Организация ссылок в неструктурированных данных в гипертекстовой системе для БД НИКА

В. А. Тищенко [I, II]

[I] Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия
[II] Образовательное частное учреждение высшего образования "Православный Свято-Тихоновский гуманитарный университет", Москва, Россия

Аннотация. На основе близости моделей ООСУБД НИКА и гипертекста делается вывод о том, что гипертекстовый интерфейс более полно раскрывает возможности БД НИКА. Применение различных спецификаций отображения к объектам БД позволяет отображать их различными способами в HTML/XML-документах. С другой стороны, существуют неструктурированные данные, которые не могут быть выделены как особые вершины БД. Для отображения таких вершин необходимо применение метаспецификаций. Так для описания ссылки на публикацию внутри текстового массива определяется метаспецификация anchor или закладка посредством текстового шаблона.

Ключевые слова: спецификация отображения, слабоструктурированная информация, метаспецификация, anchor, текстовый шаблон, контекстная ссылка.

DOI 10.14357/20718632230206

Контекст проблемы создания ссылок в текстовых массивах БД НИКА

Идентификация вершин в базе данных НИКА осуществляется посредством координаты в БД, которая содержит последовательность вершин от корня до данного объекта в БД. Гипертекстовая система ООСУБД НИКА [1] использует схему описания данных как конструктор [2] XML/HTML-документов, которые содержат объекты БД, заданные в схеме и представляемые в документе в соответствии с методами отображения объектов. Иерархия объектов отображается в иерархию гипертекстовых документов, не обязательно совпадающую с иерархией объектов, но подчиненность объектов в документах сохраняется такой же, как в БД. Гипертекстовые ссылки, связывающие документы в иерархию естественным образом, определяются в виде координат объектов в БД. Однако информация в БД НИКА может быть слабоструктурированной при наличии текстовых массивов. В связи с этим возникает актуальная проблема, связанная с возможностью построения ссылок внутри текстового массива на любые объекты в БД.

Внутренние и внешние ссылки

Текстовая информация может содержать по смыслу ссылки на соседние объекты базы данных как внутри данного документа, например, на публикацию в списке публикаций, так и на объекты в других документах, например, упоминание фамилии в другой биографической справке. Тривиальным способом задание таких ссылок является добавление гипертекстовой

разметки в сам текст, что является неудобным и неэффективным, т.к. сам объект ссылки может менять название в БД, а значит и ссылку на него. Другой способ — это *доопределение до метасистемы* гипертекстовой системы ООБД НИКА, ядро которой состоит из методов (спецификаций) отображения объектов БД, *метаметодами* (метаспецификациями).

Определение метаметодов

Объект типа массив, содержащий неструктурированные текстовые данные, может быть отображен в виде непрерывного текста. В БД такой массив определяется в виде нумерованного массива вида:

$$T(\ Text\ ) = \{\ [\ NUM, 01\ ]\ \}$$

Здесь *NUM* является целочисленным ключом и *01* — текстовой строкой. Такая структура не выделяет специальных полей данных внутри текста, которым могут быть присвоены методы отображения, и для определения ссылок внутри текстового массива нельзя воспользоваться существующим методом отображения HTTP [3] терминальных вершин в виде гипертекстовых ссылок. Для определения ссылок в текстовом массиве необходимо ввести метаметоды, которые будут назначаться определенным *фрагментам* текстового массива. Способом задания текстового фрагмента является текстовый шаблон, которому соответствует фрагмент. Метаметод определяется для текстового массива с заданной координатой в БД как подстановка вида:

M = <TextArrayDbCoordinate, TextTemplate, HtmlXmlSubstitution>

и задается текстовым шаблоном и соответствующей ему гипертекстовой разметкой. TextArrayDbCoordinate задает координату массива в БД. Например, ниже дается объявление шаблона *метаметода* LiteratureReference для произвольного текстового массива *ссылки на публикацию* в виде РБНФ-правил.

PublicationDbCoordinate обозначает относительную координату вершины в БД.

Организация ссылок в текстовом массиве

Метаметоду, определяющему гипертекстовую ссылку, назначается объект БД, на который он ссылается, причем ссылка в тексте конкретизирует элемент массива, на который она ссылается. В случае *метаметода* LiteratureReference объектом является массив “Публикации”, а элемент массива задается в самой текстовой ссылке в виде цифры в квадратных скобках. В общем случае ссылка задается как полная координата объекта от корня БД. Если объект ссылки находится внутри данного документа, то в атрибуте гипертекстовой ссылки координата в БД задается в виде якоря или закладки, например: #PublicationDbCoordinate. При этом, объекту ссылки назначается метод ANCH (anchor), который добавляет разметку, назначающую имя объекту в документе. В атрибуте name указывается координата объекта в БД. Если объект ссылки является внешним, то ссылка создается также как и в случае обычных методов отображения нетерминальных вершин или объектов БД.

```
LiteratureReferenceTemplate  ::=  “[“Number”]”.
                      Number  ::=  Digit{Digit}.
                       Digit  ::=  ”0”|”1”|…|”9”.
```

