СЛОВАРНАЯ СТАТЬЯ ЭЛЕКТРОННЫХ СЛОВАРЕЙ GLOBUS SOFTWARE HOUSE

A DICTIONARY ENTRY OF GLOBUS SOFTWARE HOUSE ELECTRONIC DICTIONARIES

С. А. Коваль

Филологический факультет СПбГУ, Санкт-Петербург

skoval@online.ru

И. Н. Ларченков

Globus Software House (холдинг «Светон»), Санкт-Петербург

soft@globus.spb.su

Описана обобщенная структура словарной статьи электронного словаря, разработанная в результате анализа наиболее массовых типов печатных словарей. Структура отражена в языке разметки DML и поддерживается программной средой MegaDictionaries, которая может служить дружественным редактором данных для этого формата.

Введение

За долгую историю создания словарей человечество выработало некоторые каноны и правила, по которым строится словарь, определяется состав информации внутри словаря и способы представления этой информации. Во многом такие нормы диктуются ограничениями, связанными с объемом печатного словаря и способами поиска информации в нем. С появлением электронных словарей ограничений на объем словаря практически не существует, а возможности поиска информации просто несравнимы с традиционными методами. При этом состав словарной статьи у большинства электронных словарей практически ничем не отличается от того, что использовался ранее в традиционной лексикографии. Однако, как нам кажется, новые возможности должны найти свое отражение в составе словарной статьи создаваемых электронных словарей.

Специалисты компании Globus Software House проанализировали большое количество классических словарей, что позволило создать обобщенную словарную статью, используемую сейчас в системе электронных словарей компании. Анализировались в основном двуязычные словари, толковые словари, словари синонимов и антонимов. Из анализа не исключались словари, по форме и содержанию являющиеся авторскими и отражающие взгляды конкретного автора (или авторов) на структуру словаря. В основном это небольшие словари, отражающие узкую специфику той или иной тематической области. В ходе анализа мы пытались выделить общие свойства всех этих разных по духу и содержанию словарей.

В результате была разработана обобщенная структура словарной статьи и, что не менее важно, создан язык ее разметки (DML – Dictionary Markup Language). Мы отдаем себе отчет в том, что в предлагаемую нами схему невозможно вписать все существующие словари, и что подобная схема неприемлема для всех типов словарей любой сложности и направленности. Язык DML преследует следующие сугубо утилитарные цели:

1) При известных и не слишком жестких ограничениях на состав словарной статьи обеспечить эффективный и удобный способ управления лингвистическими данными.
2) Обеспечить возможность коллективной работы над словарем, включая работу в удаленном режиме.
3) Создать механизм, позволяющий обмениваться данными словарям различных производителей.
4) Создать условия, при которых словари, созданные различными авторами, могут быть использованы не только по своему прямому назначению (поиск и показ информации), но и являться составной частью систем автоматического или полуавтоматического анализа текста и/или какой-либо другой обработки текстовой информации.

Структура словарной статьи Globus Software House

В обобщенную словарную статью нами были включены следующие поля.

- Базовая информация

a) Ключ записи

В качестве ключа (заголовка словарной статьи) может выступать как слово, так и словосочетание.

b) Фонетическая информация

Как правило, поле содержит транскрипцию ключа записи.

c) Набор признаков

Ключу записи можно приписать различные признаки, определяющие статью. Каждый признак представляет собой закодированное в виде символа (буквы или цифры) свойство (часть речи, род, семантический класс и т.д.).

d) Краткий комментарий или помета

Авторские комментарии (пометы), относящиеся целиком к словарной статье, например, указание на предметную область.

- Группы данных с информацией о переводах

Набор переводов к словарной статье. К одной словарной статье может быть приписано несколько групп переводов, если автор словаря считает, что имеющиеся переводы должны быть объединены по грамматическим или иным особенностям. Каждая из групп имеет следующие поля.

a) Краткий комментарий к переводам группы

Авторские комментарии к группе переводов. Рекомендуется для кратких помет типа указаний на предметную область, часть речи, грамматические особенности и т.д.

b) Список переводов

Набор переводов, составляющих данную группу.

c) Примеры использования переводов из приведенной группы

Примеров может быть несколько или не быть вообще. Информация на двух языках в примере может быть разделена специальными символами-разделителями.

