

CDS/ISIS for Windows — новый эффективный инструмент для текстовых баз данных. Опыт Теплофизического центра ИВТ РАН

М.С.Трахтенгерц

Потребности создания программных систем для автоматизированной обработки библиографической и других видов текстовой информации возникли сразу после появления ЭВМ, доступных для сравнительно широкого круга пользователей, в основном государственных учреждений, библиотек, университетов. Одним из первых удачных подходов к решению этой проблемы была разработка в 1975 г в ЮНЕСКО автоматизированной информационно-поисковой системы CDS/ISIS (Computerised Documentation Service / Integrated Set of Information Systems — Документальная Компьютеризированная Служба / Интегрированное Семейство Информационных Систем).

Не смотря на то, что в то время развивались системы управления базами данных (СУБД) различных структур (реляционные, сетевые, иерархические), ни одна из них не стала столь удобной, как ISIS, для нужд библиотек, архивов, научно-информационных центров и других подобных учреждений, в которых собираются и используются для информационного обслуживания потребителей большие объемы данных в виде текстов. СУБД действительно проявили себя эффективными инструментами для управления быстроменяющимися процессами в бизнесе, с помощью которых удается вносить оперативные изменения в базы данных. Но они не удобны в случае постоянных связей в объектах (типа автор – название произведения), разнообразия формулировок запросов потребителей, разнообразия требуемых форм представления ответов на запросы, библиографических указателей и других информационных продуктов.

Подход ЮНЕСКО при разработке CDS/ISIS оказался прагматичным и эффективным. По мере развития вычислительной техники были последовательно созданы версии для IBM/360, IBM/370 и первых персональных ЭВМ под управлением ДОС. Последняя из перечисленных, называемая также microISIS, использовалась и продолжает применяться до сих пор в различных организациях более 100 стран мира. Необходимость сохранения и доступности для использования больших объемов накопленной информации была учтена в последней версии CDS/ISIS for Windows — WinISIS, возможности которой для пользователей были существенно расширены. Среди

них — вызов полнотекстовых документов, открытие мультимедийных файлов (изображения и звук, что важно, например, при компьютеризации музеев), выход на источники в Интернет.

WinISIS в её современном виде представляет собой систему для хранения и поиска структурированной текстовой информации весьма общего характера. Имеются также ограниченные возможности обработки численных данных, если они представлены как текст, с использованием алфавитной сортировки и понятий старшинства для соотношений «больше», «меньше», «равно». Одно из важнейших преимуществ WinISIS является то, что она может управлять неограниченным числом баз данных самого разного содержания, причем её некоторые внутренние элементы также представлены в виде таких же баз данных. Последнее свойство позволяет сравнительно легко настроить систему на язык какой-либо страны или народа, т.е. провести так называемую её локализацию.

Беспрецедентно быстрое и широкое распространение WinISIS в мире в значительной мере поддерживается политикой ЮНЕСКО по её совершенствованию, бесплатному распространению, документированию на разных языках, организацией форума в интернет для обмена опытом, проведением всемирных конгрессов пользователей. Она удобна в эксплуатации. После её установки и настройки специалистом, работа на ней не требует специальных компьютерных знаний ни от библиотекаря, ни от читателя, ни от научного работника при использовании на собственном компьютере. Графический интерфейс системы достаточно очевиден. Сейчас на рынке систем для подобной обработки текста нет ничего, что бы могло сравниться с WinISIS по эффективности. Однако в России система WinISIS распространена ещё совершенно недостаточно, очевидно, из-за недостаточной информированности потенциальных потребителей.

Структурирование данных в понятиях WinISIS имеет простой смысл — текст документа заносится в поля и подполя записи, определяемые системой по меткам (тэгам), которым создатель БД придает определенный смысл. Например, в БД библиографического назначения будут выделены поля авторов, названия публикации, года выхода в свет и т.д. В исследовательской химической лаборатории список полей может быть расширен для названий веществ, типов реакций, места выполнения исследований, фамилии ученого, описания результата и т.д. Технически же все поля обрабатываются формально одинаково и это — выделение, сравнение, извлечение знаковых (буквенных) последовательностей, их сортировка в получаемых словарях и т.п. без какого-либо учета тематической специфики конкретной базы данных. Указанных операций достаточно, чтобы проводить поиск и вывод информации с использованием поисковых

терминов и булевых выражений. Этим достигаются необычайно широкие возможности применения АИПС CDS/ISIS для различных целей.

Все поля могут быть переменной длины, что существенно экономит используемую память. Это важно, например, при распространении БД на CD-ROM.

Теплофизический центр ИВТ РАН (ТФЦ) начал использовать АИПС CDS/ISIS с начала 1980-х годов. Освоение технических и системных программных средств на базе ЕС ЭВМ было осуществлено с помощью МЦНТИ, который в то время выполнял функции регионального центра ЮНЕСКО по распространению программной системы CDS/ISIS [1]. После появления IBM PC был осуществлен переход на microISIS также при содействии МЦНТИ. В 2004 году ТФЦ перевел свои информационные фонды на ISIS for Windows самостоятельно, используя общедоступные дистрибутивные каналы ЮНЕСКО.

