Реляционная база данных электронной библиотеки в Semantic Web. Представление метаданных в виде связанных данных

© Новицкий А.В.

Институт программных систем НАН Украины
ale-nov@yandex.ru

Аннотация

В статье описано представления реляционной базы данных, на которую накладываются определенные ограничения, в виде возможности публикации соответственно концепции связанных данных. В качестве сервера для публикации используется D2R Server.

1 Введение

Развитие электронных библиотек (ЭБ) непосредственно связано с развитием технологий доступа и организации информации. Семантические технологии являются наиболее динамичными в подходах представления информации в машиночитаемом виде.

Рассмотрим задачи, для которых применяются семантические технологии. Можно выделить два основных направления. Первое направление связано с развитием и внедрением сервисов, которые имеют определенное семантическое описание (WSMO, OWL-S, SAWSL т.д.) с целью решения задач их автоматической композиции. Однако это актуально лишь в том случае, когда сервисов есть множество, и поэтому целесообразно говорить о новом классе ЭБ как средстве поддержки сложных процессов коммуникации, хранения и обработки информации. Например, для науки потребность в таких библиотеках определяется через среду поддержки научных исследований. В классических ЭБ задачи автоматической композиции сервисов отсутствуют ввиду конечного множества сервисов и обозначенных целей. Второе направление связано с представлением, поиском и организацией доступа к информации в ЭБ. В рамках решения этой задачи предлагается добавлять метаинформацию об информационных объектах ЭБ. Как правило, в качестве метаинформации используют схему описания Дублинское ядро (ДЯ). Для эффективного поиска необходимо собрать и представить метаданные соответствующим образом. Semantic Web дает набор технологий, которые позволяют управлять метаданными. Для того чтобы дальше объяснять преимущества Semantic Web, введем классификацию поиска следующим образом: поиск на основе реляционной модели данных будем называть реляционным поиском. В то же время поиск в сетевой модели (к которой относятся семантические сети) данных будем называть сетевым поиском.

2 Linked Data и метаданные

2.1 Linked Data

Для ЭБ технология Semantic Web позволяет решить ряд принципиальных проблем таких как:

- интеграция информации, в различных моделях метаданных;
- обеспечение взаимодействия с другими системами (не только электронными библиотеками);
- удобного и адаптированного поиска с соответствующими интерфейсами для отображения семантики.

Одной из действующих моделей Semantic Web является модель связанных данных - Linked Data. Основные принципы Linked Data изложены в [1]. Преимущество связанных данных заключается в том, что ценность и полезность данных увеличивается, по мере увеличения количества связей с другими данными. Основные принципы связанных данных:

1. использование URIs в качестве имен для сущностей;
2. использование HTTP URIs чтобы люди могли увидеть эти имена;
3. в URI следует представлять полезную информацию, то есть они должны быть осмыслены;
4. ресурс должен содержать ссылки на другие URI с целью раскрытия дополнительной информации о сущности.

Поэтому естественно встает вопрос о представлении ресурсов ЭБ согласно концепции связанных данных. В работе [2] уже описано представления DOI как Linked data, также в [3], [4] описаны выражения элементов общепринятых схем метаданных как Linked data.

В работе мы рассмотрим проблему публикации данных представленных в ЭБ в соответствии принципам Linked Data. В качестве примера будем использовать свободное ПО Eprints. Полученные

результаты возможно распространить на Dspace и другие подобные системы.

Как известно в большинстве случаях данные в таких ЭБ хранятся в реляционной базе данных. Поэтому мы обратим внимание на общие основы отображения реляционной базы данных в Linked Data. В качестве примера ПО для публикации Linked Data используем D2R-SERVER [5], а в качестве источника данных реляционную базу данных Eprints [6].

