НТИ, сер.2, Информ. процессы и системы, 2006, №7, с.28-31

Трахтенгерц М.С.

Технология подготовки информации для баз данных в обменном формате ISO 2709

Введение

Как известно, одной из самых трудоемких работ, проводимых при полноценном функционировании библиографических и других информационно-поисковых систем (ИПС), является регулярное пополнение их информационных ресурсов новыми документами. В этой статье рассматриваются вопросы техники и организации подготовки входящего потока информации на примере информационно-поисковой системы CDS/ISIS for Windows (далее ISIS), используемой в Теплофизическом центре Института высоких температур РАН для базы данных ТЕРМАЛЬ по теплофизическим свойствам веществ [1].

В ISIS [2] это можно сделать двумя способами:

— последовательно вводить поля документа с помощью системного интерфейса и сохранить законченный документ;

— импортировать массив документов, имеющих формат стандарта ISO 2709 для обмена данных.

Второй способ значительно эффективней, но для этого нужно, чтобы массив документов был бы ранее кем-то подготовлен, например, использовался на другой ИПС и преобразован в нужный формат.

Следует учитывать еще одно немаловажное обстоятельство, связанное с глобальным расширением пользователей Интернет. С его помощью стали доступны, причем в электронном виде, многие документы (статьи, книги, материалы из периодических изданий и т.д.), которые должны отражаться в соответствующих базах данных. Это позволяет обойти ручной набор текстов в полях системного интерфейса и значительно уменьшить трудозатраты на ввод данных.

Интерфейс системы ISIS можно использовать и в этом случае, копируя строки из электронного документа сначала в буфер (clipboard) и затем из буфера в нужное поле интерфейса.

Но можно также некоторым определенным образом разметить электронный документ в отдельном редакторе и с помощью специальной программы получить эту информацию в формате обмена данных ISO 2709. Соединив подготовленные таким образом записи из разных сеансов работы и/или различных сотрудников службы

подготовки информации в достаточно большой файл, эти данные можно ввести в базу данных одним сеансом.

Сейчас не представляет проблемы также использование подобной схемы при исходных документах на бумажных носителях. С помощью сканера и программы распознавания текста легко получить в электронном виде необходимую для ввода в базу данных часть документа (авторы, название статьи, реферат и т.д.) и использовать ее для сокращения ручного набора текста, а затем и записи для базы данных в формате ISO 2709.

Имея в виду изложенное, в Теплофизическом центре (ТФЦ) была разработана схема разметки текста, соответствующая структурам баз данных в системе ISIS, и создана программа WinISO (Beta-версия), производящая преобразование размеченной записи в обменный формат ISO 2709. Поэтому описания разметки исходного текста и программы WinISO иллюстрируются примерами из реальной БД по теплофизическим свойствам чистых веществ ТЕРМАЛЬ.

