OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.9:510

ТЕХНОЛОГИИ СВЯЗЫВАНИЯ ДАННЫХ В ПРОСТРАНСТВЕ ОТКРЫТЫХ ДАННЫХ НА ПРИМЕРЕ МАТЕМАТИЧЕСКОЙ КОЛЛЕКЦИИ

Невзорова О.А., [**] Кириллович А.В. [*]

[*] *Казанский федеральный университет,*
г. Казань, Россия
alik.kirillovich@gmail.com

** *Научно-исследовательский институт «Прикладная семиотика» АН Республики Татарстан,*
Казанский федеральный университет,
г. Казань, Россия
onevzoro@gmail.com

В работе рассматриваются программные технологии связывания данных в пространстве открытых данных с помощью программных систем LIMES и SILK, и эксперименты по связыванию математического RDF-набора данных.
Ключевые слова: связывание данных; пространство открытых связанных данных.

Введение

Проект Linked Open Data (LOD)[1] является наиболее значимым по результатам примером принятия и применения принципов Linked Data (Связанные данные). На сегодняшний день технологии Linked Data все более широко используются производителями первичных данных. Главное преимущество заключается в стандартизованном подходе к структурированию и хранению интегрированных данных. Как правило, данные загружаются и представляются в виде RDF (Resource Description Framework[2]), т.е. триплетов вида «субъект — предикат — объект», из таких традиционных хранилищ как реляционные базы данных или, реже, из веб-страниц или полуструктурированных текстовых документов, используя принципы связанных данных, предложенные Т. Бернерсом-Ли[3]. Ссылки RDF соединяют данные из различных источников в единый глобальный граф RDF и позволяют браузерам и поисковым роботам Связанных данных перемещаться между источниками данных.

В сентябре 2011 г. объем данных в LOD составлял более 30 млрд. триплетов, хранящихся в примерно 300 наборах данных. Среди доминирующих предметных областей — данные из правительственных источников (43% по числу триплетов), географические данные (22%) и науки о жизни (биология, биохимия, генетика и др. - 9%).

Технологии Linked Data применяются для совместного использования данных, относящихся к широкому спектру различных тематических доменов, а также данных, относящихся к различным доменам. Характерным примером и прототипом междоменных Связанных данных является DBpedia – набор данных, автоматически извлекаемый из общедоступного ресурса Википедия. Одним из перспективных направлений использования Связанных данных являются приложения в области образования, науки и культуры. Существенное продвижение имеется, например, в задачах интеграции библиотечных каталогов на глобальной основе, установления взаимных связей между содержанием библиотечных каталогов по теме, географическому расположению или историческому периоду, установление взаимных связей между библиотечными каталогами и данными третьих сторон (архивы изображений и видео, базы знаний типа DBpedia). Научные коллекции представлены в облаке LOD пока достаточно фрагментарно и поддерживаются, например, в наборах данных

[1] http://lod2.eu
[2] http://www.w3.org/RDF/
[3] http://www.w3.org/DesignIssues/LinkedData.html

ACM[4], DBLP[5] и CiteSeer[6]. Основным содержимым этих наборов данных являются стандартные метаданные статей (название, год публикации, информация об авторах и др.).

В статье рассматриваются программные технологии связывания данных из математического RDF-набора, подготовленного на основе статей математического журнала «Известия ВУЗов. Математика», издаваемого в Казанском (Приволжском) федеральном университете. Можно ожидать, что публикация в облаке LOD данных математической коллекции позволит распространить концепцию открытых данных в математическом сообществе, поможет профессиональным исследователям и студентам получить удобный доступ к специальным знаниям, интегрированным в глобальную систему знаний.

1. Стек программных технологий подготовки математического RDF-набора данных

Программные технологии подготовки математического RDF-набора данных подробно описаны в [Невзорова и др., 2012].

В основе подхода лежит представление коллекции математических документов в виде единого семантического графа, в котором как вершины – объекты математического знания, так и ребра – связи между ними, определяются множеством специализированных словарей (онтологий), которые специфицируют как термины из области математики, так и элементы логической структуры математического документа.

Стек программных технологий, разработанный для публикации математических данных в LOD, включает набор программных систем, предназначенных для выполнения следующих задач:

- преобразование формата данных;
- аннотирование текста;
- семантическое аннотирование;
- аннотирование метаданных;
- генерация RDF-набора;
- связывание данных.

