УДК 004.82

Г.Ю. Проскудина, К.А. Кудим

О ТЕХНОЛОГИИ ИСПОЛЬЗОВАНИЯ ВНЕШНИХ ДАННЫХ ПРИ СОЗДАНИИ И РЕДАКТИРОВАНИИ ЭНЦИКЛОПЕДИЧЕСКИХ ТЕКСТОВ

В работе обсуждается развивающийся проект Викиданные, веб-сервис запросов и язык запросов. Работа веб-сервиса, языка запросов и форм вывода результатов демонстрируется на многочисленных примерах. Разработана технология использования Викиданных сторонними системами. Учитывая это, рассматривается расширение ExternalData, разработанное для программного обеспечения MediaWiki. В ходе тестовой эксплуатации расширение ExternalData было доработано. Расширение используется для вставки запросов данных к внешним источникам, в нашем случае к базе знаний Викиданные, и их результатов в вики-разметку создаваемых текстов статей. Разработана процедура создания страницы-списка.

Ключевые слова: электронная энциклопедия, база Викиданные, сервис запросов, язык запросов.

Введение

Технология создания и поддержки функционирования *электронной Большой украинской энциклопедии*[1] предполагает оперативное и своевременное ее наполнение из уже имеющихся внешних источников. Важным и необходимым источником данных может стать *Википедия*, *Свободная энциклопедия* – один из крупнейших и популярнейших веб-сайтов мира[2].

Концепция Википедии хорошо известна и довольно проста – это открытая энциклопедия знаний, где любой человек может вносить и редактировать информацию. На сегодняшний день ее англоязычный раздел (английская Википедия) насчитывает более 5 млн. статей, 11 Википедий, в том числе немецкая, французская, русская, итальянская и некоторые другие насчитывают от 1 до 4 миллионов статей. Украинский раздел Википедии по состоянию на 9 января 2017 года содержит 671 977 статей различной тематики; по данному показателю украинская Википедия занимает 16 место из более чем 287 языковых разделов.

В настоящее время незаметно для большинства своих читателей эта популярная онлайновая система претерпевает значительные изменения, учитывая то, что связанный с Википедией проект, *Викиданные*[3], управляя фактологической информацией Википедии, вводит новую многоязычную так называемую "*Википедию данных*" – свободную и открытую базу данных, которая собирает многоязычные структурированные данные Википедии в одном месте и представляет их в общий свободный доступ для копирования, использования и распространения, открывая новые возможности для многих других приложений[4].

Как и Википедия, Викиданные организованы постранично. Каждый экземпляр данных имеет свою страницу, на которой можно редактировать его свойства. Свойства описывают объект и связывают его с другими страницами данных, например, с классом, представителем которого он является, как то: персона, место, событие и многие другие. Например, элемент города *Рим* может иметь свойство *Население* со значением 2.777.979. Можно задавать горные вершины, места и географические координаты зданий. Можно связать человека с его или ее местом рождения, родом занятий или с его номером в базе данных органа управления; связать поли-

[1]Работа выполнена в рамках первого этапа проекта "Разработка проекта компьютерного варианта и технологии создания и поддержки функционирования Большой украинской энциклопедии" программы информатизации НАНУ.

[2]https://ru.wikipedia.org/wiki/Википедия

[3]https://ru.wikipedia.org/wiki/Викиданные

[4]https://www.wikidata.org/wiki/Wikidata:Introduction/ru

тика со своей политической партией; связать населенный пункт со своей более высокой административной единицей; связать страну с ее высшим руководством и ее национальным гимном и т. д. Данная информация может быть показана на любом языке, даже если данные собраны на другом языке. При доступе к этим значениям вики-клиент покажет самые последние данные в актуальном состоянии.

Сегодня почти каждая страница Википедии на разных языках включает в себя содержимое от Викиданных. Все больше и больше редактируемых вручную *инфобоксов* (таблицы с основной, фактической информацией по теме статьи) используют Викиданные в качестве базы данных серверной части, поэтому отображаемая информация будет одинакова во всех изданиях Википедии.