Подстановочная часть метаметода LiteratureReference задает гипертекстовую разметку:

```
LiteratureReferenceSubstitution  ::=  “[“OpenRefElemTag Number CloseRefElemTag ”]”.
                  OpenRefElemTag  ::=  “<”AnchorElem RefAttr”>”
                 CloseRefElemTag  ::=  “<”, ”/”, AnchorElemr, ”>”
                     AnchorElemr  ::=  “a”
                         RefAttr  ::=  “href”, ”=”, PublicationDbCoordinate
```

Оптимизация длин ссылок в документе

В общем случае ссылки отображаются в виде полной координаты в БД. Все нетерминальные вершины на одном уровне БД ссылаются на подчиненные им вершины (кроме вершин типа ссылка). Во всех координатах таких вершин дублируется общая часть пути в БД до объекта, которому подчинены эти вершины. Гипертекстовый язык разметки позволяет сокращать общую часть пути, которая может быть указана в заголовочной части документа в атрибуте href элемента base. Такой подход позволяет сокращать все подобные ссылки, начиная их с координаты, в которой они отличаются. Общей частью путей в БД для ссылок в документе является координата в БД того объекта, который отображается в данный документ. Этот объект является базовой вершиной для всех остальных объектов, отображенных в документе. Исключения составляют объекты БД типа ссылка на значение. У вершин такого типа ссылка на объект БД хранится в виде пути в схеме описания данных. Этот путь, конкретизированный ключами из БД, составляет полную координату к объекту в БД.

Пример ссылок в текстовых массивах

На Рис. 1 в биографической справке имеется текстовый массив под вершиной Кончина, содержащий ссылки на публикацию “[1]” и заявителя “[Зв1]”. Для примера показана разметка с указанием *закладок* на элементах массивов

NIKA_ROOT INDEX▼ ФИО▼ Димитрий (Пьяных Дмитрий Васильевич)▼

Димитрий (Пьяных Дмитрий Васильевич)
Год рождения 1902
День рождения 19
Месяц рождения 10
Место рождения Пермская о., Верхне-Муллинский р., д.Лягушино
иеромонах

...

Кончина
Неточная дата 1951–1955гг.
погиб в ИТЛ (?)
Место Приморский край, Севвостлаг (?)

По данным [Зв1] погиб в Севвостлаге в ноябре 1943г. По данным [1] — в следственном деле УВД Севвостлага на 03.01.1951г. значился в живых. По данным [Зв1] 06.10.1955г. архивное дело осужденного иеромонаха Димитрия в УВД Севвостлага было уничтожено

Реабилитация
Дата 00/00/1989
Кем реабилитирован по Указу Президиума ВС СССР от 16.01.1989г.
По году репрессий 1939

Публикации
****1.**** Душу не погублю. Исповедники и осведомители в документах и... о методах агентурной работы/ Сост., предисловие В.А.Королева. М.: Содружество Православный Паломник, 2001. 240с., ил.
С.55–57,59–61,64–65,67,70–71.
****2.****Пермский Государственный Архив Новейшей истории. Фонд N 643.. *<http://www.permgani.ru/funds/index.php?act=fund&id=495&opis=10956&page=31>*

Заявители
****1.**** Костина (Пьяных) Валентина Александровна
внучатая племянница

Рис. 1. Фрагмент биографической справки

Публикации и Заявители. В качестве имен закладок использованы координаты относительно координаты базовой вершины в БД. Аналогично для ссылки на элемент массива Документы используется обозначение [Дn] и имя закладки Документ.n.

Указание объектов ссылок через форматную строку

Для задания закладок используются нетерминальные вершины, которые отображаются в порядке левого обхода иерархии вершин в БД. Для создания в документе форматированного вывода используются форматные строки, применяемые к терминальным вершинам посредством спецификации STL (style). На Рис. 1 в массивах Публикации и Заявители вместо ключей используются терминальные вершины Публикация и ФИОЗаявителя со спецификацией STL с атрибутом в виде форматной строки:

<**a** name=”%<rdbp>”>^*dn.</**a**> %s%<http>,

где %<rdbp> обозначает относительную координату в БД (Публикации.1, Публикации.2, Заявитель.1 и т.д.), ^*dn — значение ключа для вершины с системным номером dn, %s — значение текущей терминальной вершины, %<http> — гипертекстовая ссылка, которая может быть задана в терминальной вершине (в случае, если терминальной вершине присвоена спецификация HTTP). Аналогично для массива Документы вместо ключей используется терминальная вершина ОписаниеДокумента со спецификацией STL с той же форматной строкой. Терм %<http> указывается только в форматной строке для вершины Публикация.

