ISSN 0206-0477

RUSSIAN ACADEMY OF SCIENCES

ZOOLOGICAL INSTITUTE
BOTANICAL INSTITUTE

INFORMATION RETRIEVAL SYSTEMS IN BIODIVERSITY RESEARCH

St. PETERSBURG
1999

Volume is devoted to 275th Anniversary of the Russian Academy of Sciences

Symposium Organizers gratefully acknowledge the financial support of the Russian Foundation for Basic Research, assistance of the Linnean Society, London, UK and technical assistance of the Lintec Computer company

Information Retrieval Systems in Biodiversity Research. (Abstracts of the International Symposium). - *Proceedings of the Zoological Institute RAS*. 1999. Vol. 278. 139 pp.

Abstracts of Reports presented at the International Symposium "Information Retrieval Systems in Biodiversity Research" and 5 reviews on the main fields of data-retrieval systems used in biodiversity study are included in this book. Symposium has been held from 23th to 27th May 1999 in the main building of the Zoological Institute, Russian Academy of Sciences.

Zoological Institute RAS, 199034, St. Petersburg, Universitetskaya Embankment, 1
Phone (812) 3280011, 3280311 Fax: (812) 3282941
Symposium WWW-site: http:/www.zin.ru/conferences/irsb

немецкой фирмой dialobis edition для подготовки руководств по жукам и деревьям на CD-ROM. Новая версия BIKEY, создание которой средствами C++ Builder завершается сейчас, предназначена для работы в Windows. В этой версии сняты ограничения на число таксонов и признаков в одном ключе. Число возможных состояний одного признака увеличено до 16. Сделан переход от жестко детерминированного диагноза (когда таксоны удаляются из списка возможных уже при одном несовпадении состояния признака) к вероятностному диагнозу (когда таксоны остаются в списке возможных до достижения заданного пользователем порога числа несовпадений). Предоставлена возможность выбора нескольких состояний одного признака. Как и в предущих версиях для хранения данных используются базы данных стандартного формата DBF. Система BIKEY8 предназначена для автоматизации многих процессов работы с диагностической информацией о биологических таксонах. Для конечного пользователя наиболее интересной в этой системе является программа диалогового определения PICKEY8. В ее новой версии оставлен привычный для пользователей PICKEY интерфейс с минимизированным числом органов управления и с максимальным использованием площади экрана под изображения. Это делает реализованные с его помощью определители доступными не только для специалистов, но и для неквалифицированных пользователей. Программа PICKEY8 имеет ряд преимуществ перед известными аналогичными программами: при выборе признаков сама показывает поясняющие их рисунки, не заставляя пользователя вызывать их специально; при выходе определяющего на любую группу таксонов есть возможность окончания определения просто по тотальным рисункам, без использования конкретных признаков. - **Зоологический институт РАН, Университетская наб., 1, 199034, Санкт- Петербург.**

J. DIEDERICH[1], R. FORTUNER[2] & J. MILTON[1]. Concepts and Approach for a General Identification System. Many methods and systems for computer identification have been proposed in the last 30 years, but their usage remains very limited. It seems that this is due to several types of problems: Reliance on a single approach per tool (e.g., elimination in the case of multi-entry keys). Use of unreliable data (e.g., by dichotomous and multi-entry keys). Black box aspect of some approaches (e.g., neural networks) . Use of unfamiliar principles (e.g., Bayesian systems). Large amount of data entry needed to use some ID tools (e.g., statistic-based tools). Slowness of the tool compared to a printed key. Large amount of work needed to create the database. The data gathered for one tool cannot be used by other tools. Database not kept up to date. Lack of freedom for the user who must use the characters selected by the author of the tool. Lack of freedom for the user who must obey the machine (traditional expert systems). *Genisys* is a long term theoretical project that aims at defining the concepts for a general identification system able to overcome these difficulties. This it does mainly by two major innovations: 1 - The database should be created from the literature and include all the characters described for all the species in a group. Published data lacks uniformity, but uniformity can be restored by using a character format that is both uniform and representative. A tool (Terminator) has been designed (not currently available) to extract characters from published description and put them into a Genisys database. A version of the Terminator has been prototyped and tested demonstrating the feasibility of the task. Improvements and redesign would depend on funding. 2 - The ID tool must be in fact a set of tools, each designed to help the identifier do one of the possible tasks in an identification session (elimination, similarity, dissimilarity, instant recognition, statistics, probability, fuzzy logic, non-morphological approaches, etc.). Other tools could export Genisys data into a different format (e.g., data matrix or Delta-coded data) so that existing tools could be used as well as the tools developed for Genisys. The problems listed above would be solved as follows: Genisys will be a set of tools helping the user with many different approaches, including, but not limited to, elimination. Elimination will use reliable characters only; reliability of all characters will be evaluated based on metadata. Neural networks, Bayesian systems, etc., if included, will be only some of several available approaches. The set of tools will include some approaches that need very little input from the user (but input-intensive approaches will be