Цель данной работы – использовать внешний надежный источник данных такой, как Википедия, при создании страниц электронной Большой украинской энциклопедии. Для чего предлагается технология включения внешних данных, которую можно рассматривать в качестве одной из составляющих технологии создания статей для энциклопедии. Это позволяет на своем ресурсе запрашивать нужную информацию из Википедии и формировать на основе полученных данных свои страницы или фрагменты страниц.

Разработке такой технологии и посвящена настоящая работа. Здесь подробно обсуждается новый перспективный и развивающийся проект Википедии – Викиданные, веб-сервис запросов к базе данных Викиданные и язык запросов. Работа вебсервиса, языка запросов и форм вывода результатов демонстрируется на многочисленных примерах. Затем речь пойдет о том, как можно использовать Викиданные в сторонних системах, а не только в Википедии. При этом используется разработанное для программного обеспечения MediaWiki расширение ExternalData для вставки запросов данных к *внешним источникам* (в нашем случае к базе данных Викиданные) и их результатов в викиразметку создаваемых текстов статей. В ходе тестовой эксплуатации расширение ExternalData было доработано. Разработана и приводится процедура создания страницы-списка.

В нашей экспериментальной энциклопедии[5] можно посмотреть ряд статей, имеющих на своих страницах запросы к внешней базе данных Викиданные: "Реки и озера Украины", "Национальные парки Украины", "Города Украины", “ВУЗы Украины", "Писатели и поэты, родившиеся в Украине", "Знаменитые музыканты Украины", "Выдающиеся ученые и инженеры Украины".

1. Викиданные

Первоначально задуманная как текстовый ресурс, Википедия собирает растущее количество *структурированных данных*: числа, даты, координаты, разные типы отношений. Эти данные стали ресурсом огромной ценности с потенциальным применением во всех областях науки, техники и культуры.

Такое развитие событий не удивительно, учитывая, что Википедия представляет "всеобщее видение мира, в котором каждый человек может свободно обмениваться суммой всех знаний". Тогда не шла речь о том, что эти знания должны включать в себя данные, которые могут быть найдены, проанализированы и использованы повторно.

Может вызвать удивление тот факт, что Википедия не обеспечивает прямого доступа к большинству этих данных, ни через сервисы запросов, ни через выгружаемый экспорт данных. Фактически использование данных происходит редко и часто ограничено очень специфическими частями информации, такими, например, как гео-теги статей Википедии, используемые в Google Maps. Причина этого поразительного разрыва между видением и реальностью является то, что данные Википедии спрятаны внутри 30 млн. статей на 287 языках, откуда очень трудно извлечь нужную информацию.

Такая ситуация, во-первых, не устраивает тех, кто будет использовать данные, но и, во-вторых, возрастает опас-

[5] http://sew.isofts.kiev.ua

ность для основной цели Википедии – обеспечение современными и точными энциклопедическими знаниями. Одна и та же информация часто появляется в статьях на разных языках и во многих статьях в пределах одного языка. Численность населения Рима, например, можно найти в английской и итальянской статье о Риме, но также и в английской статье о городах Италии. Все эти цифры могут быть отличаться друг от друга.

Цель Викиданные – преодолеть эти проблемы путем создания новых средств Википедии управлять своими данными в глобальном масштабе. Результат этих продолжающихся усилий можно увидеть на сайте wikidata.org.

Викиданные – самый новый проект Викимедиа, это совместно редактируемая, свободная база знаний, которую можно читать и редактировать людьми и машинами. Хороший обзор на эту тему представлен в [1]. На данный момент достигнуты следующие результаты:

– централизация связей между разноязычными изданиями Википедии и другими сайтами проекта Викимедиа. К примеру все статьи Википедии об "энциклопедии" (на любом языке) связаны с одним элементом Викиданных с идентификатором Q5292. Эти так называемые *ссылки на сайты* и другие данные о сущности, известной как "энциклопедия", можно посмотреть на странице https://www.wikidata.org/wiki/Q5292;

– централизация инфобоксов. Все больше и больше измененных вручную инфобоксов, таблиц с основной, фактической информацией по теме статьи, намереваются использовать Викиданные в качестве базы данных серверной части, поэтому отображаемая информация будет одинакова во всех изданиях Википедии;

– обеспечение интерфейса для различных запросов. Содержание Викиданных можно запросить через открытый интерфейс SPARQL на сервисе https://query.wikidata.org. В дальнейшем результаты запроса планируется интегрировать на страницы в Википедии и других проектов, как списки, таблицы, карты и другие формы.