Способы создания ссылок в гипертексте

Элементы ссылок являются концептуальными конструкциями и представляют собой связь между двумя ресурсами, одним из которых является текущий документ [4]. К ним относятся элементы a, area, form и link. Элемент link создает ссылку на внешний ресурс, первые три элемента создают гиперссылки (Табл. 1).

Все координаты в БД в документе кодируются посредством 6-битного кодировщика (аналог Base64). Альтернативным способом создания гипертекстовых ссылок является форма. Применение метода POST снимает возможные ограничения (как со стороны клиента, так и со стороны сервера) на длину URL, а значит и на длину координаты в БД. В этом методе данные формы помещаются в текст запроса.

Табл. 1. Примеры ссылок в БД с использованием ссылочных элементов

Элемент	HTML-разметка	Примечание
Link	<**link** rel=stylesheet type="text/css" href="/bin/loadsty.exe/nika_css/html.css">	Ссылка на стилевой файл html.css
A	<**a** href=”?NIKA_ROOT**.**INDEX**.**ГодАреста”> ГодАреста </**a**>	Гипертекстовая ссылка в виде координаты в БД
Area	<**map** name=graph> **…** <**area** shape=circle coords=”x37,y37,r” alt=”1937” href="?NIKA_ROOT**.**INDEX**.**ГодАреста**.**1937"> **…** </**map**>	Гипертекстовая ссылка на карте изображения графика арестов
Form	<**form** action=**…** method=POST> <**input** type=hidden name=query value="NIKA_ROOT**.**INDEX**.**ГодАреста"> <**input** type=submit value="ГодАреста" style="border:0;cursor:pointer;text-decoration:underline;color:blue;background-color:white"> </**form**>	Гипертекстовая ссылка методом POST в виде координаты в БД

Использование пользовательских атрибутов при определении ссылок

В соответствии с классификацией Холмского [5] язык HTML до версии 5 относится к *контекстно-свободным языкам*, т.к. является подмножеством языка стандартной обобщенной разметки SGML. Язык SGML, а также его подмножество — язык XML, определяют схему документа, являющуюся аналогом задания грамматики формального языка в БНФ-нотации. Язык HTML 5 *не является* контекстно-свободным языком, он содержит атрибуты data-*, задаваемые пользователем. Эти атрибуты не могут быть описаны в общем виде посредством формы БНФ. В качестве примера можно привести фрагмент массива, содержащий список ФИО, в котором каждый ключ – это гипертекстовая ссылка на подчиненный уровень, а также для каждого ключа добавлен пользовательский атрибут data-href-classif-{prefix}, в котором сохраняется ссылка в виде координаты в БД на соответствующий префикс в алфавитном классификаторе [6]. Каждый ключ в HTML-документе будет иметь следующий вид:

{key}

Имя атрибута **data-classif-href-{prefix}** может отличаться у разных ключей в зависимости от начального префикса. Предполагается, что префикс ключа в названии атрибута кодирован строчными английскими буквами. Не фиксированное название атрибуты не позволяет его определить посредством SGML схемы или xsd-схемы, а значит не может быть определено в БНФ-нотации.

В неструктурированном тексте (Рис. 1) ссылка на заявителя “Зв1” может содержать кроме закладки пользовательскую *контекстную* ссылку на классификатор по полю ФИОЗаявителя:

<**a** href="#Заявитель.1" **data-classif-href-fiozayavitelya-{prefix}**= "INDEX.ФИОЗаявителяAlpha.{prefix-path}">{key}

Контекстная ссылка **data-classif-href-fiozayavitelya-{prefix}** ссылается на префикс, с которого начинается ключ {key}, в многоуровневом индексе по полю ФИОЗаявителя.

Заключение

Одним из основных элементов гипертекстовой системы являются гипертекстовые ссылки и способы их организации. При отображении вершин БД в гипертекстовые документы можно выделить следующие способы.

1. Ссылки на подчиненные вершины для нетерминальных вершин текущего уровня в БД, на базовые вершины при отображении пути в БД, для элементов массивов ссылки на первый, предыдущий, следующий и последний элементы, а также сетевые ссылки для вершин типа ссылки на значение.
2. Ссылки в текстовых массивах, объектами которых являются любые вершины схемы описания данных (кроме терминальных и ссылочных вершин).
3. Ссылки в текстовых массивах, объектами которых являются подстроки любых текстовых массивов в БД.