2.2 Отображение реляционной базы данных в Linked Data

D2R Server [5] является инструментом для публикации реляционных баз данных в Semantic Web. Он позволяет RDF и HTML браузерам перемещаться по содержанию базы данных, а приложениям запрашивать информацию из базы данных, используя язык запросов SPARQL. Детальное исследование использования данного ПО для публикации веб-сайтов как Linked Data можно найти в [7].

Рассмотрим, каким образом происходит отображение реляционной базы данных в Linked Data. Как известно, базовыми понятиями реляционной базы данных являются: тип данных, домен, атрибут, кортеж, первичный ключ и отношение. Отношение (relation) - это вся структура целиком, набор записей (в обычном понимании - таблица). Кортеж - это каждая строка, содержащая данные. Более распространённый, но менее формальный термин - запись. Атрибут - это столбец в отношении.

Попробуем определить общий метод, согласно которому реляционная база данных публикуется в виде связанных данных. Для этого отобразим основные понятия реляционной базы к основным понятиям связанных данных.

В связанных данных каждый ресурс, который имеет уникальный URI, описывается с помощью модели данных RDF. Ресурсом в RDF может быть любая сущность - как информационный ресурс (например, веб-сайт или изображения), так и не информационный ресурс (человек, город или некое абстрактное понятие). Ресурс состоит из списка утверждений в виде «субъект - предикат - объект», каждое такое утверждение называется триплетом. Для обозначения субъектов, предикатов и объектов в RDF используются URI. Множество RDF-утверждений образует ориентированный граф, в котором вершинами являются субъекты и объекты, а ребра являются предикатами. Модель данных RDF предусматривает собой использовать онтологии. В LD общепринято использовать онтологии для предикатов. Схема отражения имеет следующий вид:

Реляционная база данных	**Связанные данные**
Тип данных	XML schema datatypes
Атрибут	Предикат, который определяется через общепринятые словари и онтологии
Кортеж	Предикат-объект
Первичный ключ	Субъект
Отношение	Ресурс
Домен	

Для того чтобы более детально понять процесс отражения обратимся к рис. 1, на котором показано как элементы отношения представляются RDF графом.

Table Metadata

Id (key)	title	date
1	Book about DC	02-02-2010
2	Book about PC	03-02-2011
3	Book about Notebook	25-09-2011

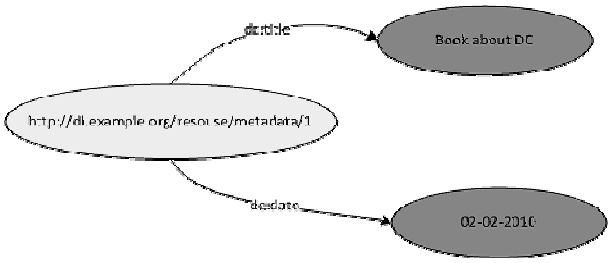


Рис.1 Отношение и представления кортежа через RDF.

В примере на Рис 1. представлено некоторое отношение, которое содержит информацию об метаданных. Данное отношение может быть описано 3-мя информационными ресурсами, каждый, из которых, образуется 2-мя триплетами.

2.3 Публикация связанных данных

Для упрощения, сначала смоделируем отображение реляционной базы данных ПО Eprints к модели связанных данных. Ниже на Рис. 2 представлены схема базы данных Eprints 3.2.2. Из множества таблиц выбраны только те, которые непосредственно имеют отношение к метаданным, их оказалось 27. Для примера мы выбрали только 2 таблицы.