Основной информационный продукт ТФЦ — база данных по теплофизическим свойствам веществ ТЕРМАЛЬ.

В тематику свойств БД включены следующие свойства:

- термодинамические;
- транспортные (неэлектромагнитные);
- упругие;
- электрические (проводимость, диэлектрическая проницаемость и термоэдс);
- оптические (константы).

Рассматриваются следующие вещества:

- все элементарные вещества и индивидуальные неорганические соединения;
- ряд технически важных смесей, а именно: смеси и сплавы щелочных металлов, воздух и смеси его основных компонентов (N_2 , O_2 , Ar, CO_2 , H_2 , а также смеси CO_2 -He, H_2 -Ne);
- простейшие органические соединения, а именно:

а) Углеводороды - предельные и непредельные - от C_1 до C_6 и все их галогензамещенные метанового и этанового ряда (фреоны).

б) Оксиды и сульфиды углерода, карбиды и карбониды элементов.

с) Цианистые соединения.

д) Простейшие органические кислоты (до уксусной) и все их соли.

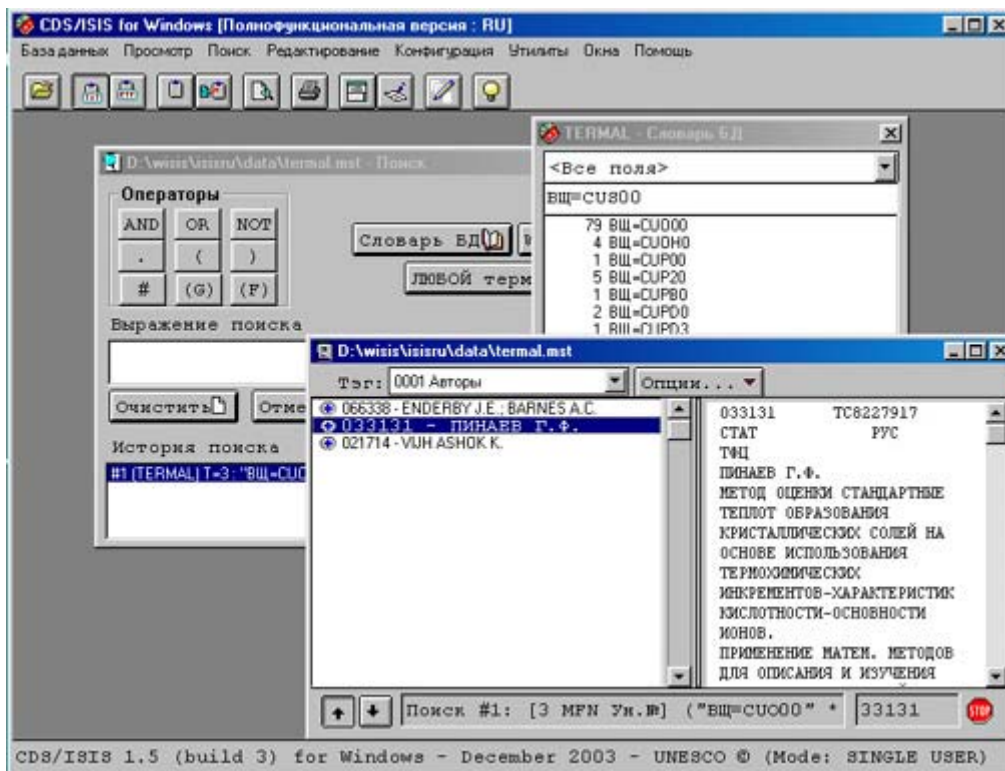
е) Некоторые технически важные органические соединения — ацетон, простейшие спирты - метиловый и этиловый.

Каждое вещество с целью унификации кодируется по специальным правилам, исключающие возможность обозначения разных веществ одинаковым ключевым словом, и обратное.

Для каждой работы — статьи, монографии, обзора и т.п., составляется документ (запись), состоящая из именованных полей. При подготовке документа вместо имени поля используется его цифровая метка — тэг. В БД ТЕРМАЛЬ определены следующие поля:

Тэг	Имя поля	Указ. Пп	Тэг	Имя поля
001	Авторы	а	017	Номер ТФЦ
002	Назв.рус.		018	Вид док.
003	Назв.ориг		019	Язык
004	Источ.	ab	020	Год публикации
005	Конференция		021	Референт
006	Реферат		022	T_ниж
007	Находится в		023	T_вер
008	Свойства	а	024	P_ниж
009	Носитель	а	025	P_вер
010	Фаза	а	026	Имя полнотекстового
011	Фазовый пер.	а		электронного источника
012	Тип свойства	а		Здесь буквы справа — указатели разбиения поля на повторяющиеся подполя
013	Физич. поле	а		
014	Вид работы	а		
015	Хим.формула	а		
016	Класс вещества	а		

В качестве примера удобного интерфейса новой системы приведем окна, используемые при поиске в БД ТЕРМАЛЬ.