available as well, if needed). Fast identification (e.g., by instant recognition) will be supported. The database will be created from published data. A single database will be used by all the tools: it will be worth it to keep it up to date. The user will be free to use any set of characters. The user will be in charge of the identification process. The *Genisys* concepts were first developed for nematodes (Nemisys project) but soon expanded to identification in general. *Genisys* is currently a set of high level principles and specifications for biological databases and identification. A summary of these principles and a list of articles already published on *Genisys* can be seen in the following Web site: http://math.ucdavis.edu/~milton/genisys.html. These principles have not yet been implemented, but the project is seeking funding. - **[1]University of California, Department of Mathematics, Davis, California 95616, USA, Tel. JD: (1) (530) 752-0892, JM: (1) (530) 752-3657; Fax (1) (530) 752-6635, E-mail dieder@math.ucdavis.edu & milton@math.ucdavis.edu; [2]11 place Frèzeau de la Frèzelliôre, 86420 Monts sur Guesnes, France, Tel. (33) 5 49 22 87 18; Fax (33) 5 49 22 74 10; E-mail fortuner@wanadoo.fr**

Дж. ДИДЕРИЧ[1], Р. ФОРТЮНЕР[2], Дж. МИЛЬТОН[1]. Концепции и подход к общей идентификационной системе. Несмотря на множество идентификационных систем предложенных за последние 30 лет, их использование ограничено. Видимо, это связано со следующими проблемами: Уверенность в одном подходе (например, элиминация в случае многовходовых ключей). Использование ненадежных данных (в дихотомических и многовходовых ключах). Подход черного ящика (в нейронных сетях). Использование непривычных принципов (в Байесовских системах). Большое число данных требует использования некоторых инструментов ID (например, инструментов, основанных на статистике). Медлительность, сравнимая с использованием печатного ключа. Большая работа по созданию БД. Данные отобранные одним инструментом не могут быть использованы другим. БД устаревает. Отсутствие свободы пользователя, который должен использовать признаки выбранные автором ключа. Отсутствие свободы пользователя, связанное с диктатом машины (традиционные экспертные системы). *Genisys* - долговременный теоретический проект, направленный на определение концепций общей идентификационной системы, способен преодолеть эти трудности. Это достигается включением 2 главных инноваций: 1 - БД должна быть создана по литературным данным и включать все признаки описанные для всех видов группы. Опубликованные данные нестандартизированы, однообразие может быть воссоздано с помощью представительного и стандартного формата признаков. Создан инструмент (Терминатор) для изъятия признаков из публикаций и введения в БД. Версия Терминатора была испытана на выполнимость задачи. 2 - Инструмент ID в действительности должен быть набором инструментов, каждый из которых предназначен для помощи в одной из нескольких задач идентификационной сессии (элиминация, сходство, несходство, моментальное распознование, статистика, вероятность, логика, неморфологический подход, и т.д.). Другие инструменты могут экспортировать данные *Genysis* в различные форматы (например, матрицу данных или формат Delta), чтобы можно было использовать существующие данные и данные приготовленные в *Genisys*. Вышеперечисленные проблемы могут быть решены следующим образом: *Genisys* будет набором инструментов для различных подходов, включая элиминацию. Элиминация будет использовать только надежные признаки, надежность признаков будет оценена по метаданным. Нейронные сети, Байесовские системы, если будут включены, явятся только некоторыми из употребляемых подходов. Набор инструментов будет нуждаться в небольшом числе вводов пользователя (но и интенсивный ввод в необходимых случаях возможен). Быстрая идентификация (например, узнаванием) будет поддерживаться. БД будет создана по литературным данным. Одна БД будет использоваться всеми инструментами. Пользователь будет свободен в выборе любого набора признаков, он будет управлять идентификационным процессом. Концепция *Genisys* впервые разработана для нематод (проект *Nemisys*), но вскоре распространится на идентификацию вообще. В настоящее время это система принципов и спецификаций для биологических БД и идентификации. Сводка принципов