

УДК 681.3

А.О. Ордин (5 курс, каф. АиВТ), Н.В. Соколова, к.т.н., доц.

ПРИМЕНЕНИЕ ТЕХНОЛОГИИ SGML ДЛЯ ФОРМАТИРОВАНИЯ ДОКУМЕНТОВ В ЭЛЕКТРОННОЙ БИБЛИОТЕКЕ

Электронная библиотека (ЭБ) предоставляет средства как для хранения и поиска электронных документов, так и для их доставки в соответствии с требованиями пользователей. Мировой опыт показывает, что эти службы со временем становятся коммерческими, и немаловажным фактором в этой ситуации становится возможность формирования заказа на электронную доставку только части документа (диапазон страниц, глава или перечень глав, параграф и пр.). Следовательно, при создании ЭБ необходимо предусмотреть механизм структуризации документа и выбора его части в соответствии со структурным разбиением.

Представленное исследование было посвящено выбору технологии и разработке программного обеспечения (ПО) для структуризации электронных документов, выбору источника представления документов, выбору языка программирования, который удовлетворяет требованиям адаптивности к различным операционным системам.

Передовой технологией структурирования электронных документов на сегодняшний день является SGML. Технология SGML является основой для других спецификаций, то есть SGML - это не отдельный язык разметки, а целое семейство языков, позволяющих структурировать текст. XML, HTML - это наиболее известные примеры языков структурирования.

SGML позволяет:

- Отдельно описывать структуру данных, не привязывая это описание к способу отображения.
- Определять форму представления данных независимо от содержания.
- Преобразовывать содержание и структуру данных
- Согласовывать практику применения данных с национальным законодательством
- Описывать стандартные коммуникативные форматы данных, принятые в различных предметных областях.

В России для представления документов в электронном виде наиболее широко используется MS Word. В совокупности с простотой и удобством работы данный редактор быстро завоевал популярность среди пользователей и в настоящее время стал стандартом де-факто на средства обработки текстовой информации. Помимо чтения форматов большинства программ MS Word может сохранять результаты работы во множестве форматов текстовых файлов, а также и в формате HTML, позволяющем публиковать документы в Интернете. Однако осуществлять поиск нужной части документа среди больших массивов информации и получение выборочных частей документа представляется довольно трудоемкой задачей. По этой причине задача преобразования формата MS Word в SGML является актуальной.

В результате исследования был создано ПО, позволяющее конвертировать документы MS Word в формат SGML с заданными характеристиками (с возможностью подключения различных таблиц стилей). В докладе подробно рассматриваются функциональные возможности программной системы.

Созданное ПО было сопоставлено с аналогичными программными продуктами, используемыми в настоящее время для этих целей. Произведенный анализ показал имеющиеся недостатки и достоинства, которые будут учтены при модификации программной системы.

Заключительной фазой работ является внедрение данной технологии и ПО в Фундаментальной библиотеке СПбГТУ, где необходимо выполнять структурирование документов для организации службы электронной доставки.