Схема разметки текста

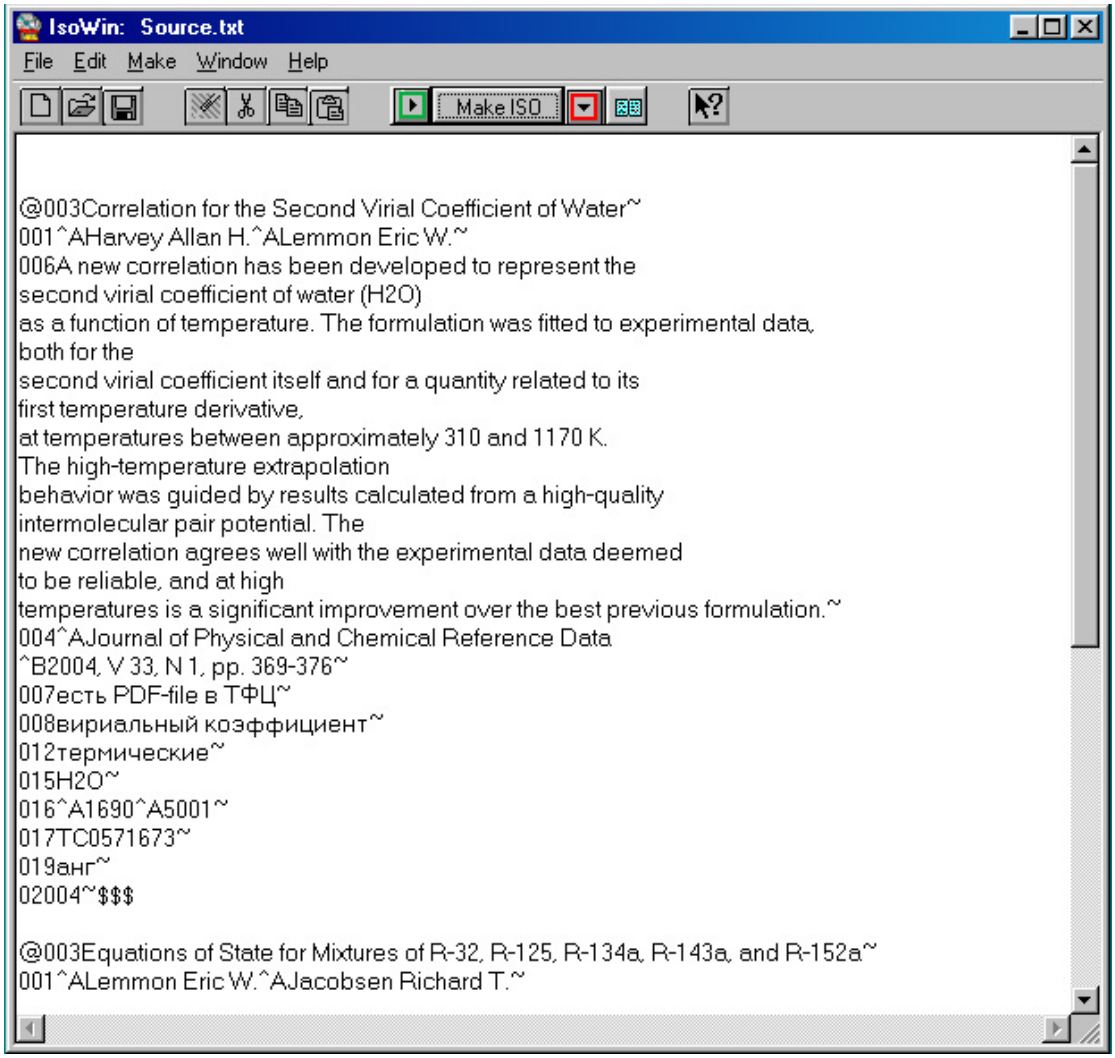
Поскольку каждая из баз данных в системе ISIS имеет свою структуру записей, вводимую с помощью Таблицы описания полей (FDT), она должна быть отражена в процессе разметки исходного текста. Это означает, что при разметке должны использоваться те же по написанию и смыслу метки полей — тэги, и указатели подполей, если они присутствуют в полях. Таким образом обеспечивается совместимость вводимых новых данных и базы данных, для которой они предназначены. Использование всех меток FDT не обязательно, приводятся только заполненные поля.

Например, FDT для БД ТЕРМАЛЬ имеет следующую структуру:

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

Здесь буквы справа — указатели разбиения поля на повторяющиеся подполя

На Фиг.1 показано редакторское поле программы IsoWin с текстом, размеченным в соответствии со структурой записи БД ТЕРМАЛЬ.



Фиг.1 Пример разметки электронной записи в программе IsoWin для импорта в БД ТЕРМАЛЬ.

Правила разметки текста заключаются в следующем:

- Текст, подлежащий преобразованию в формат ISO 2709 заключается между первым символом @ и завершающей комбинацией символов $$$. Все, что находится вне этих символов, исключается из преобразования и может быть использовано для комментариев и рабочих указаний как в начале текста, так и между документами.
- Информация, подлежащая внесению в некоторое поле БД, начинается с тэга этого поля (состоящего из трех цифр в системе ISIS) и заканчивается символом тильда ~.

- В случае присутствия подполей они маркируются символом ^ с последующей буквой — указателем подполя в FDT.
- Если в поле имеется лишь один экземпляр подполя, его разделитель можно опустить.
- Несколько полей могут помещаться в одной строке текста.
- Последовательность полей в записи значения не имеет.