Отметим, что синтаксис *языка DML позволяет определять списки групп для всех поддерживаемых типов данных, если эти типы определяются как группа.*

- Группы связанных записей

В виде таких записей рекомендуется представлять устойчивые выражения, в которых участвует ключ словарной статьи, или аналогичную информацию. В каждой такой записи может быть сгруппировано несколько вариантов перевода, что имеет следующую структуру.

a) Идентификатор данных

Как правило, это оборот, в составе которого участвует ключ записи.

b) Комментарий к связанной записи

Краткие комментарии к связанной записи, образующей данную группу. Рекомендуется для кратких помет типа указаний на предметную область, часть речи, грамматические особенности и т.д.

c) Список переводов

Набор переводов, составляющих данную группу.

d) Примеры использования переводов, включенных в группу

Примеров может быть несколько или не быть вообще. Информация на двух языках в примере может быть разделена с использованием специальных символов-разделителей.

- Группы тематически связанных записей

Каждому ключу может быть приписана информация о других словах или выражениях, имеющих с ним смысловую связь. Используется при создании тематических и подобных им словарей. Информация может быть разделена на группы следующей структуры.

a) Идентификатор данных

Слово или выражение, выступающее в роли ключа группы.

b) Комментарий

Краткие комментарии к вложенной записи, образующей данную группу. Рекомендуется для кратких помет типа указаний на предметную область, грамматические особенности и т.д.

c) Список переводов

Набор переводов, составляющих данную группу.

d) Примеры использования переводов, включенных в группу

Примеров может быть несколько или не быть вообще. Информация на двух языках в примере может быть разделена с использованием специальных символов-разделителей.

- Комментарий к записи

Содержит форматированный текст (формат RTF) с информацией для ключа в целом. Текст может содержать данные об особенностях использования ключа или другую грамматическую информацию. Формат RTF позволяет осуществлять шрифтовое, стилевое или цветовое оформление данных.

- Толкование

Форматированный текст (формат RTF). Рекомендуется для хранения информации о значениях и оттенках значений слова.

- Синонимические ряды

Ключу может быть приписана информация о соответствующих синонимических рядах. Используется при создании словарей синонимов. Каждый ряд оформляется как отдельная группа данных и имеет следующую структуру.

a) Комментарий

Краткие комментарии к группе.

b) Синонимический ряд

Набор синонимов, составляющих группу.

c) Примеры использования или описание данных

Примеры или дополнительные авторские комментарии.

- Группы антонимов

Используются при создании словарей антонимов. Имеют следующую структуру.

a) Комментарий

Краткие комментарии к группе.

b) Список антонимов

Набор антонимов, составляющих группу.

c) Примеры использования или описание данных

Примеры или дополнительные авторские комментарии.

- Данные, определяемые пользователем

При необходимости составитель словаря может определить свой собственный тип данных, если требуется ввести и, впоследствии, управлять данными, которые не могут быть отнесены ни к одному из предопределенных типов. В рамках одной записи может быть определено несколько пользовательских типов (аспектов, в которых описывается ключ записи). Информация в пределах каждого пользовательского типа может иметь следующую структуру.

a) Имя типа (например, «Этимология»)

b) Комментарий

Краткий текстовый комментарий к данным этого типа, включенным в запись.

c) Данные

Основное содержание данных этого типа. Текстовая неформатированная информация.

d) Примеры

Примеры в виде неформатированного текста.

- Поля мультимедиа

Включают графическую и звуковую информацию, приписанную записи в целом.

Отметим, что *только ключевое поле является обязательным* в этой структуре словарной статьи. Все остальные группы и типы могут быть использованы по желанию автора словаря в любой комбинации. По нашим оценкам, до 80% широко используемых в настоящее время словарей укладывается в приведенную нами схему. Но только при создании электронного словаря поддержка всех или большинства вышеописанных типов позволит создать словари огромной информативности.

Язык разметки словарной статьи DML

Сама по себе идея языков разметки словарных статей далеко не нова. В том или ином виде, прямо или косвенно, такой язык присутствует практически в каждом электронном словаре. Чтобы не было разночтений в понимании того, что является языком разметки, договоримся, что под этим мы будем понимать такой способ представления информации, который полностью описывает структуру словарной статьи и может быть прочитан любым текстовым редактором.