В 1-ом способе перечислены ссылки, которые создаются спецификациями отображения при отображении вершин в гипертекстовые документы. Первый способ естественным образом определяется на основе схемы описания данных. Преимущество этого способа перед двумя другими состоит в том, что вершинам схемы присваиваются методы отображения вершин [3], а таже существует возможность определения циклических ссылок на шаблон [6]. Алфавитный классификатор является примером использования циклических шаблонов. 2-ой и 3-ий способы требуют задания ссылок в текстовых массивах посредством метаспецификаций. В 3-ем способе необходимо дополнительно указать закладку в текстовом массиве для объекта ссылки. Перечисленные способы могут определять произвольные ссылки. Это используется на практике, например, при создании в текстовых массивах ссылок на элементы массивов Публикации, Заявители, Документы. Метаспецификации для создания ссылок обогащают гипертекстовую систему БД НИКА возможностью определения произволь-

ных ссылок. Кроме того, метаспецификации могут использовать элементы семантики на уровне текста, при необходимости не отображать отдельные части текстового массива.

Литература

1. Емельянов Н.Е., Тищенко В.А. Представление гипертекста в СУБД НИКА // Технология программирования и хранения данных / Сб. трудов ИСА РАН. Т.45. Под ред. чл.-корр. РАН Арлазарова В.Л. и д.т.н. проф. Емельянова Н.Е. - М. 2009. С. 17-36.
2. Соловьев А.В., Тищенко В.А. Языки XSL и CSS как конструкторы версии для печати в гипертекстовой системе ООСУБД НИКА // Труды ИСА РАН. 2020. Т. 70. № 3. С. 65-74. DOI: 10.14357/20790279200308
3. Емельянов Н.Е., Тищенко В.А. Принципы построения web-сервера на основе объектно-ориентированной базы данных // Информационные технологии и вычислительные системы. 1997, N 4. С.90-99.
4. HTML Living Standard — Last Updated 7 October 2022. Copyright (c) WHATWG. https://html.spec.whatwg.org/multipage/
5. Тищенко А.В. Математические основы информатики = Mathematical Fundamentals of Informatics: Учебное пособие; Финуниверситет. — М.: Финуниверситет, 2014 — 128 с.
6. Тищенко В.А. Структура OPC-trie как новый тип индекса в СУБД НИКА // Труды ИСА РАН, 2021. Т. 71.Вып. 4. С.76-81.

Тищенко Владимир Александрович. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», Москва, Россия. Научный сотрудник, кандидат технических наук, сотрудник кафедры Информатики ПСТГУ. Область научных интересов: средства создания и поддержки электронных изданий и информационных систем. E-mail: vtischenko@isa.ru

Building of References in Unstructured Data in a Hypertext System for the NIKA Database

V. A. Tishchenko[I,II]

[I] Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia
[II] St. Tikhons”s Orthodox University, Moscow, Russia

Abstract. Based on the closeness of OODBMS NIKA and hypertext models, a conclusion about a more complete disclosure of the possibilities of the NIKA database through a hypertext interface is made. Applying different specifications of representation to database objects allows them to be displayed in different ways in HTML/XML documents. On the other hand, there are unstructured data that cannot be defined as special database nodes. To display such vertices, the use of metaspecifications is necessary. So, to describe a reference to a publication inside a text array, an anchor metaspecification is defined using a text template.

Keywords: specification of representation, semi-structured information, metaspecification, anchor, text template, context reference.

DOI 10.14357/20718632230206

References

1. Emelyanov N.E., Tishchenko V.A. Representation of hypertext in the NIKA DBMS // Technology of programming and data storage / Sat. Proceedings of the ISA RAS. V.45. Ed. Corresponding Member RAS Arlazarov V.L. and Doctor of Technical Sciences prof. Emelyanov N.E. - M. 2009.S. 17-36.
2. Soloviev A.V, Tishchenko V.A. XSL and CSS as printable version constructors in a hypertext system for NIKA DBMS // Information technology / Proceedings of the ISA RAS 2020. V.70: No. 3. P. 65-74. DOI: 10.14357/20790279200308
3. Emelyanov N.E., Tishchenko V.A. Principles of building a web server based on an object-oriented database // Information technology and computing systems. 1997, N 4. P. 90-99.
4. HTML Living Standard — Last Updated 7 October 2022. Copyright (c) WHATWG. https://html.spec.whatwg.org/multipage/

5. Tishchenko A.V. Mathematical Fundamentals of Informatics: Tutorial – Moscow, Financial University under the Government of the Russian Federation, 2014. 128 p.
6. Tishchenko V.A. OPC-trie: specification of the optimal classifier for the NIKA DBMS // Proceedings of ISA RAS, 2021. V. 71. N. 4. P.76-81.

Tishchenko V. A. Researcher, Candidate of Technical Sciences. Federal Research Center “Computer Science and Control” of Russian Academy of Sciences. Employee of department of Informatics, PSTGU. Number of printed works: more than 30. Research interests: means of creation and support of electronic publications and information systems. E-mail: vtischenko@isa.ru