Модель данных Викиданных не реляционная и не на основе RDF, хотя и существуют отображения в RDF, но она отражает стратегию Викиданных на хранение утверждений вместо фактов. Каждое утверждение должно быть получено с помощью ссылок, а противоречивые утверждения намеренно не запрещены. Утверждения могут дополнительно контролироваться уточнителями, такими как домен или дата действия, в конечном счете, поддерживая п-арные отношения между сущностями (элементами) Викиданных. Свойства Викиданных определяются консенсусом сообщества. Например, P571 идентифицирует свойство *зарождение* (*inception*), чтобы заявить дату, когда-то что-то было создано или основано. Названия (labels) и область применения (scope notes) могут быть отредактированы независимо от утверждений с поддержкой синонимов и омонимов.

1.1. Руководящие принципы базы данных Викиданные. Приведем перечень конструктивных решений, характеризующий подход, принятый в базе данных Викиданные.

Открытое редактирование. Также как и Википедия, Викиданные позволяют каждому пользователю сайта расширять и редактировать сохраненную информацию, даже без создания учетной записи. Интерфейс на основе форм делает редактирование легким и удобным.

Контроль сообщества. Под контролем сообщества вкладчиков находятся не только фактологические данные, но и схема данных. Авторы, редактирующие численность населения Рима, в первую очередь на их взгляд вносят самое правильное число.

Множественность. Поскольку многие факты оспариваются или просто не определены, было бы наивно ожидать глобального соглашения об "истинных" данных. Викиданные позволяют противоречивым данным сосуществовать и обеспечивают механизмы для организации такого множества данных.

Вторичные данные. Викиданные собирают факты, опубликованные в пер-

вичных источниках, вместе со ссылками на эти источники. Там нет такого понятия, как "истинное население Рима", но есть – "население Рима, опубликованное в городе Риме в 2011 году".

Многоязычность данных. Большинство данных не привязаны к одному языку: цифры, даты и координаты имеют универсальное значение; заголовки (labels), например, Рим или Население, переведены на многие языки. Викиданные – это многоязычный проект. Есть только один сайт Викиданные, в то время как Википедия имеет независимые издания для каждого языка, т.е. Википедия имеет множество сайтов.

Легкий доступ. Цель Викиданные – предоставлять данные не только Википедии, но и другим внешним приложениям. Данные экспортируются через вебсервисы или API в нескольких форматах, включая XML, JSON, RDF. Данные публикуются в соответствии с юридическими условиями по лицензии CC0[6], позволяющие максимально широкое повторное использование.

Непрерывная эволюция. В лучших традициях Википедии, Викиданные растут вместе с сообществом и задачами. Вместо того, чтобы разработать совершенную систему, которая была бы представлена миру через несколько лет, новые возможности разворачиваются постепенно и как можно раньше. Все эти свойства характеризуют Викиданные как специфический вид curated (специально отобранных) баз данных [2].

1.2. Краткая история проекта Викиданные. Проект Викиданные был запущен в октябре 2012 года. Тогда редакторы могли только создавать элементы (items) и соединять их со статьями Википедии. В январе 2013 года три Википедии, сначала венгерская, затем еврейская (на иврите) и итальянская, подключились к Викиданные. Между тем, сообщество уже создало более трех миллионов элементов. В феврале присоединилась английская Википедия, а в марте 2013 года уже все существующие Википедии были подключены к БД Викиданные.

По состоянию на февраль 2014 года Викиданные получали информацию от более чем 40 тыс. участников. Начиная с мая 2013 года с Викиданные постоянно работали более 3.5 тыс. активных участников – это те вкладчики, которые делают по крайней мере пять изменений в течение месяца. Учитывая это можно сделать вывод, что в настоящее время Викиданные – один из наиболее активных проектов Викимедиа.

В марте 2013 года в качестве языка сценариев Википедии введен язык Lua[7], который может использоваться для автоматического создания и обогащения некоторых частей статьи, например, упомянутых инфобоксов. Скрипты Lua могут получить доступ к Викиданные, позволяя редакторам Википедии извлекать, обрабатывать и отображать эти данные.