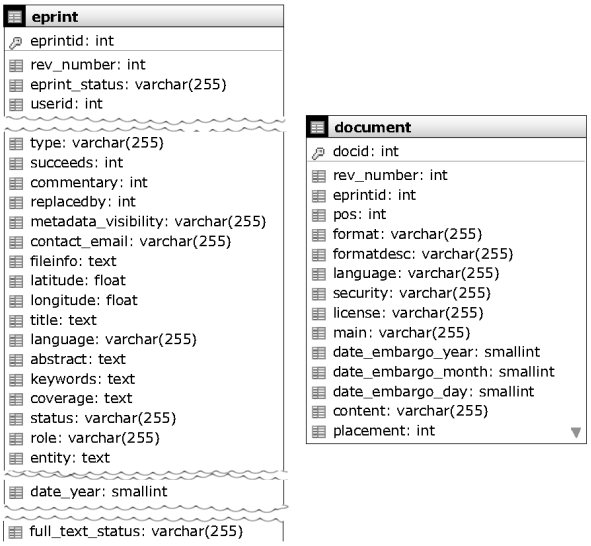


Рис .2 Часть схемы реляционной базы данных Eprints 3

Как было ранее отмечено, для публикации LD нами используется ПО D2R server. Сопоставление между реляционной моделью данных и LD происходит путем определения специального файла на основе спецификации [8].

Онтологии сопоставляются со схемой базы данных, используя d2rq:ClassMaps и d2rq:PropertyBridges. Центральным объектом в D2RQ, а также объектом, с которого начинается построение новой схемы отображения D2RQ является ClassMap. Понятие ClassMap представляет собой класс или группу аналогично классам онтологии. ClassMap также определяет способ идентификации экземпляров класса. ClassMap имеет наборы PropertyBridges, которые определяют свойства экземпляров класса.

Приведем пример файла, отвечающий за мапинг. В файле на основе спецификации [8] нами определено пространство имен для ДЯ и Квалификаторы ДЯ. Например, для описания формата файлов используем словарь DCMI Metadata Terms, который определяется пространством имен: @prefix dct:.

В дальнейшем мы задали соответствие первичного ключа к субъектам, а атрибутов отношения к предикатам. Продемонстрируем части созданной нами схемы отображения. Определение формата файла ресурса dct:format (имеется ввиду формат документа в ЭБ), используя сопоставление свойств из различных таблиц:

```
map: eprint__format a d2rq: PropertyBridge;
d2rq: belongsToClassMap map: eprint;
d2rq: property dct: format;
d2rq: column "document.format";
d2rq: join "document.eprintid = eprint.eprintid";
```

Определим схему отображения заголовка (title), типа (type), и описания ресурса (description)