В статье рассматриваются основные результаты по связыванию данных существующими инструментами и проблемы связывания с наиболее популярными Интернет-ресурсами.

2. Программные инструменты связывания данных

В Linked Data основополагающим принципом связывания идентичных ресурсов является использование ссылок на другие URI для поиска фактов относительно интересующего объекта. Типичным решением является формирование ссылочного утверждения owl:sameAs между различными семантически похожими URI, обозначающими эти ресурсы.

Целью процесса сопоставления является поиск семантически связанных сущностей и установление типовых ссылок между ними, большинство из которых имеют формальные свойства (такие как транзитивность, симметрия и т.д.), которые могут использоваться в механизмах рассуждений и других приложениях для выведения новых знаний.

Можно выделить две категории систем и методов выявления связей между сущностями (классами, свойствами или экземплярами) баз знаний:

- первая категория использует методики сопоставления онтологий и имеет целью установление ссылок между онтологиями, лежащими в основе двух источников данных;
- вторая, более сложная категория занимается попарным сопоставлением семантически сходных экземпляров и выявляет наличие связей между экземплярами в различных источниках данных.

В статье рассматривается задача выявление связей между экземплярами объектов, заданных различными URI.

Для указанных целей предлагается применять подходы, реализованных в программных системах SILK[7] и LIMES[8]. Далее кратко рассмотрим решения по связыванию данных в системах LIMES и SILK.

2.1. Система LIMES

Система LIMES (Link Discovery Framework for MEtric Spaces) - система выявления связей для метрических пространств – ориентирована на решение задачи сопоставления экземпляров на основе эффективных (по времени) методов выявления связей между источниками Связанных данных. Проблема масштабируемости процесса выявления связей решается с использованием аксиомы треугольника в метрических пространствах для определения оценок сходства экземпляров. Официальная страница системы LIMES находится на сайте исследовательской группы Agile Knowledge Engineering and Semantic Web (AKSW) по адресу[11].

Общая последовательность выполняемых функций по решению задачи связывания в системе LIMES представляется в виде схемы (рис.1). В систему поступает информация об исходном источнике данных (Source), целевом источнике данных (Target) и спецификации устанавливаемой связи (Link Specification) в виде конфигурационного файла.

[4] http://acm.rkbexplorer.com/
[5] http://dblp. rkbexplorer.com/
[6] http://citeseer.rkbexplorer.com/

[7] http://www4.wiwiss.fu-berlin.de/bizer/silk/
[8] http://aksw.org/Projects/limes

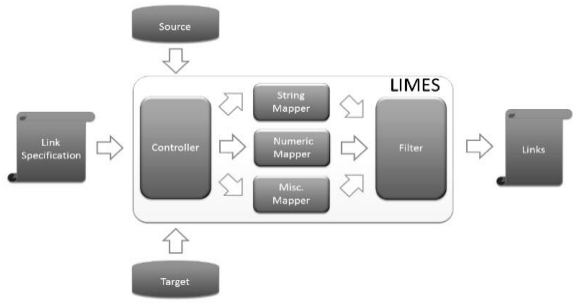

Рисунок 1 – Функции системы LIMES

Информация поступает в контроллер (Controller), где происходит разделение данных из источников по разным типам. Далее строки обрабатываются в строковом преобразователе (String Mapper). Числовые значения (а также все значения, которые могут быть эффективно преобразованы в векторное пространство) отображаются в метрическое пространство и обрабатываются с помощью алгоритма HYPPO (Numeric Mapper). Все прочие значения преобразуются в смешанном преобразователе (Miscellaneous Mapper). Результаты работы всех преобразователей фильтруются и объединяются с помощью эффективных по времени алгоритмов (Filter). Затем генерируются связи (Links).

Для связывания данных пользователю системы LIMES необходимо предоставить следующую информацию: источники данных; объекты, подлежащие связыванию, типы объектов для связывания; условия связывания объектов. Вся эта необходимая информация должна быть представлена в виде конфигурационного файла, написанного на конфигурационном языке LIMES (LIMES Configuration Language). На рис. 2 представлен графический интерфейс конфигурационного файла. В версии LIMES 0.5RS1 пользователь заполняет соответствующие поля формы, после чего конфигурационный файл формируется автоматически.