Как видно, эти правила просты и очевидны. В примере на Фиг.1 основная часть записи на английском языке была заимствована из внешнего электронного источника (поля 001 — 006), а другая часть (поисковые термины, названия веществ и свойств из тезауруса БД Термаль на русском языке и другие) были внесены специалистами-референтами на основании анализа содержания публикации в целом.

Программа IsoWin

Программа преобразования размеченного соответствующим образом текста в формат обмена ISO 2709 IsoWin работает в среде Windows 95 и выше. Она состоит из двух основных функциональных частей — собственного текстового редактора для разметки текста и конвертора текста.

Вообще говоря, в качестве редактора может быть использован также любой другой, который сохраняет текст в файле точно в таком же виде, каком он виден в окне этого редактора. Пример подходящего редактора — Microsoft Notebook. В размеченном файле не должно быть никаких символов, управляющих видом и размером используемых шрифтов, табуляцией и т.п., за исключением команд «перевод строки» – «перевод каретки». В соответствии с принятой разработчиками системы ISIS в ЮНЕСКО позицией, обеспечивающей сохранение прежних информационных фондов пользователей в кодировке ДОС (кодировка OEM), расширенные многоязычные кодировки Unicode в ней, и, соответственно, в этой программе не используются.

В Beta-версии редактор IsoWin может выполнять все типичные для текстовых редакторов операции, но для упрощения работы в внешними источниками имеет следующие особенности:

- Коды табуляции, попадающие в редактор при переносе документов из других источников, в ходе дальнейшей обработки игнорируются.
- Знаки переноса части слова на следующую строку при преобразовании устраняются, т.е. их не нужно редактировать

вручную. Последовательная цепочка пробелов также сокращается до одного.

Редактор имеет количественные ограничения:

- Длина строки текста не превышает 80 символов.
- Длина одного поля в документе не должна превышать 30 Кб.
- Длина файла, вводимого в окно редактора, также не должна ориентировочно превышать 30 Кб. При преобразовании в формат обмена длина файла увеличивается и может выйти за пределы допустимой, причем по отношению к исходной величине она становится тем больше, чем меньше длина каждого документа и, соответственно, больше их число в файле. В случае появления сообщения о нехватке памяти необходимо разбивать вводимый файл на части и затем соединять полученные в результате конвертирования.

После того, как размеченный файл подготовлен в поле редактора или внесен в него из другого приложения, преобразование его в формат обмена производится по команде GetISO (кнопка на панели или команда в меню Make). Полученный результат виден в том же окне редактора. Структура формата обмена здесь не обсуждается. Отметим только, что запись каждого документа начинается с длинной последовательности цифр, кодирующих его структуру, и таким образом можно выделить зрительно отдельные записи и получить некоторое представление о содержании файла. Минимальное редактирование результата, вообще говоря, возможно в режиме клавиатуры «замена» без изменения длин полей и подполей.

Как уже говорилось выше, тексты во внутренних записях баз данных в ISIS имеют кодировку DOS (OEM), а готовятся практически всегда в Windows. Поэтому конвертирование в обменный формат может выполняться программой в двух режимах — с преобразованием кодировки из Windows в DOS (ANSI to OEM) и без такого преобразования. Второй режим используется в том случае, когда исходный файл готовится в кодировке OEM. Надо отметить, что, хотя преобразование ANSI to OEM предусмотрено в системе ISIS, в случае русского языка оно работает неудовлетворительно. Лучше, когда для операции импорта данных в ISIS подается файл в кодировке OEM. Возможно, что такое же происходит и другими языками. В данной версии преобразование ANSI to OEM производится только для русского языка по собственной таблице соответствий. Буквы русского алфавита в исходном файле, кодированном в OEM в каком-либо другом редакторе, в редакторе IsoWin выглядят не читаемыми и не редактируются. Нормальный режим работы — набор и редак-

тирование текста в стандартном режиме Windows и конвертирование результата в кодировку OEM.

