Описание спецификации формата mnsXML и примеры его практического использования

А. Н. Желонкин, В. А. Романенко

Удмуртский государственный университет, Ижевск, Россия

Одним из путей интенсивного развития электронных полнотекстовых коллекций, созданных на транскрипционной основе, является взаимовыгодный обмен данными между системами хранения текстовой информации, предоставляющими пользователю различные функциональные возможности для поиска, выборки и анализа данных. Известно, что формат текста включает в себя кодировку символов текста и структуру текста в виде одной или нескольких разметок.

Формат mnsXML предназначен для стандартизации загрузки электронных текстов в ИАС «Манускрипт», хранения их вне базы данных в «универсальном формате» и для обмена текстами с другими электронными полнотекстовыми коллекциями текстов.

Формат должен обеспечить: (1) загрузку новых текстов в ИАС «Манускрипт» из разных исходных форматов, (2) выгрузку текстов из ИАС «Манускрипт» для передачи в другие коллекции, (3) создания резервных копий текстов, (4) выполнения массовых операций над текстами вне базы данных, в конечном счете – полноценный обмен текстами, метаданными, аналитической и лингвистической информацией между различными коллекциями.

Описание спецификации формата mnsXML

В исходном тексте может существовать несколько видов разметки (иерархий): геометрическая, морфологическая, синтаксическая, функционально-структурная, иерархия по писцам, по авторам и др. Геометрическая иерархия (деление на листы, страницы, строки, может присутствовать и более глубокое деление – до знаков) является базовой. В рамках каждой из иерархий выделяется минимальный тип единиц, как правило, являющийся словоформой или знаком.

Текст, выгруженный в формате mnsXML, может состоять из одного или нескольких файлов. Каждый файл представляет одну из разметок текста и

является xml-документом. Единицы в файле базовой иерархии имеют уникальные номера (идентификаторы), с помощью которых осуществляется связь с этими же единицами в других иерархиях. Иными словами, размеченный текст можно представить в виде текста с базовой разметкой, на которую наложены один или несколько дополнительных слоев разметки.

Разметка текста выполнена при помощи тегов и во многом основана на рекомендациях консорциума TEI (Text Encoding Initiative) (TEI: P5 Guidelines. URL: http://www.tei-c.org/Guidelines/P5/). В тех случаях, когда в TEI отсутствовали теги и их атрибуты для представления структуры некоторых иерархий (например, морфологической иерархии), семейство тегов было дополнено.

Каждый файл в формате mnsXML состоит из трех основных разделов:

1. заголовок – часть, содержащая метаинформацию о тексте (код иерархии, дата выгрузки, автор текста, дата создания и др.);

2. тело – основная часть, содержащая разметку текста;

3. окончание – часть, содержащая вспомогательные информационные материалы (список используемых шрифтов, словарей и др.).

Одним из примеров использования формата mnsXML является загрузка новых текстов. Загрузка текста выполняется в два этапа: сначала текст в исходном формате преобразуется специализированной программой-конвертором в файлы формата mnsXML, которые далее, на втором этапе, стандартной программой-конвертором загружается в базу данных ИАС «Манускрипт». Такое разделение дает возможность обрабатывать специальной программой все специфические особенности текста (кодировку символов, особые виды разметки), приводя их к стандартному виду, что позволяет сделать процедуру загрузки унифицированной и дает возможность управлять доступом к ней.

На вход программы-конвертора помимо исходного текста поступают таблица конвертации символов в кодировке Menaion (используемой в ИАС «Манускрипт» в качестве базовой) и файл-заголовок с метаданными текста. На выходе получаем набор файлов в формате mnsXML и журнал ошибок.

После проверки безошибочности формирования текста в промежуточном виде текст посредством конвертора из mnsXML загружается в базу данных.

Другим примером использования mnsXML может служить обмен текстами, мета- и аналитической информацией между различными коллекциями. Основными сложностями при таком обмене между двумя коллекциями являются различия в базовой разметке (геометрической иерархии), различные системы идентификации единиц и использование в каждой коллекции своих словарей.

Различия в базовой (и других) разметках должны быть сформулированы в виде правил конвертации между форматом данной коллекции (текста) и общим форматом обмена (например, mnsXML).

В настоящее время тестирование конвертеров осуществляется на материале произведений М. В. Ломоносова, которые сканируются, распознаются, сверяются с исходным печатным материалом Полного собрания сочинений, предварительно размечаются и сохраняются в текстовом формате ASCII. Подготовленные таким образом тексты с помощью подготовленного инструментария загружаются в ИАС «Манускрипт» и демонстрируются на портале (URL: http://manuscripts.ru/mns/portal.main?p1=31).

Подводя итоги, можно сказать, что данный формат необходим для работы с транскрипциями письменных памятников: он позволяет снизить трудозатраты, связанные с загрузкой/выгрузкой текстов, и – как следствие – обмениваться текстами с другими электронными коллекциями.

Благодарности

Работа выполняется при поддержке Российского гуманитарного научного фонда (РГНФ), проект № 07-04-12147в («Большой корпус русского языка XVIII в.»).

Description of the mnsXML specification format and examples of practical use

Alexey N. Zhelonkin, Vitaliy A. Romanenko

Udmurtia State University, Izhevsk, Russia

The purpose of this paper is to present a new storage format for ancient texts, which is intended for standardization of electronic text loading into the Manuscript information-analytical system, for storage and pre-processing outside of the database, and for interchange with other electronic full-text collections of ancient texts. The structure of the format, its capabilities, and areas of application will be described.