Назначение основных полей графического интерфейса:

- **Endpoint** – точка доступа SPARQL к источнику данных. Версия LIMES 0.5RS1 поддерживает файлы формата CSV, N3 (для файла-источника данных в поле Endpoint указывается абсолютный путь);
- **Var** – переменная, обозначающая источник данных;
- **Pagesize** – максимальное количество триплетов, возвращаемое точкой доступа SPARQL на каждый запрос. Если это количество не ограничено или источником данных является файл, значение поля Pagesize устанавливается равным -1;
- **Restriction** – ограничения на входные данные, позволяющие уменьшить количество сравнений, например, за счет указания принадлежности экземпляров данных к определенному классу, а также выполнения стандартной предобработки данных;
- **Property** – указание классов (типов объектов), которые будут участвовать в связывании.

Выбор типов объектов, которые могут участвовать в связывании, осуществляется на этапе описания источников данных. Правила идентификации объектов являются важной составляющей конфигурационного файла.

В версии LIMES 0.5RS1 для составления условия связывания используются поля, показанные на рис. 3. В поле Metric описывается метрика, которая вычисляется для пары объектов. Система LIMES поддерживает набор встроенных метрик:

- для строк – расстояние Левенштейна (Levenshtein), триграммы (Trigrams), коэффициент перекрытия (Overlap), коэффициент Жаккарда (Jaccard), косинусный коэффициент (Cosine);

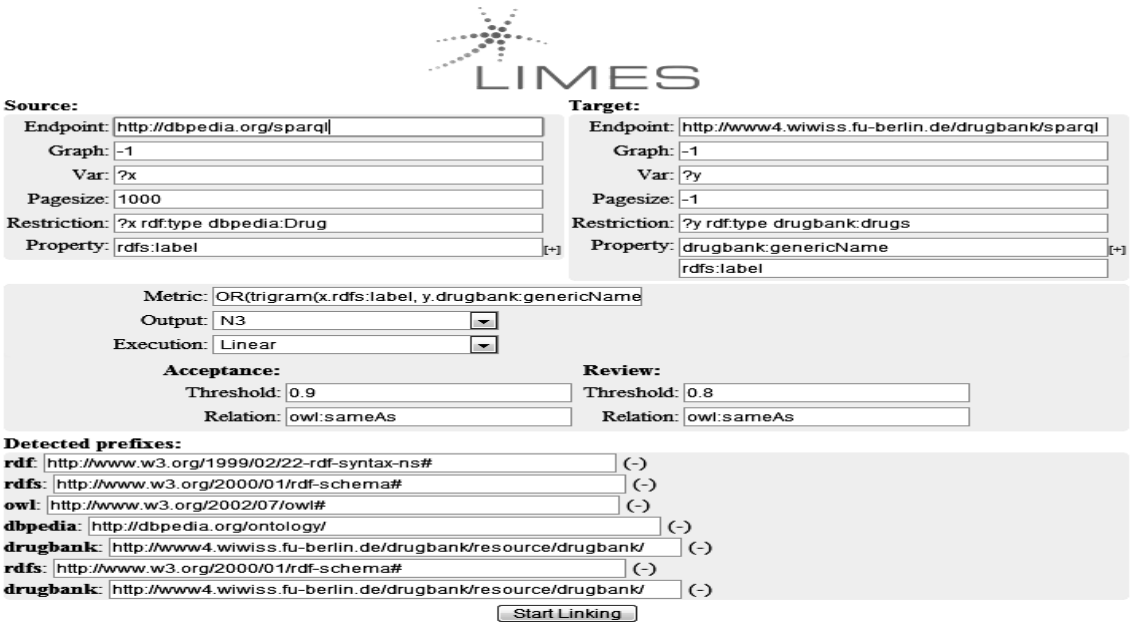

Рисунок 2 – Графический интерфейс создания конфигурационного файла

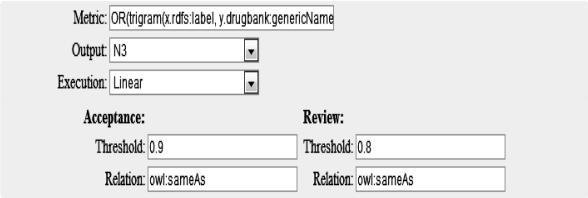

Рисунок 3 – Поля для составления условия связывания данных

- для числовых векторов – Евклидово расстояние (Euclidean), wgs84.

Система LIMES оперирует формальными выражениями для построения сложных метрик путем комбинирования метрических и логических бинарных операций. К метрическим операциям относятся встроенные метрики, а также операторы нахождения минимума (MIN), максимума (MAX), суммы (ADD), произведения (MULT) и др. Логические операции (AND, OR, DIFF и др.) позволяют объединять или фильтровать результаты метрических операций.