По нашему мнению, языки разметки делятся на два класса по своим функциональным возможностям и назначению.

1) Языки, решающие, в основном, задачи визуального представления лексикографической информации.
2) Языки, созданные для логического управления информацией, находящейся в словарной статье.

Разработанный нами язык DML принадлежит ко второй группе. Все возможности языка DML направлены на структурное разбиение информации и практически не содержат данных о способе ее представления. Мы считаем этот подход наиболее интересным и перспективным, так как, предоставляя огромные возможности по использованию словарей для решения широкого спектра задач, он, при необходимости, способен решать и задачи, перекрывающие возможности языков первого типа.

В задачи данной статьи не входит сравнительный анализ различных языков разметки. Таких языков достаточно много и не все они открыты и опубликованы. Каждый из них решает свой спектр задач (как правило, в рамках одного проекта) и не может претендовать на какую-либо универсальность. Не претендует на роль стандарта и язык DML, поскольку, по нашему мнению, этот вид человеческой деятельности вообще едва ли подлежит стандартизации. DML является опубликованным, открытым и бесплатным средством логической организации лексикографической информации.

DML создан по образу и подобию известных теговых языков. Одним из правил, используемых при построении тега, является возможность по начальному тегу построить конечный тег блока. Например, для тега <ENTRY>, конвертор или компилятор языка DML в состоянии понять, что тег, которым закончится этот блок, будет выглядеть как </ENTRY>. Это позволяет программному обеспечению, работающему с форматом DML игнорировать не поддерживаемые теги без ущерба для остальной информации. Кроме того, это обеспечивает (в разумных пределах) совместимость форматов не только сверху вниз (от младшей версии к старшей), но и снизу вверх, что является чрезвычайно важным при использовании DML в долгосрочных проектах. Принцип, согласно которому конечный тег строится от начального добавлением косой черты после треугольной скобки, соблюдается во всех тегах языка DML без исключения.

Текущей версии формата DML присвоен номер версии 2.0. Ниже (табл. 1) мы приводим фрагмент сводной таблицы тегов языка этой версии. Полную таблицу тегов Вы можете найти на сайте компании

(www.megadict.ru) и в руководстве пользователя электронных словарей Globus Software House.

Начальный тег/Конечный тег	Описание
<ENTRY> </ENTRY>	Данные каждой записи располагаются между этими тегами. Эти теги можно считать разделителями между записями.
<KEY> </KEY>	Ключевое поле записи. Это поле является обязательным для всех записей.
<DOMAIN> </DOMAIN>	Теги определяют авторский комментарий к записи в целом.
<PHONETIC> </PHONETIC>	Теги содержат фонетическую информацию о ключе. Как правило, это транскрипция, однако DML формат не накладывает никаких ограничений на состав и свойства фонетической информации.
<MARKER> </MARKER>	Тег включает информацию о строке признаков, приписанных записи в целом. Максимальная длина строки признаков равна 25 символам.
<TRANSLATEDATA> </TRANSLATEDATA>	Контейнер тегов, включающий данные о переводах. Таких блоков данных, приписанных одному ключу, может быть несколько (группы переводов). Внутри этого контейнера могут располагаться только теги одной группы переводов.
<TDOMAIN> </TDOMAIN>	Тег определяет краткий комментарий автора словаря к текущей группе переводов. Используется только внутри контейнера <TRANSLATEDATA>
<TTRANSLATE> </TTRANSLATE>	Тег включает в себя информацию непосредственно о списке переводов текущей группы. Используется только внутри контейнера <TRANSLATEDATA>
<TSAMPLE> </TSAMPLE>	Содержит информацию о примере использования текущего перевода. Используется только внутри контейнера <TRANSLATEDATA>

Табл. 1. Теги языка DML 2.0 (фрагмент сводной таблицы)

