Боровинский Арсен Исаевич
Основатель библиотеки ELiS
Россия, Пермь
arsen@elibsystem.ru

Связанные данные как основа хранения сущностей в ядре электронной библиотеки

Аннотация
В докладе рассматривается возможность построения электронной библиотеки на основе связанных данных, использование связанных данных как внутренней структуры хранения библиографических данных взамен MARC-структуры, некоторые вопросы хранения и использования сущностей в виде связанных данных.

Ключевые слова: *связанные данные, каталогизация, структура библиографических данных, сущности, BIBFRAME, RDA.*

Основным стандартом распространения данных в интернете является RDF от консорциума W3C. На базе этого универсального стандарта описания ресурсов, Американская библиотечная ассоциация (ALA) выпустила стандарт RDA, а Библиотека Конгресса США BIBFRAME. Кроме указанных стандартов, были предприняты попытки создать стандарт библиографического описания в рамках консорциума W3C и библиографическое расширение структурированного описания ресурсов Schema.org.

Последние годы активизировался переход на стандарты BIBFRAME и RDA в качестве замены MARC-формата. В итоге ожидается вывод MARC из активной эксплуатации. В связи с чем возникает необходимость радикальной смены процессов каталогизации в библиотеках и описания документов, в том числе в электронных библиотеках.

В рамках разработки новой версии электронной библиотеки ELiS встала задача пересмотра внутренней структуры хранения данных для придания ей большей гибкости. Решено реализовать подход, основанный на моделировании предметной области графами, в которых узлами являются сохраняемые сущности со своими локально-уникальными идентификаторами, а ребрами – связи между сущностями. Сущности и связи сохраняются в базу данных. В качестве базы данных рассматриваются прежде всего реляционные базы данных, а не графовые.

Эта схема является переложением RDF на реляционную модель. При этом нужны всего две таблицы: таблица сущностей и таблица связей между сущностями.

В ELiS и сущности и связи являются типизированными.

В RDF ресурс может иметь не только связи с другими сущностями, но и типизированные (именованные) атрибуты-значения.

С точки зрения реляционной модели, такие сущности будут иметь столбцы, имена которых соответствуют пространствам имен (типам) атрибутов узла графа.

Так как типов сущностей в реальности потребуется много и у многих сущностей будут собственные типизированные атрибуты, в реляционной модели потребуется большое количество столбцов, а таблица сущностей будет разреженной.

С учетом разреженности таблиц, в качестве баз данных разумно рассматривать не только традиционные реляционные базы данных, но и нереляционные с колоночной архитектурой хранения таблиц, такие как Apache Cassandra.

Однако с точки зрения кода информационной системы, работать с такой структурой данных неудобно, поскольку средства разработки не обеспечивают автодополнение кода, что скажется на сложности написания бизнес-логики приложения и потенциально влечет большое количество ошибок.

Преодолеть этот недостаток можно написанием кода на объектно-ориентированном языке, создавая на каждый тип хранимой сущности класс.

Бизнес-логику тогда можно привязывать к классам исходного кода, а не к хранимым типам сущностей и связям. В классы можно встроить валидаторы связей и значений (например, что в связь «автором книги является» нельзя было поместить класс не являющийся агентом). Для удобства обращения к часто-необходимым полям можно сделать методы-помощники, которые инкапсулируют цепочки извлечения данных из связанных сущностей, ну а в сервисах бизнес-логики работать с объектами классов, а не с типизованными сущностями с множеством связей с другими сущностями.

Путем наследования классов и типов связей можно удобно реализовать наследование поведения сущностей и наследование поведения связей.

Недостатком подхода является запрет в большинстве объектно-ориентированных языков множественного наследования, в то время как в RDF оно возможно по стандарту и часто используется. Соответственно, удобство программирования потребует наложить некоторые

ограничения на возможности RDF в части описания сущностей и связей для библиографического описания.

С точки зрения совместимости с RDA и BIBFRAME представляется, что обе эти систему не обеспечивают достаточную подробность описания данных для тех задач, под которые создается новая библиотека ELiS. Для ELiS требует больше как классов, так и атрибутов. Поэтому в новой версии ELiS планируется реализовать собственную схему данных, но из которой можно будет для многих сущностей получить описание и в BIBFRAME и в RDA и в других форматах, включая Dublin Core, Schema.org и Microdata.