В поле Output назначается формат результирующих данных (поддерживаются форматы N3 и TAB). В разделе Acceptance указывается значение порога метрики (поле Threshold), при достижении которого устанавливается связь для данной пары объектов и результат записывается в выходной файл. В разделе Review указывается порог, значение которого меньше, чем в разделе Acceptance. Соответственно, связи, для которых значение метрики попадет в промежуток между этими двумя порогами, будут записаны в другой файл для последующей ручной обработки. В поле Relation указывается тип связи объектов, например, owl:sameAs.

Результатом работы системы LIMES является файл, содержащий данные об установленных связях.

В разделе 3 подробно рассматриваются эксперименты по связывания данных математического RDF-набора данных в системе LIMES.

2.2. Система SILK

Официальная страница проекта находится на сайте Свободного Университета Берлина: http://www4.wiwiss.fu-berlin.de/bizer/silk/.
Система SILK предназначена для нахождения отношений между элементами из разных наборов Связных Данных. Система принимает на входе схему связывания, на выходе возвращает набор найденных связей. Для связывания данных пользователю системы SILK необходимо предоставить следующую информацию: источники данных; объекты, подлежащие связыванию, типы объектов для связывания; условия связывания объектов.

Взаимодействие с системой SILK осуществляется посредством командной строки (Silk Single Machine), либо с помощью графического интерфейса (SILK Workbench).

Источники данных задаются указанием SPARQL-точки доступа, либо как локальный RDF-файл в разных форматах (RDF/XML, N3, Turtle и др.).

Схема связывания задается как XML-файл в формате Silk-LSL (Silk Link Specification Language), либо конструируется при помощи графического интерфейса.

3. Эксперименты по связыванию математического RDF-набора данных в системах LIMES и SILK

В подготовленном наборе данных IVM [Невзорова и др., 2012] представлены метаданные 1456 статей журнала «Известия ВУЗов. Математика», издаваемого в Казанском федеральном университете, за 1997-2009 годы, которые включают:

- название публикации (на русском и английском языках), выпуск журнала и его год, страницы в журнале, список литературы (сокращенный), авторы;
- автор публикации: имя (на русском и английском языках, сокращенная и полная формы), адрес электронной почты, адрес страницы на ресурсе mathnet.ru, организация;
- организация (как место работы): название (несколько вариантов написания, на русском и английском);
- адрес публикации на ресурсе mathnet.ru.

Для описания метаданных статей используется популярная схема AKT Reference Ontology (AKT). В наборе IVM представлено около 24000 триплетов метаданных. Доступ к данным IVM осуществляется через локальный RDF-файл.

Для представления математических знаний в ходе разработки проекта по математическому RDF-набору IVM с помощью экспертов-математиков Казанского федерального университета была подготовлена онтология профессиональной математики OntoMath^{pro}. Главная цель OntoMath^{pro} – предоставить информационный терминологический ресурс для автоматизированной обработки электронных профессиональных математических публикаций на русском языке. Онтология OntoMath^{pro} содержит определения как общепринятых математических понятий, так и развивающуюся терминологию, в основном, из следующих разделов математики: теория чисел, теория множеств, алгебра, геометрия, математическая логика, дискретная математики, теория алгоритмов, математический анализ, дифференциальные уравнения, численные методы, теория вероятностей и математическая статистика. Источниками для определения семантики концептов OntoMath^{pro} служили: классические учебники соответствующих разделов математики,

электронные ресурсы – Wikipedia и Cambridge Mathematical Thesaurus, научные статьи журнала «Известия вузов. Математика», а также профессиональные знания и опыт экспертов-математиков.

В качестве языков представления OntoMath^{pro} выбраны языки OWL-DL/RDFS, которые предоставляют высокие выразительные логические средства и алгоритмически разрешимы не только теоретически, но и практически, с помощью возможностей таких инструментов логического вывода, как Pellet и Fact++. В частности, разделы математики и элементы математического знания выражались с помощью концепта owl:Class. Уникальный идентификатор ресурса (URI) каждого класса представляет собой суррогатный ключ, который составлен из пространства имен онтологии и кода, однозначно идентифицирующего класс внутри онтологии.