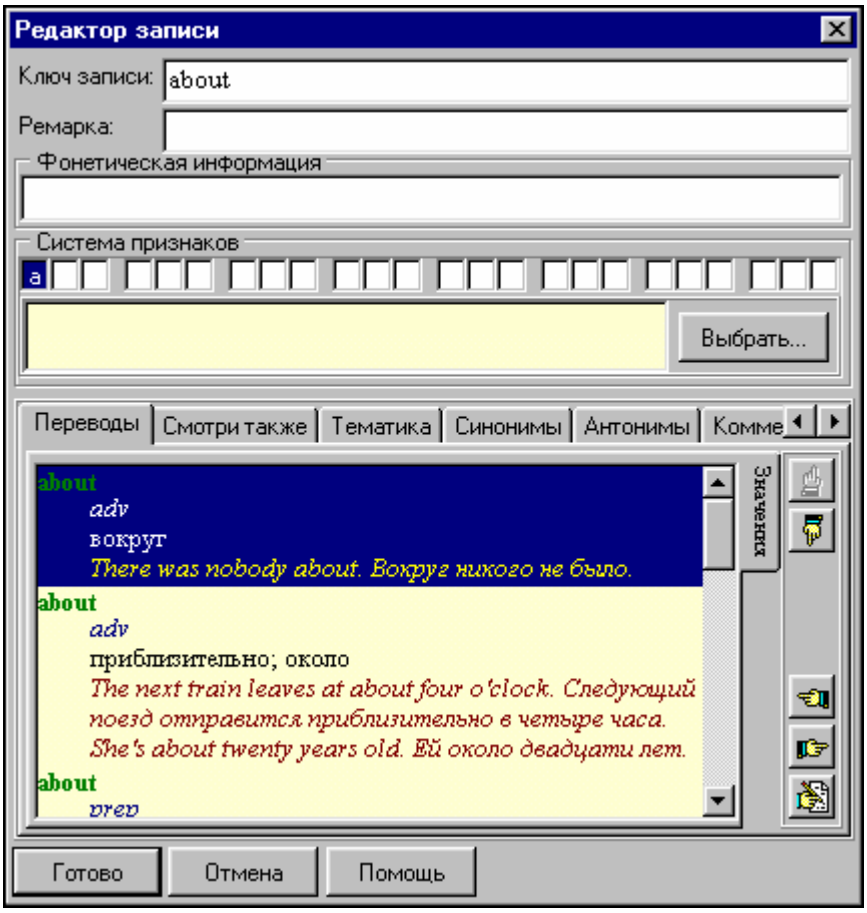

Рис. 1. Редактор DML данных

Любой словарь, созданный по правилам языка DML, может быть без каких-либо изменений конвертирован во внутренний формат электронного словаря Globus Software House или использован каким-либо иным образом (например, для преобразования данных в форматы словарей других производителей).

Электронные словари Globus Software House (распространяемые под торговой маркой MegaDictionaries) позволяют осуществлять целый ряд важных операций над данными в формате DML. Прежде всего, стоит упомянуть компилятор с языка DML, встроенный во все версии этого программного обеспечения. Алгоритм диалогового пополнения словаря позволит редактировать словари в дружественной диалоговой среде (см. рис. 1), даже не изучая теги языка DML. При этом автор словаря имеет возможность экспорта данных из внутреннего формата в открытый формат DML. При наличии соответствующих прав пользователь пакета MegaDictionaries может экспортировать словарь в стандартные форматы RTF и HTML, что позволяет осуществлять предпечатную подготовку словаря к классической публикации на бумажных носителях. Таким образом, формат DML обеспечен мощным и удобным средством для управления данными в этом формате.

Изначально язык DML проектировался таким образом, чтобы написание всевозможных фильтров или конверторов не составляло большой технической проблемы. Тем не менее, для популяризации формата и облегчения работы с ним компания Globus Software House в рамках проекта «Открытый код» (Open Source) выпускает в свободное обращение исходные коды своих конверторов из нескольких текстовых форматов в формат DML.

Кроме того, для создания электронных словарей, которые могут быть использованы при решении широкого спектра научно-исследовательских задач (в частности задачи статистического анализа текста), в рамках Globus Software House планируется разработать открытую систему лингвистических признаков и выпустить первый словарь с использованием этой системы в рамках концепции Open Data (т.е. непосредственно в текстовом формате DML). Все словари, распространяемые под лицензией Open Data, могут быть свободно и бесплатно использованы в любых некоммерческих проектах.

Заключение

Мы надеемся, что предлагаемые нами инициативы еще более увеличат спектр применимости электронных словарей вообще и словарей компании Globus Software House в частности, а концепция Open Data заинтересует научные коллективы, занимающиеся вопросами автоматической обработки и анализом текстовой информации.