ИСПОЛЬЗОВАНИЕ МИКРОДАННЫХ В КОМПЬЮТЕРНОЙ ЛЕКСИКОГРАФИИ

Терешко Анна Витальевна

УО «Брестский государственный университет имени А.С. Пушкина»

Современная компьютерная лексикография предполагает использование языков разметки электронного текста в качестве инструментария задания необходимых структур и отношений между элементами словарных статей и всего словаря с помощью наборов дескрипторов. Разработка данного инструментария является актуальной лексикографической задачей.

Цель работы – оптимизация структурно-семантической разметки электронных словарей, выступающих в качестве структурных компонентов сетевых лексикографических сервисов.

Научная новизна работы заключается в предложении и реализации структурно-семантической разметки электронных словарей на основе микроформатов и микроданных.

Объектом исследования являются электронные словари, предметом исследования выступает структура электронных словарей. В качестве методов исследования используются лингвистическое моделирование и конструирование на основе использования языков метатекстовой разметки.

Внешняя разметка (метаданные) содержит сведения о тексте и авторстве. Структурная разметка маркирует элементы мета-, макро- и микростуруктуры словаря. Техническая разметка определяет кодировку, даты обработки, источники. Стандартом для структурно-семантической разметки электронных словарей сегодня является XML (eXtensible Markup Language), на основании которого разработаны конкретные языки и форматы лексикографического описания (DSL, XDXF, TBX). Однако использование XML и его приложений во многих случаях неэффективно из-за их высоких трудозатрат работы с ними, что объясняется их функциональной избыточностью.

Предлагается использовать для структурно-семантической разметки лексикографической информации концепцию проекта Schema.org [1], объединившего ведущие поисковые системы в разработке единой схемы семантической разметки на основе эффективной структуризации информационных ресурсов. Предлагаемые Schema.org схемы представляют собой семантическую разметку, предназначенную для поисковых роботов, и

могут быть непосредственно проанализированы ими с целью извлечения и обработки информации о содержимом сетевых ресурсов, в том числе лексикографических. Основным форматом разметки в Schema.org являются микроформаты (дескрипторы поверх HTML), позволяющие описывать любую информацию на веб-страницах. Спецификация микроформатов представляет собой способ разметки содержания для определения любых типов информации.

Разметка микроформатами происходит непосредственно в html-коде страниц. Код микроформатов прост для написания в любом текстовом редакторе. Наиболее обобщенный тип сущности – это thing (нечто), у которого есть свойства: name (название), description (описание), url (ссылка) и image. Каждый тип информации описывает определенный тип элемента. В настоящее время поисковые системы уже поддерживают микроформатную разметку веб-страниц в результатах персоналий, событий, обзоров, товаров и множества других онтологий [2]. Стандарт schema.org предусматривает возможность добавлять свойства и дочерние типы для имеющихся типов сущностей.

Предлагается на основе микроформатов определить новую онтологическую сущность для описания собственно лексикографической информации. Таким микроформатом может быть объявлен XLD (XHTML Lexicography Data) – микроформат для пометки лингвистических метаданных. XLD можно использовать как на лексикографических интернет-ресурсах, так и для разметки любого словарного контента. Необходимо сообщить браузерам и поисковикам, что страница поддерживает XLD. Для этого в теге <head> веб-страницы надо добавить атрибут profile:

<head profile="http://gmpg.org/xld/17">

Для каждой гиперссылки на странице нужно добавить атрибут rel. Пример:

...

Значений атрибута rel может быть несколько, в таком случае они перечисляются через пробел.

Далее должен быть приведен список допустимых категорий атрибута rel с указанием их значений. Такие категории могут включать лексикографическую информацию об электронном словаре и его элементах с любой желаемой полнотой. Например, они могут содержать метаданные различных уровней лексикографической концепции и структуры, обеспечивая простую, изящную и очень гибкую словарную метаразметку.

Предлагаемая микроформатная разметка может быть реализована на любом естественном языке или быть мультиязычной, что важно для создания переводных словарей. Микроформаты представляют собой открытый формат и могут быть свободно использованы в любых лексикографических проектах. Словарные данные, размеченные микроформатами по стандарту schema.org,

становятся общедоступными и могут быть извлечены и использованы любыми сетевыми поисковыми сервисами.

1. Schema.org [Электронный ресурс]. – Режим доступа: http://schema.org/. – Дата доступа: 20.03.2018.

2. Концевой, М. П. Семантическая разметка интернет-ресурсов на основе микроформатов / М. П. Концевой // Информатика: проблемы, методология, технологии : мат. XIII межд. науч.-метод. конф., Воронеж, 7–8 февраля 2013 г. ; 4–х т. – Воронеж : ВГУ, 2013. – Том 2. – С. 186–189.