В настоящее время продолжаются работы, относящиеся к поддержке произвольных поисковых запросов с возможностью использовать их результаты при автоматическом обновлении различных списков в статьях Википедии.

1.3. Из многих, один. Первоначальной задачей Викиданных было согласовать 287 языковых разделов Википедии [1]. Для Викиданных, чтобы быть действительно многоязычными, объект, представляющий «Рим», должен быть одним и тем же во всех языках. К счастью, Википедия уже имеет механизм тесно связанный с этим вопросом: ссылки на язык, отображаемые слева каждой статьи, соединяют статьи на разных языках. Эти ссылки были созданы из отредактированных пользователем исходных текстов в нижней части каждой статьи, приводят к квадратному числу таких ссылок по теме: каждая из 207 статей о Риме содержит список из 206 ссылок на все другие статьи о Риме – это в общей сложности 42,642 строк текста. В итоге до появления проекта Викиданные в статьях Википедии содержалось больше

[6] https://creativecommons.org/choose/zero/

[7] https://www.lua.org/pil/p1.html

текста для разноязычных ссылок, чем фактического содержания статьи.

Соответственно, лучше хранить и управлять разноязычными ссылками в одном месте, и это была первая задача для Викиданных. Для каждой статьи Википедии, создана отдельная страница на Викиданных, где ссылки на соответствующие статьи Википедии указаны на разных языках. Такие страницы на Викиданных называются *элементами* (items). Первоначально для каждого элемента могло храниться только ограниченное число данных: список языковых ссылок-связей, имена, список псевдонимов, краткое описание. Имена, псевдонимы и описания могут определяться отдельно (на данный момент до 358 языков).

Сообщество Викиданных создало *боты* для того, чтобы переместить языковые ссылки из Википедии в Викиданные, вследствие чего из Википедии можно было удалить более 240 миллионов ссылок. И сегодня, большинство языковых ссылок, отображаемых в статьях Википедии, подаются с Викиданных. В статью еще можно добавить пользовательские ссылки, которые необходимы в тех редких случаях, когда ссылки не являются двунаправленным, например, некоторые статьи связаны с более общими статьями на других языках, при этом Викиданные намеренно соединяет только те страницы, которые охватывают один и тот же предмет. Импортируя языковые ссылки, Викиданные получили огромное множество исходных элементов, которые "обоснованы" реальными страницами Википедии.

1.4. Модель данных. Как и Википедия, Викиданные организованы постранично, и такая организация также совпадает со структурой самих данных [3]. Каждый *предмет* (subject), по которому Викиданные структурируют свои данные называется *сущностью* (entity), и каждая сущность имеет свою страницу. Система пока что различает два типа сущностей: *элементы* (items) и *свойства* (properties). Практически каждая статья Википедии на любом языке имеет соответствующий элемент, представляющий собой предмет (или тему) данной статьи. Каждый элемент имеет страницу, на которой пользователи могут просматривать и вводить данные. Так, например, страницу элемента английского писателя Дугласа Адамса можно увидеть по ссылке: https://www.wikidata.org/wiki /Q42 (рис. 1).

Каждый элемент Викиданных имеет *название* (label), *описание* (description) и, вероятно, один или несколько *псевдонимов* (aliases). *Ссылки на сайты* (sitelinks) связывают каждый элемент с соответствующими статьями на всех клиентах Википедии. *Утверждения* (statements) описывают детальные характеристики для каждого элемента. Каждое утверждение состоит из *свойства* (property) и его *значения* (value).

В нашем примере, название страницы – "Q42", а не "Дуглас Адамс", так как Викиданные – многоязычный сайт. Поэтому элементы не идентифицируются названием на конкретном языке, а нейтральным идентификатором элемента, который автоматически назначается при его создании, и который не может быть изменен в дальнейшем. Идентификаторы элемента всегда начинаются с буквы "Q" с последующим числом. Каждая страница элемента содержит следующие основные части:

– название (например, "Дуглас Адамс");

– краткое описание (например, "английский писатель и юморист");

– список псевдонимов (например, "Дуглас Ноэль Адамс");

– список утверждений (самая обширная часть данных, см. далее);

– список ссылок на сайты (ссылки на страницы Википедии и другие проекты).

