

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. М. Т. КАЛАШНИКОВА

Информационные технологии и письменное наследие

El'Manuscript-2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
И741

Изданы при поддержке гранта РФФИ (проект № 12-06-06061-г),
гранта РГНФ (проект № 12-04-14154-г) и в рамках реализации комплекса
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:

В. А. Баранов, д-р филол. наук, проф.

А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

Информационные технологии и письменное наследие [Текст] :
И741 материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября
2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ;
Ижевск, 2012. — 328 с.

ISBN 978-5-8021-1402-5

Сборник содержит материалы конференции, посвященной современ-
ным электронным средствам хранения, описания, обработки, исследования
и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978-5-8021-1402-5

© Петрозаводский государственный
университет, 2012
© Ижевский государственный технический
университет им. М. Т. Калашникова, 2012

ИСПОЛЬЗОВАНИЕ TEI И RDF В ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЕ «РУССКАЯ ЛИТЕРАТУРА XVIII века»¹

А. В. Андреев

Институт лингвистических исследований РАН, Санкт-Петербург

The conceptual data model of the information retrieval system «Russian Literature of XVIIIth century» challenges efficient indexing. The system is based on a set of TEI P4 conformant documents where available search criteria are derived from the mark-up. Initial relational approach proved to be very inadequate and was replaced by a RDF-based system. Standard RDF schemata are used as much as possible but are too coarse-grained in many cases, so a custom schema was devised.

Информационно-поисковая система «Русская литература XVIII века» обеспечивает поиск текстов и фрагментов текстов по ряду параметров, как-то: элементы библиографического описания, упоминаемые имена собственные, стиховедческие параметры и т.п. [Андреев и др., 2009] Основой ИПС является корпус текстов, размеченных в соответствии с рекомендациями TEI P4 (нами используется SGML-версия, а не XML, чтобы облегчить ручную разметку). Концептуальной особенностью ИПС является то, что набор поисковых параметров не задан жестко, а определяется используемыми тегами TEI.

Несмотря на относительно небольшой объем корпуса, вопрос эффективной организации инвертированных индексов встал уже на ранних этапах разработки. Первоначально нами использовалась реляционная модель, как наиболее распространенная, однако опыт практического использования системы выявил следующие слабые места:

- требование нормализации модели накладывает ограничения на используемую разметку, либо приводит к чересчур громоздким схемам БД
- динамическая природа набора параметров плохо сочетается с необходимостью иметь жесткую схему БД
- областью действия параметра может быть как текст, так и произвольный его фрагмент, что также плохо выражимо в реляционной модели
- обновление поисковых индексов оказывается нетривиальной задачей.

Поэтому после ряда экспериментов было принято решения следовать подходу, намеченному в [Tummarello et al., 2005] и представлять поисковые индексы как набор RDF-троек. В отличие от указанной работы мы не ставили себе целью получить RDF-представление, полностью эквивалентное TEI-разметке, а только обеспечить возможность быстрого нахождения текстов/фрагментов по заданному поисковому предписанию.

Мы старались по возможности следовать существующим стандартам и рекомендациям в области метаданных, однако не для всех параметров это оказалось возможным. В принципе, использовались классы и свойства следующих схем:

Dublin Core Terms для представления библиографического описания и структурирования текстов

FOAF для представления имен собственных

Для некоторых параметров, в первую очередь стиховедческих, готовых RDF-элементов, кажется, не существует вообще, а для некоторых — стандартные элементы оказываются недостаточно детализованными.

Так, в рамках принятой концептуальной модели, необходимо различать несколько видов имен собственных (например, имена исторических лиц vs имена мифологические). С другой стороны, отношение `dcterms:isPartOf`, естественным образом используемое нами для репрезентации структуры текста, не позволяет выразить существование особых частей текста, например, эпиграфов. Поэтому нами была разработана собственная RDF-схема, (а) реализующая недостающие элементы и (б) вводящая подклассы и подсвойства для интересующих нас категорий. Нами использовался метаязык RDFS, поскольку использование OWL в нашем случае приводило бы только к увеличению накладных расходов.

После определения модели данных встал вопрос о выборе платформы реализации, которая бы позволяла легко оперировать как SGML, так и RDF. Такая платформа была найдена — ей оказалась система программирования SWI-Prolog, имеющая как встроенный SGML-парсер (реализующий подмножество полного SGML, достаточное для обработки документов TEI P4), так и развитые средства работы с RDF-тройками.

Список литературы

- Андреев и др., 2009 — Андреев А. В., Бухаркин П. Е., Матвеев Е. М., Пономарева М. В. О разработке новой теоретической модели репрезентации истории литературы (на материале русской литературы XVIII века) // Литературная культура России XVIII века. Вып. 3. СПб, 2009. С. 303–310
- Tummarello et al., 2005 — Tummarello G., Morbidoni C., Pierazzo E. Toward textual encoding based on RDF // Proceedings of the 9th ICCI International Conference on Electronic Publishing. Leuven-Heverlee, 2005. Pp. 57–64.