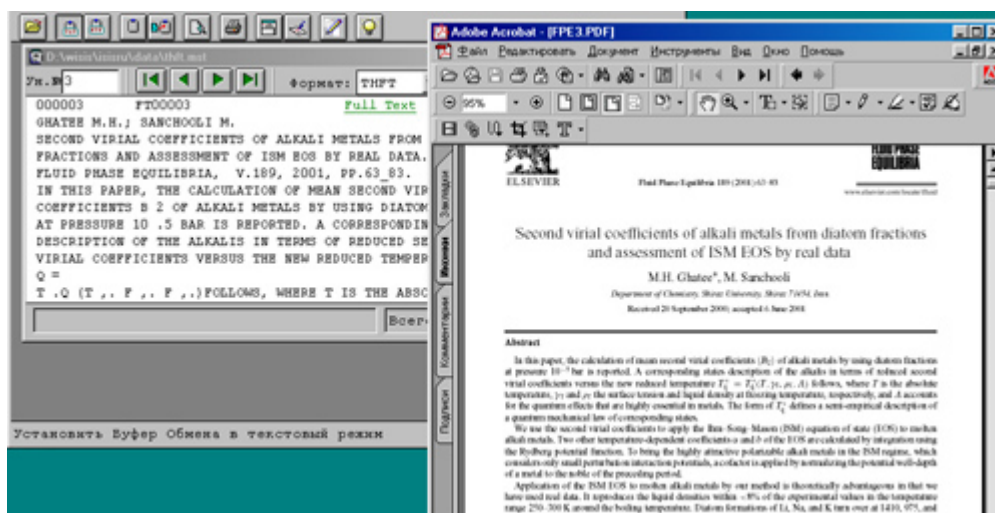
Фиг.1 Формы интерфейса поиска данных в ТЕРМАЛЬ.

Порядок действий пользователя или оператора системы может быть следующий:

Первым шагом при поиске вызывается форма для составления запроса, в окне которой «Выражение поиска» можно сразу записать подготовленную его формулировку. Однако, чтобы убедиться, что необходимые для запроса ключевые слова присутствуют в БД, целесообразно сначала ознакомиться со словарем ключевых слов, кнопка вызова которого находится на той же форме. Двойной щелчок мыши на термине переносит его в поле поиска, а кнопки операторов устанавливают необходимые отношения между ними, если терминов несколько. Обращение к словарю показывает также, сколько в БД имеется документов с выбранным термином, что позволяет оценить количество данных на выдаче. После команд «Выполнить» и «Показать» открывается форма со списком отобранных документов (приводятся авторы) и окном для просмотра указанного из них. В файле результатов можно сохранить как все подобранные документы, так и только специально помеченные.

Столь же удобны и понятны формы и для других операций на БД ТЕРМАЛЬ.

В новой версии ТЕРМАЛЬ реализована удобная для пользователей возможность получения полного текста документа в электронном виде непосредственно в ходе проведения поиска.



Фиг.2 Вызов полного текста документа из результата поиска в БД ТЕРМАЛЬ.

Для этой операции в ТЕРМАЛЬ назначено поле с тэгом «026». Если полный текст документа, описание которого найдено при поиске, имеется в электронном виде, его файл помещен в специальную директорию и ссылка на него внесена в поле 026, то в окне найденной записи с помощью условного оператора появляется надпись “Full Text” (выделена зеленым цветом). Щелкнув по ней стрелкой мыши, пользователь запускает программу открытия файла, выбираемую по его расширению, и полный текст документа появляется на экране. На примере в Фиг.2 мы видим окно программы Acrobat Reader, открывшей документ в формате PDF. Теперь его можно сохранить, передать по электронной почте или распечатать.

ISIS предоставляет и многие другие возможности, не затронутые в настоящей статье. Освоенная Теплофизическим центром новая технология работы с базами данных показала свою эффективность и может быть рекомендована для широкого использования как учреждениями, так и лично специалистами.

В завершение статьи укажем, что желающие получить программу CDS/ISIS for Windows должны самостоятельно обратиться в UNESCO для регистрации и получить пароль для её скачивания (бесплатно). Веб адрес:

http://portal.unesco.org/ci/en/ev.php-RL_ID=5330&URL_DO=DO_TOPIC&URL_SECTION=201.html

К сожалению, локализация программы на русский язык содержит грубые ошибки (в частности, неправильное кодирование русских букв), и ТФЦ пришлось самостоятельно устранять их. По условиям лицензионного соглашения ТФЦ не имеет права передавать кому-либо эту программу, но мы готовы поделиться этой разработкой с теми, кто имеет легальную программу от ЮНЕСКО. Фактически для нового получателя ISIS это сводится к замене примерно полутора десятков файлов на новые с теми же именами. Мы предполагаем разместить их на нашем портале www.thermophysics.ru, а также предложим их ЮНЕСКО для внесения в распространяемый пакет.

Литература

1. Трахтенгерц М.С., Нун А. Создание тематической базы данных ТЕРМАЛЬ с помощью CDS/ISIS/ЕС. — НТИ, Сер.2, 1989, №11, С.16.