Первые три части данных (название, описание, псевдонимы) известны под общим названием термины. Они в основном используются для поиска и отображения элементов. Элемент может иметь название на любом языке, поддерживаемом Викиданными. То, что отображается на страницах, зависит от настройки языка

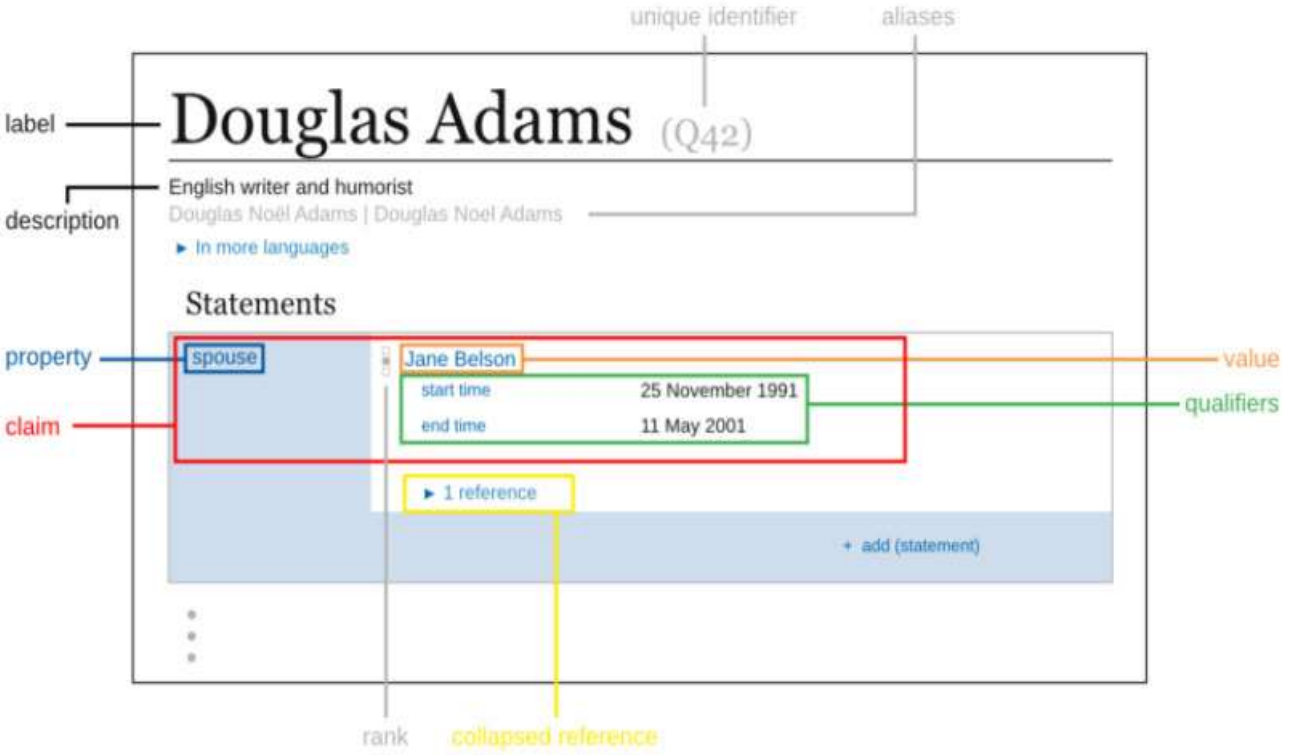

Рис. 1. Пример элемента Викиданные "Дуглас Адамс" Q42

пользователя. Ссылки могут быть предоставлены для любой из 286 языковых версий Википедии, а также для нескольких родственных проектов, таких например, как Викигид и Викисклад. Ссылки на сайты являются функционалом (не более одной ссылки на сайт) и обратным функционалом (не более одного элемента для любой ссылки). В отличие от прежней системы языковых ссылок Википедии, ссылки должны использоваться только для статей, которые точно на эту тему, а не на более широкую или более узкую, или иным образом связанных с темой. Некоторые элементы не имеют каких-либо ссылок, например, элемент "Женский" (Q6581072), который используется в качестве возможного значения для пола лиц.

1.4.1. Свойства и типы данных. На рис. 2 показан простой пример утверждения (statement), которое близко напоминает RDF-тройку, где предмет – “Дуглас Адамс” (Q42), свойство (property) – супруга (spouse) и значение (value) – Jane Belson.

