Сравнительный анализ подходов к созданию электронных филологических коллекций

Андреев А. В., Бухаркин П. Е., Пономарева М. В.

13 декабря 2011 г.

В отличие от чисто лингвистических корпусов текстов, принципы построения которых достаточно хорошо изучены, методологии разработки электронных коллекций, ориентированных на историко-литературоведческий анализ, фактически не существует.

В настоящее время на материале русской литературы развиваются несколько электронных коллекций, из которых аиболее известны, вероятно, проекты «ФЭБ-Веб», «Российская виртуальная библиотека», электронная библиотека «ImWerden»[1]. В данном работе мы не будем касаться вопросов, связанных с комплектацией фондов электронных коллекций, а также эдиционных принципов, лежащих в основе этих коллекций; мы ограничимся лишь проблемами организации текстов внутри коллекции.

Большинство электронных коллекций, включая и вышеперечисленные, в целом следуют традиционному «библиографическому» принципу, при котором тексты в коллекции структурируются посредством нескольких каталогов, в первую очередь. Так, в «Русской виртуальной библиотеке» тексты организованы в 4-уровневую иерархию (раздел – автор – издание – конкретное произведение); в библиотеке «ImWerden» издания упорядочены с помощью алфавитного (автор/издание[2]) и неформального систематического каталогов. Ни одна из этих коллекций не имеет средств автоматизированного библиографического поиска – пользо-

[1]Международный проект «Гутенберг», одна из наиболее известных и полных электронных библиотек, текстов на русском языке практически не содержит

[2]Библиотека «ImWerden» оперирует не отдельными произведениями, а изданиями целиком, оцифрованными в формате PDF

ватель просто имеет дело с упорядоченными списками. Проект «ФЭБ-Веб» имеет несколько более сложную структуру: тексты распределены по нескольким *электронным научным изданиям (ЭНИ)*, внутри которых собраны издания, относящиеся к одному автору или к одной теме; структура каждого издания вопроизводит структуру оригинала, с разбиением на тома, главы и т. п. Помимо этого, «ФЭБ-Веб» представляет сводный алфавитный каталог и средства для библиографического поиска по названию произведения, автору и дате публикации (ценность которых, к сожалению, значительно снижается из-за поиска только в пределах одного ЭНИ, но не во всей коллекции).

Организация текстов в виде каталогов ориентирована только на внешний поиск текста (т. е. на поиск по тому или иному элементу метаданных — библиографического описания). Наряду с этим, все рассмотренные коллекции предоставляют в том или ином виде и средства *внутреннего* поиска, т. е. поиска по содержимому текста. В первую очередь в этой роли выступают механизмы полнотекстового поиска. Полнотекстовый поиск — это мощный инструмент анализа, применение которого стало в полной мере возможным только после перехода к цифровой форме представления текстов; однако применительно к задачам филологического исследования художественной литературы с его использованием связаны две существенные проблемы:

1. слова, не входящие в основной фонд русского языка, как-то: имена собственные, окказионализмы, диалектизмы и т. п., случаи употребления которых как раз представляют особый интерес для филолога, могут быть лемматизированы с большим трудом

2. в коллекции могут содержаться и тексты, грамматика которых отличается от современной нормативной (например, произведения русской литературы XVIII в.)

К сожалению, во всех трех рассмотренных коллекциях эти проблемы полностью вытесняются гораздо более серьезной проблемой, а именно: выбором в качестве инструмента поиска стандартных средств Google (ImWerden) и Yandex (РВБ и ФЭБ-Веб). Поскольку алгоритмы индексации и поиска, используемые в этих средствах, являются коммерческой тайной соответствующих компании, результаты поиска принципиально не верифицируемы и, следовательно, *категорически* не могут быть использованы как основа для серьезного научного исследования.

Помимо такого общего поиска, РВБ и ФЭБ-Веб также содержат научно выверенные конкордансы к отдельным произведениям и авторам; также из средств внутреннего поиска в РВБ присутствуют метрико-строфические указатели к отдельным авторам.

Следует отметить, что большинство электронных коллекций являются достаточно закрытыми — не в плане ограничения доступа к информации, а в плане возможности интеграции со сторонними инструментами анализа и обработки данных. Ни одна из рассмотренных коллекций не предоставляет никаких программных интерфейсов (например, Web-сервисов) для работы с текстами коллекции. Более того, ни одна из этих коллекций не предоставляет доступа к машинно-читаемым версиям текстов[3] и библиографических описаний. Таким образом, пользователь имеет дело исключительно с отформатированными HTML-страницами и со встроенными средствами поиска и навигации, которые, как было выше указано, имеют достаточно ограниченную функциональность.

[3]Для проектов ФЭБ-Веб и РВБ существование таких версий заявлено, однако пользователю они не доступны; более того, только для РВБ существует описание формата этих внутренних версий.