннотация (Abstract): База данных описывает одну из крупнейших в мире коллекцию Зоологического института РАН (ЗИН). Коллекция включает свыше 100 тысяч образцов 26 тысяч видов морских беспозвоночных. Создание коллекционной БД по беспозвоночным, начатое в 1987 г., очень перспективно для изучения биоценотических отношений и экосистемы морской фауны. С целью систематизации фаунистических, экологических и коллекционных данных в ЗИН РАН разработаны ИПС ОКЕАН, ЗООКОД и ЭКОАНТ. Формат БД dBase.

Цель **(Purpose):** Определить видовой состав, численность и распространение фауны беспозвоночных Арктического региона.

Дополнительная информация (Supplemental Information): Данные использованы в работе: Smirnov I.S. et al. 2007. Creation of the information retrieval system for collection of the marine animals (fish and invertebrates) at the Zoological Institute of the Russian Academy of Sciences // In Vanden Berghe, E. et al. (Eds). Proceedings 'Ocean Biodiversity Informatics' - International Conference on Marine Biodiversity Data Management, Hamburg, German, 29 November-1 December 2004 IOC Workshop Report No. x / BSH/ VLIZ Special Publication 20, p. 177-186.

Период накопления содержания БД (Time Period of Content)

Временной интервал (Time Period Information): 1861-2004

Состояние на текущий момент (Currentness Reference): в процессе создания

Частота поддержания и обновления (Maintenance and Update Frequency): периодически

Пространственная зона (Spatial Domain)

Описание географического охвата (Description of Geographic Extent): Арктический регион

Координаты границ (Bounding Coordinates)

Западная граница (West Bounding Coordinate): 0.00

Восточная граница (East Bounding Coordinate): 170.00 W

Северная граница (North Bounding Coordinate): 85.00 N

Южная граница (South Bounding Coordinate): 65.00 N

Ключевые слова (Keywords)

Тема (Theme)

Ключевое слово темы (Theme Keyword Thesaurus): зообентос

Ключевое слово темы (Theme Keyword): видовой состав

Ключевое слово темы (Theme Keyword): численность

Ключевое слово темы (Theme Keyword): распространение

Место (Place)

Ключевое слово места (Place Keyword): Арктический регион

Ключевое слово места (Place Keyword): арктические моря Евразии

Таксономия (Taxonomy)

Таксономические ключевые слова (Taxonomic Keywords) (каждый вид + научное название рассматривается одним таксономическим ключевым словом): список включает 160 видов

Таксономические процедуры (Taxonomic Procedures)

Таксономические процедуры (Taxonomic Procedures): Определение проведено в лаборатории специалистами по разным группам животных (Ю.И.Галкин,

Б.И.Сиренко, С.В.Василенко, И.С.Смирнов, А.В.Смирнов, Н.Л.Цветкова, А.А.Голиков, ЗИН РАН) на основании определительных таблиц, опубликованных Академией наук России. Все экземпляры были определены до видов. Большинство экземпляров хранится в лаборатории морских исследований ЗИН РАН, Санкт-Петербург, Россия.

Ограничения доступа (Access Constraints): нет

Ограничения использования (Use Constraints); Данные не могут быть переданы третьему лицу

Финансовая поддержка создания БД (Data Set Credit): Работа поддержана РФФИ и ArcOD Program

Программное окружение (Native Data Set Environment): FoxPRO, Windows

Информация о местоположении и направлении станций (Point and Vector Object Information)

Число станций (Point and Vector Object Count): 7436

Информация об объекте и его характеристиках (Entity and Attribute Information)

Подробное описание (Detailed Description)

Тип объекта (Entity Type)

Название объекта (Entity Type Label): staye.dbf

Определение типа объекта (Entity Type Definition): пакет данных является файлом MS-DOS, созданным на основании файла FoxPRO формата dbf

Атрибут (Attribute): Название поля

Наименование атрибута (Attribute Label): NUMREC

Определение атрибута (Attribute Definition): Номер записи

Источник определения атрибута (Attribute Definition Source): ЗИН РАН

Атрибут (Attribute): Название поля

Наименование атрибута (Attribute Label): VOY

Определение атрибута (Attribute Definition): Номер рейса исследовательского судна

Источник определения атрибута (Attribute Definition Source): ЗИН РАН

Сходным образом дана характеристика еще 46 полей базы staye.dbf, а также баз данных anistay.dbf с 5 полями и collect.dbf с 15 полями)

Способы представления и передачи данных (Standard Order Process)

Цифровая форма (Digital Form Windows)

Название формата (Format Name): Foxpro

Цифровые способы передачи данных (Digital Transfer Option): networks, ftp, website, email, CD

Сетевой ресурс (Network Resource Name): локальная сеть, интернет

Справочная информация о метаданных (Metadata Reference Information) (Metadata Reference Information)

Дата Формирования метаданных (Metadata Date): 17 марта 2007 г.

Контакты (Metadata Contact):

Контактное лицо (Contact Person): Игорь Сергеевич Смирнов

Контактная организация (Contact Organization): Зоологический институт РАН

Контактный адрес (Contact Address): 199034 Санкт-Петербург, Университетская наб., 1

Контактный телефон (Contact Voice Telephone): 7 812 328 1311

Название пакета метаданных (Metadata Standard Name):

Коллекция Зоологического Института РАН арктических бентосных беспозвоночных. Версия стандарта ArcBIC-ZIN-0017.03.2007

Ограничения доступа к метаданным (Metadata Access Constraints): нет

Ограничения использования метаданных (Metadata Use Constraints): нет

Литература

[1] Smirnov Igor S. , Andrei L. Lobanov, Alexei A. Golikov, Elena P. Voronina & Alexey V. Neyelov. Creation of the information retrieval system for collections of the marine animals (fish and invertebrates) at the Zoological Institute of the Russian Academy of Sciences // In: Vanden Berghe, E., W. Appeltans, M.J. Costelio, Pissierssens P. (Eds). Proceedings of Ocean Biodiversity Informatics: an international conference on marine biodiversity data management Hamburg, Germany, 29 November - 1 December, 2004. Paris, UNESCO/IOC, VLIZ, BSH, 2007. P. 177-186.

[2] Taxonomic Databases Working Group, 2005 Annual Meeting, 11-18 September 2005, St. Petersburg, Russia. Abstracts. (Edited by W.G. Berendsohn and Adrian Rissone). SPb. 2005. P. 1-42.

[3] Смирнов И.С., А.Л. Лобанов, А.Ф. Алимов, А.А. Голиков, А.Г. Кирейчук. Международные Интернет-проекты по созданию биологических электронных коллекций // Научный сервис в сети ИНТЕРНЕТ: Труды Всероссийской научной конференции (18-23 сентября 2006 г., г. Новороссийск). - М.: Изд-во МГУ, 2006. с. 216-218.

[4] Смирнов И.С., А.Ф.Алимов, А.Г.Кирейчук, Е.П.Воронина, А.Л.Лобанов. Международные проекты по созданию электронных коллекций морских животных: первые результаты // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Седьмой Всероссийской научной конференции (RCDL'2005). Ярославль, 4 - 6 октября 2005 г. - Ярославль: Ярославский