Свойства, как предметы, описаны на страницах и имеют идентификаторы, начинающиеся с "P". Например, свойство супруг(а) на самом деле P26. Свойства тоже имеют названия, псевдонимы и описания, но у них нет ссылок на сайты.

Кроме того, свойства Викиданных также имеют тип, ограничивающий принимаемое значения. Тип данных *дата рождения* – время, *супруг*(*а*) – связан с другим элементом. В табл. (рис. 2, [3]) в левой колонке приведен список всех допустимых типов данных. Общие медиа (Commons Media) представляет собой особый тип данных для ссылок на медиафайлы в хранилище медиа-ресурсов Викисклад, использующееся всеми Википедиями. Типы данных определяют структуру значений, принимаемых свойствами. Свойство может иметь простое значение (как в случае типа *элемент*, *item* или как для типа *строка*) или комплексное значение, которое требует несколько полей, как для времени, сферических координат и количества. Колонка справа (рис. 2) показывает возможные компоненты каждого значения.

Table 1. Wikidata datatypes and their current member fields and field types

Datatype	Member fields
Item	item id (IRI)
String	string
URL	URL (IRI)
Commons Media	article title (string)
Time	point in time (dateTime), timezone offset (int), preferred calendar (IRI), precision (byte), before tolerance (int), after tolerance (int)
Globe coordinates	latitude (decimal), longitude (decimal), globe (IRI), precision (decimal)
Quantity	value (decimal), lower bound (decimal), upper bound (decimal)

Рис. 2. Типы данных в Викиданных, их наборы и типы полей

Для времени, сохраняется дополнительное смещение (в минутах) для часового пояса и ссылка на календарную модель, являющаяся предпочтительной для отображения (например, по Юлианскому календарю, Q1985786). Также можно указать точность, чтобы выразить неопределенные значения, такие как "сентябрь 1547" или "3-е столетие". Детали здесь не существенны. Для наиболее распространенных типов неточностей (точность до дня, месяца, года), чтобы закодировать эту информацию непосредственно в литералах, задавая основной момент времени, используются специальные типы данных схемы XML (xsd:date, xsd:gYearMonth, xsd:gYear).

Для получения сферических координат, в таблице представлено необычное поле – *шар* (globe), которое задает небесное тело, например, координаты относятся к Земле (Q2).

1.4.2. Сложные утверждения и ссылки. Полная модель данных утверждений Викиданных немного сложнее, чем можно предложить из рис. 1. С одной стороны, утверждения могут иметь так называемые *квалификаторы* (или уточнители), предоставляющие дополнительную контекстную информацию для данного утверждения. С другой стороны, каждое утверждение может включать в себя одну или несколько ссылок, в поддержку этого утверждения. Утверждение, где представлены оба аспекта, показано на рис. 3.

Основная пара свойство-значение в данном утверждении – "spouse: Jane Belson" (P26: Q14623681), но здесь существует и контекстная информация.

Уточнители на рис. 3 – "дата начала: 25 ноября 1991" и "дата окончания: 11 мая 2011", утверждают, что Дуглас Адамс был женат на Джейн Белсон с 1991 года до своей смерти в 2011 году. Здесь используются свойства *дата начала* (P580) и *дата окончания* (P582) соответствующих типов времени. Эти пары свойство-значение относятся к основной части утверждения, а не к элементу на странице (Дуглас Адамс).

В Викиданных уточнители используются в нескольких ситуациях. Наиболее распространенным является указание на время действия утверждения, так что случай на рис. 3 – довольно типичный. Тем не менее, Викиданные использует многие другие виды аннотаций, предоставляющих контекстную информацию об утверждении. Например, *автор таксона* (P405, важный контекст для биологических названий таксонов) и таксономия *астероидов* (P1016, контекстуализировать

Рис. 3. Часть сложного утверждения о жене Дугласа Адамса,
как показано в Викиданных

спектральную классификацию астероидов). В некоторых случаях, уточнители-предоставляют дополнительные аргументы отношений, которые имеют более двух участников. Например, свойство (P553) учетная запись веб-сайта определяет веб-сайт (например, Twitter, Q918), но, оно как правило, используется с уточнителем P554, задающего имя учетной записи, используемого элементом на этом сайте. Можно утверждать, что это тернарные отношения, но граница между аннотацией контекста и n-арной связью размыта. Например, американский сериал, *Звёздный путь: Следующее поколение* (Q16290) имеет свойство *в ролях* (P161) *Брент Спайнер* (Q311453) с двумя значениями уточнителя *играет роль персонажа* (P453): *Дейта* (Q22983) и *Lore* (Q2609295). Обратите внимание, что такое же свойство может быть использовано в нескольких уточнителях в одном утверждении.