Опишем ряд экспериментов по связыванию, выполненных с помощью инструментальных средств SILK и LIMES для RDF-набора IVM и онтологии OntoMath^{pro}.

3.1 Связывание научно-исследовательских организаций в системе LIMES

Связывание данных о научно-исследовательских организациях выполнено для RDF-набора IVM и набора связанных данных CORDIS (Community R&D Information Service: RKBExplorer).

Доступ к данным CORDIS может осуществляться как через локальные RDF-файлы, так и через SPARQL-точку доступа. Список ссылок на локальные RDF-файлы содержится по адресу: http://dblp.rkbexplorer.com/sitemap.xml. Каждый из файлов содержит информацию о научных организациях, сотрудники которых имеют в соответствующий год представленную в CORDIS публикацию. SPARQL-точка доступа имеет адрес: http://cordis.rkbexplorer.com/sparql/.

Многие элементы из IVM имеют эквивалентные им элементы из CORDIS, т.е. соответствуют одной и той же реальной организации. Необходимо найти пары эквивалентных элементов и соединить их ссылками owl:sameAs. Эквивалентность организаций определяется по схожести только одного признака — названия организации.

Оба набора данных используют онтологию AKT, организации в обоих наборах имеют тип akt:Organization. В наборе IVM название организации задается свойством akt:name, в наборе CORDIS — свойством akt:has-pretty-name.

На рис. 4 приведено решение по связыванию набора IVM и локального файла cordis-projects-2004.rdf набора CORDIS, подготовленного с помощью графического интерфейса системы LIMES.

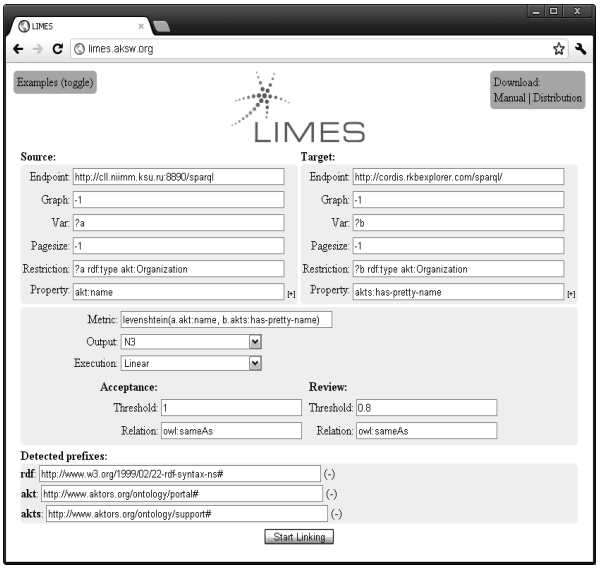

Рисунок 4 - Решение по связыванию наборов IVM и Cordis

Построенный список содержит 87 найденных гарантированно правильных связей при ограничениях на размер выборки из SPARQL-точки доступа Cordis.

3.2. Связывание классов онтологии OntoMath^{pro} с ресурсами DBPedia в системе SILK

Связывание классов онтологии OntoMath^{pro} с данными из ресурсов DBPedia осуществлялось на основе следующих признаков:

- название класса (rdfs:label). Название класса из OntoMath^{pro} сравнивается с аналогичным названием ресурса из DBPedia;
- ссылка на статью английской Википедии в поле комментария (rdfs:comment). Адрес ссылки сравнивается со значением входящего свойства foaf:primaryTopic ресурса из DBPedia;
- ссылка на статью русской Википедии в поле комментария (rdfs:comment). Название статьи (т.е. та часть адреса ссылки, которая идет после «http://ru.wikipedia.org/wiki/») сравнивается с названием ресурса (rdfs:label) из DBpedia.

При этом из всего множества ресурсов DBpedia отбираются для связывания только ресурсы из категории *category:Mathematics* и категорий разделов математики до 4-го уровня вложенности. Введенное ограничение позволяет обеспечить необходимые требования по точности и полноте связывания. Например, точность можно повысить за счет исключения из рассмотрения связей с классами-омонимами из других (нематематических) областей знаний.

Например, #E64 (Якорь из области теории групп) ≠ dbpedia:Anchor (Морской якорь).

Ограничение на уровень вложенности ресурсов вызвано рядом причин. По техническим причинам в системе SILK невозможно задать транзитивный запрос на выборку ресурсов с произвольным уровнем вложенности. Поэтому используется

запрос, в котором уровень вложенности фиксирован (ограничен 4 уровнями) и в котором явно перечисляются условия на выборку ресурсов из категории *category:Mathematics*. Можно также отметить ряд проблем, связанных с транзитивностью системы категорий DBpedia.