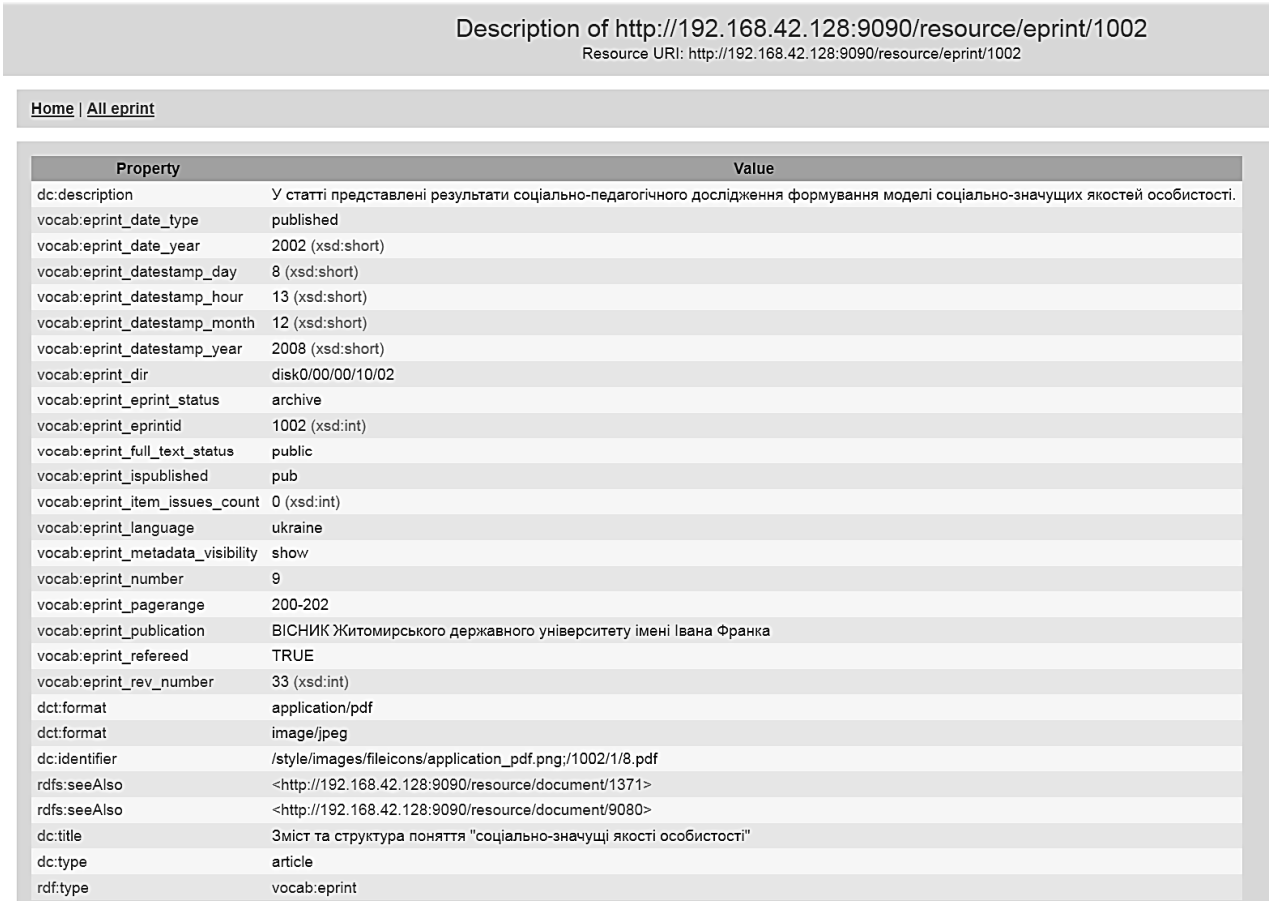
```
map: eprint_title a d2rq: PropertyBridge;
d2rq: belongsToClassMap map: eprint;
d2rq: property dc: title;
d2rq: propertyDefinitionLabel "eprint title";
d2rq: column "eprint.title";

map: eprint_type a d2rq: PropertyBridge;
d2rq: belongsToClassMap map: eprint;
d2rq: property dc: type;
d2rq: propertyDefinitionLabel "eprint type";
d2rq: column "eprint.type";
map: eprint_abstract a d2rq: PropertyBridge;
d2rq: belongsToClassMap map: eprint;
d2rq: property dc: description;
d2rq: propertyDefinitionLabel "eprint abstract";
d2rq: column "eprint.abstract";
```

На данном этапе можно сказать, что мы построили схему отображения база данных Eprints в модель LD, тем не менее, схема не полна. Для того, чтобы наше представление полностью соответствовало концепции связанных данных, а именно, чтобы ресурс был связан с другими ресурсами, необходимо, чтобы объект являлся субъектом для другого триплета. Мы дополнительно добавили в схему конструкцию, которая позволила связать ресурсы между собой.

```
map: eprint__label a d2rq: PropertyBridge;
d2rq: belongsToClassMap map: eprint;
d2rq: refersToClassMap map: document;
d2rq: property rdfs: seeAlso;
d2rq: join "document.eprintid => eprint.eprintid";
```

Определив таким способом ключевые аспекты схемы отображения, мы получили модель LD для реляционной базы данных Eprints.

На Рис 3. представлено вид информационного ресурса с ИД 1002 опубликованного при помощи D2R server и на основе разработанной нами схемы отображения.