Алгоритм перекодировки ANSI->OEM и обратно выбран с учетом того, что документы в БД ТЕРМАЛЬ являются двуязычными — в одном документе, как правило, имеются записи на обоих языках. Это связано с тем, что большая часть релевантной информации опубликована (или отреферирована) в англоязычных источниках. В то же время поисковые термины для удобства отечественных пользователей приведены на русском языке. Это видно на примере документа в Фиг.1. Перекодировке подвергаются лишь буквы русского алфавита, поскольку латинские буквы и знаки препинания воспроизводятся правильно в обоих представлениях. Она производится с помощью строк соответствия для Windows SWin="ёЁТюабцдефгхийклмнопярстужвьызшэщчъЮАБЦДЕФГХИЙКЛМНОПЯРСТУЖВЬЫЗШЭЩЧЪ" и для DOS, которая здесь не приводится из-за её не читаемости. Последняя получается путем преобразования SWin в вид OEM и просмотра результата в редакторе для Windows. Такой алгоритм позволяет достаточно просто произвести локализацию программы IsoWin для других языков, алфавит которых содержит буквы, отличающиеся от простых латинских. Сохранение интерфейса на английском языке не является помехой, поскольку он привычен в компьютерном мире.

Режим преобразования при конвертировании файла устанавливается в подменю Options меню File.

При выполнении команды GetISO производится автоматическое сохранение исходного текста в файле Source.txt в той же директории (папке), где находится программа IsoWin. Его можно сохранить также до выполнения команды конвертирования с помощью команды Save as в меню File. После выполнения конвертирования исходный файл в окне редактора заменяется результатом, который также автоматически сохраняется в файле Result.iso в той же директории. Буквы русского алфавита при этом выглядят не читаемыми. Этот файл можно также сохранить в другом месте командой Save as.

Файл результата можно просмотреть после обратного его преобразования в формат Windows с помощью команды AnsiToOem в меню Window или кнопки на инструментальной панели. Сохранять его в таком виде не имеет смысла, т.к. ISIS не воспримет его правильно, но можно использовать для обнаружения допущенных ошибок. Поэтому файл Result.iso сохранятся на прежнем месте неизменным. Неизмененный файл вызывается в окно редактора командой Result в меню Window или кнопкой на инструментальной панели (справа от кнопки GetISO). Если в конвертированном файле были обнаружены ошибки, то в окно редактора можно вызвать исходный

файл в последнем виде и внести необходимые исправления. Используется команда Source в меню Window или кнопка на инструментальной панели (слева от кнопки GetISO).

Остальные команды меню и соответствующие им кнопки типичны для простых редакторских программ и не нуждаются в специальных пояснениях.

Конечно, программа IsoWin может использоваться и для других систем управления базами данных (СУБД), принимающих информацию в формате ISO 2709 и кодировке ANSI.

Заключение

Программа IsoWin практически может стать основным инструментом рабочего места оператора подготовки информации для ввода в БД. В ТФЦ работа организована таким образом, что готовые документы от операторов поступают к администратору БД, который составляет массивы для очередного пополнения БД через операцию импорта. Он же осуществляет контроль правильности записей, в частности, путем просмотра дополнений в файл поисковых терминов. Ошибки в терминах легко обнаруживаются по их смещениям в списке, упорядоченном по алфавиту. Ошибочные документы либо корректируются администратором ввода данных, либо возвращаются операторам.

Освоенная Теплофизическим центром ИВТ РАН новая технология работы с базами данных ISIS for Windows, которая распространяется UNESCO, показала свою эффективность и может быть рекомендована для широкого использования научными институтами, библиотеками, а также отдельными специалистами, имеющими дело с большими объемами информации в своей работе.

Теплофизический центр предоставляет возможность свободно воспользоваться программой IsoWin полностью работоспособной версии Beta всем заинтересованным организациям и частным лицам. Для этого она выставляется в Интернет на портале www.thermophysics.ru и сопровождается необходимой документацией и информацией для контактов с разработчиком.

Литература

1. Трахтенгерц М.С. ТЕРМАЛЬ — база данных в Теплофизическом центре ИВТ РАН. Доклад на XI-ой Российской конференции по теплофизическим свойствам веществ, 4 – 7 октября 2005 г., Санкт-Петербург. См. также www.thermophysics.ru

2. http://www.unesco.org/isis/files/winisis/windows/doc/english/WINISIS15.pdf