Иерархия категорий DBpedia является несовершенной и в ней часто нарушается принцип транзитивности, то есть ресурс, относящийся к достаточно вложенной подкатегории некоторой категории, к самой категории может и не относится.

Например, категория *category:Algebra* содержит вложенную подкатегорию *category:Museu_Picasso* в иерархии Algebra → Convex_geometry → ... → Cubes → Cubism → Pablo_Picasso → Museu_Picasso.

Более того, некоторые категории содержат очень общие подкатегории (Space, Time, Structure), и поэтому, на очень глубоком уровне вложенности содержат почти все ресурсы DBpedia.

Онтология OntoMath^{pro} содержит 3450 классов. В результате процесса связывания классов онтологии с ресурсами DBpedia были найдены 842 связи. Число связанных классов составило 828 (некоторые классы были связаны сразу с несколькими ресурсами DBpedia).

Для оценки полноты связывания (при ограничении до 4 уровней иерархии**)** была выполнена оценка количества ресурсов, участвующих в связывании при увеличении уровней иерархии до 5-8. В результате не была найдена ни одна дополнительная связь, что обосновывает выбранное ограничение глубины, которое существенно не влияет на полноту связывания.

Оценки точности связывания составляют 95% (примерно 43 связи (5%) были оценены как некорректные или не совсем корректные). Эти некорректные и не совсем корректные связи возникли по следующим причинам:

- неточные ссылки в онтологии (ссылка на статью Википедии для связанных или более общих понятий). Например, #E203 (Компактный слой) ≠ dbpedia:Novikov's_compact_leaf_theorem (Теорема о компактном слое), или пример связи с более общим понятием: #E3004 (Признак Лейбница сходимости знакочередующегося ряда») ≠ dbpedia:Alternating_series (Знакочередующийся ряд);
- некорректные интервики (ссылки между языковыми разделами) в Википедии. Например, #E1317 (Сравнение по модулю натурального числа) ≠ dbpedia:Modular_arithmetic (Модульная арифметика);
- наличие омонимичных статей в пределах четвертого уровня вложенности, например, #E1408 (Отображение спаривания) ≠ dbpedia:Mating (Сексуальное спаривание), которое попало в категорию category:Statistics через следующую цепочку подкатегорий: category:Fertility → category:Demography → category:Fields_of_application_of_statistics → category:Statistics.

Заключение

Публикация данных в пространстве Linked Open Data позволяет каталогизировать данные, повысить ценность данных за счет связывания их с другими данными, понизить степень дублирования данных, облегчить доступ к данным заинтересованным сторонам. Процесс связывания данных выполняется специальными инструментами связывания, которые в интерактивном режиме позволяют формировать различные схемы связывания данных, осуществлять настройку методов связывания, в том числе управляя ограничениями на условия связывания данных.

В результате проделанной работы были выявлены ряд важных ограничений на связывание данных с ресурсами DBPedia, получены оценки точности и полноты связывания данных с данным ресурсом, изучены основные технологические решения систем связывания SILK и LIMES. Проведенные эксперименты выявили также ряд особенностей математического RDF-набора, затрудняющего эффективное связывание, что позволило наметить пути улучшения качества построенных решений.

Благодарности

Работа выполнена при финансовой поддержке РФФИ, грант № 11-07-00507а.

Библиографический список

[Невзорова и др., 2012] Невзорова О.А., Жильцов Н.Г., Заикин Д.А., Жибрик О.Н., Кириллович А.В., Невзоров В.Н., Биряльцев Е.В. Прототип программной платформы для публикации семантических данных из математических научных коллекций в облаке LOD // Ученые записки КГУ. Серия Физико-математические науки. 2012. В печати.

DATA BINDING TECHNOLOGY IN THE SPACE OF OPEN DATA ON THE EXAMPLE OF MATHEMATICAL COLLECTIONA

Nevzorova O.A., ** Kirillovich A.V. *

** Kazan Federal University, Kazan, Russia*

alik.kirillovich@gmail.com

*** Research Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic, Kazan Federal University, Kazan, Russia*

onevzoro@gmail.com

The paperis presented the software technologies of data binding in the space of Open data with use of LIMES and SILK systems, and experiments on the binding of mathematical RDF-dataset.