Description of http://192.168.42.128:9090/resource/eprint/1002
Resource URI: http://192.168.42.128:9090/resource/eprint/1002

Home | All eprint

Property	Value
dc:description	У статті представлені результати соціально-педагогічного дослідження формування моделі соціально-значущих якостей особистості.
vocab:eprint_date_type	published
vocab:eprint_date_year	2002 (xsd:short)
vocab:eprint_datestamp_day	8 (xsd:short)
vocab:eprint_datestamp_hour	13 (xsd:short)
vocab:eprint_datestamp_month	12 (xsd:short)
vocab:eprint_datestamp_year	2008 (xsd:short)
vocab:eprint_dir	disk0/00/00/10/02
vocab:eprint_eprint_status	archive
vocab:eprint_eprintid	1002 (xsd:int)
vocab:eprint_full_text_status	public
vocab:eprint_ispublished	pub
vocab:eprint_item_issues_count	0 (xsd:int)
vocab:eprint_language	ukraine
vocab:eprint_metadata_visibility	show
vocab:eprint_number	9
vocab:eprint_pagerange	200-202
vocab:eprint_publication	ВІСНИК Житомирського державного університету імені Івана Франка
vocab:eprint_refereed	TRUE
vocab:eprint_rev_number	33 (xsd:int)
dct:format	application/pdf
dct:format	image/jpeg
dc:identifier	/style/images/fileicons/application_pdf.png;/1002/1/8.pdf
rdfs:seeAlso	<http://192.168.42.128:9090/resource/document/1371>
rdfs:seeAlso	<http://192.168.42.128:9090/resource/document/9080>
dc:title	Зміст та структура поняття "соціально-значущі якості особистості"
dc:type	article
rdf:type	vocab:eprint

Рис 3. Информационный ресурс с ИД 1002 опубликованного при помощи D2R server

При разработке отображения было обнаружено проблему с неправильным подходом к проектированию базы данных. В результате данная схема реляционной базы данных ПО Eprints, оказалась не подготовленной к представлению ее в виде LD, для всех атрибутов. Для примера: если нам необходимо сгруппировать в RDF объект все статьи, опубликованные в определенном печатном издании, то атрибут, который отвечает за это, не вынесен в отдельные отношения, что делает невозможным такое группирование, что, в свою очередь, не позволяет создать ресурс, который имеет свой URI и содержал бы описание неинформационных ресурсов о печатном издании. Причиной является неправильное выделение типа сущностей реляционной базы данных. В данном случае, типом сущности есть элементы ДЯ, которые необходимо принимать в качестве атрибута, и вносить в одном отношении только на основе уникальности экземпляра сущности.

Говоря другими словами, в случае проектирования базы данных для ЭБ, необходимо принимать во внимание, чтобы атрибуты соответствующих элементам ДЯ содержатся в одной таблице только тогда, когда значение этого атрибута является уникальным для электронного ресурса описываемого ДЯ. В противном случае, такие атрибуты следует группировать в отдельные таблицы и связывать по ключу. Ниже приведена таблица требований для ДЯ:

Элемент дублинского ядра	Принадлежность атрибута к одному отношению	Вынесение атрибута в отдельное отношение
Title	+	-
Creator	-	+
Subject	-	+
Description	+	-
Publisher	-	+
Contributor	-	+
Date	+/-	+/-
Type	-	+
Format	-	+
Identifier	+	-
Source	-	+
Language	+	-
Relation	+	-
Coverage	+	-
Rights	-	+

Следует отметить, что для другого популярного ПО Dspace, проблема также актуальна, поскольку схема базы данных этого ПО подобная и имеет те же проблемы.

2.4 Поиск информации в сети связанных данных

Полученная модель связанных данных построена на данных реляционной структуры. В данном случае запрос к семантической модели (сетевой поиск) данных транслируется в реляционный поиск. Продемонстрируем это на примерах запросов и ответов:

запрос
SELECT?x?y
WHERE {
?x rdfs: seeAlso?y.
?x dc: type "article".
}
LIMIT 3

Результат выполнения данного запроса будет иметь следующий вид

x	y
<http://192.168.42.128:9090/resource/eprint/61>	<http://192.168.42.128:9090/resource/document/4030>
<http://192.168.42.128:9090/resource/eprint/61>	<http://192.168.42.128:9090/resource/document/11313>
<http://192.168.42.128:9090/resource/eprint/311>	<http://192.168.42.128:9090/resource/document/2417>

запрос
SELECT?s?p?o
WHERE {
?s?p?o.
?s dc: type?o
}

Результат будет иметь следующий вид

s	p	o
<http://192.168.42.128:9090/resource/eprint/129>	<http://purl.org/dc/elements/1.1/type>	"Article"
<http://192.168.42.128:9090/resource/eprint/13>	<http://purl.org/dc/elements/1.1/type>	"Article"
<http://192.168.42.128:9090/resource/eprint/15>	<http://purl.org/dc/elements/1.1/type>	"Book_section"
<http://192.168.42.128:9090/resource/eprint/59>	<http://purl.org/dc/elements/1.1/type>	"Article"

Однако следует подчеркнуть, что D2R Server не поддерживает вывод знаний при запросах.

3. Выводы

В результате работы были сформированы некоторые отображения реляционной модели данных в модель LD а также принципы проектирования реляционной базы данных для ЭБ, с возможностью отображения в Linked Data. Указаны также ограничения, которые необходимо соблюдать при проектировании баз данных для ЭБ. Конечно, описанные проблемы можно решить путем построение программных медиаторов, которые смоделируют необходимую схему данных, но это повлечет дополнительную сложность проектирования ЭБ. Дальнейшая работа связана с проектированием виртуальных таблиц и их отражением в Linked Data.

Литература

[1] Berners-Lee, T.: Linked Data. In: World Wide Web Consortium (W3C). (Accessed 2009) Available at: HYPERLINK "http://www.w3.org/DesignIssues/LinkedData.html" http://www.w3.org/DesignIssues/LinkedData.html

[2] Erickson, J.: DOIs, URIs and Cool Resolution. In: Bitwacker Associates. (Accessed 2010) Available at: HYPERLINK "http://bitwacker.wordpress.com/2010/02/04/dois-uris-and-cool-resolution/" http://bitwacker.wordpress.com/2010/02/04/dois-uris-and-cool-resolution/

[3] Ed Summers, A.: LCSH, SKOS and Linked Data. Proc. Int'l Conf. on Dublin Core and Metadata Applications, 25-33 (2008)

[4] Baker, T.: Tutorial: Dublin Core -Building blocks for interoperability. In : Dublin Core and Linked Data, Tokyo (2010) http://www.meta-proj.jp/ev-1/ev-p3.pdf.

[5] Chris Bizer, R.: D2R Server - Publishing Relational Databases on the Semantic Web. In: WWW4 - research application server of the Lehrstuhl für Wirtschaftsinformatik. Available at: HYPERLINK "http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/" http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

[6] University of Southampton: EPrints - Digital Repository Software. Available at: HYPERLINK "http://www.eprints.org/" http://www.eprints.org/

[7] Wang, X.: Investigating the Use of Linked Data for Exposing the Data from the Catalhoyuk Web Site., Southampton (2009) http://2tree.brinkster.net/Resource/Dissertation/Master.pdf.

[8] Chris Bizer, R.: The D2RQ Plattform v0.7 - User Manual. In: WWW4 - research application server of the Lehrstuhl für Wirtschaftsinformatik. Available at: HYPERLINK "http://www4.wiwiss.fu-berlin.de/bizer/d2rq/spec/" http://www4.wiwiss.fu-berlin.de/bizer/d2rq/spec/

Simulation of a relational database of digital libraries in Semantic Web. Provision of metadata in the form of Linked Data

Novytskyi Oleksandr

The article describes the modeling of a relational database, which imposed certain restrictions, as the possibility of publishing concepts, respectively, of Linked Data. As a server for publishing using D2R Server.