Мы использовали пары свойство-значение во многих местах: и в качестве основных частей утверждений, и в качестве уточнителей, и в ссылках. В каждом из этих случаев Викиданные также поддерживают два специальных “значения”: ни один и некоторый (none и some). Значение ни один используется, например, в утверждении: "Королева Англии Елизавета I не имела супруга". Что позволило получить простую форму для отрицания и отличить его от случаев, когда информация является просто неполной. Это также позволяет добавлять ссылки на негативные утверждения. Иногда это используется, когда известно, что это свойство имеет значение, но нет возможности предоставить более подробную информацию, как, например, в утверждении: "Папа Линус имел дату рождения, но она нам неизвестна". Оба данных специальных "значения" можно использовать во всех местах, где разрешены обычные значения свойств, поэтому они, как правило, не упоминаются в явном виде.

1.4.3 Порядок и ранг. Все данные в Викиданных упорядочены – псевдонимы, утверждения, пары свойство-значение в качестве ссылки и т. д. Информация о порядке в Викиданные используется только для представления, и не считается значимой для ответа на запросы.

Даже если в ответах на запрос не нужно использовать порядок утверждений, иногда необходимо выделять некоторые из утверждений от остальных.

Например, Викиданные содержат много исторических данных с подходящими уточнителями, например, численность населения городов в разное время. Такие данные имеют множество применений, но простой запрос для населения города не должен возвращать длинный список чисел. Чтобы упростить базовую фильтрацию данных, утверждениям Викиданных можно присвоить один из трех рангов: *нормальный* (используется по умолчанию), *привилегированный* (когда нужно выделить предпочтительные значения) и *вызывает возражение* (когда нужно пометить неправильно или непригодную информацию, но по какой-то практической причине ее хранят в системе).

1.5. Викиданные в цифрах. С момента своего запуска в октябре 2012 года база данных Викиданные значительно выросла. Некоторые статистические факты о ее текущем содержании показаны на рис. 4.

Статистика Викиданных сгенерирована внешним приложением, в данном случае – SQID (https://tools.wmflabs.org/sqid/#/status). Внешние инструменты – программы, которые запускаются не на серверах Викиданных, а на сторонних серверах. Большинство из них полезны для извлечения данных, предоставляемых Викиданными [4–5].

1.6. Примеры использования Викиданных сторонними приложениями. Информация, собранная в Викиданных интересна в своем собственном виде, поэтому для более удобного и эффективного доступа к ней могут быть построены многие внешние приложения. Например, приложения общего просмотра данных, такие как на рис. 5. Здесь страница Иоганн Себастьян Бах автоматически сгенерирована на основе данных, что были извлечены из Викиданных. Или инструменты специального назначения, например, древо жизни, таблицы элементов, а также различные инструменты для отображения.

Приложения могут использовать API Викиданных для просмотра, запроса и даже редактирования данных. Если не достаточно возможностей простых запросов, то в таком случае требуется локальная копия базы Викиданных.

Statistics based on Wikidata dump 31.10.2016

	Items	Properties	Total
Number	24324155	2844	24326999
Statements	120600907	20106	120621013
Labels	125873851	69604	125943455
Descriptions	201000976	30174	201031150
Aliases	12800682	39137	12839819
Site links	53711422	0	53711422

Data Freshness

Statistical data is computed from the data dump about once per week. Basic statisics (class and property names, usage counts for properties, direct instances of classes) are refreshed more frequently, about once per hour. All other data is live.

Dump date	31.10.2016
Property data refresh	07.11.2016, 12:12:06
Class data refresh	07.11.2016, 11:12:16

Рис. 4. Статистика базы данных Викиданные

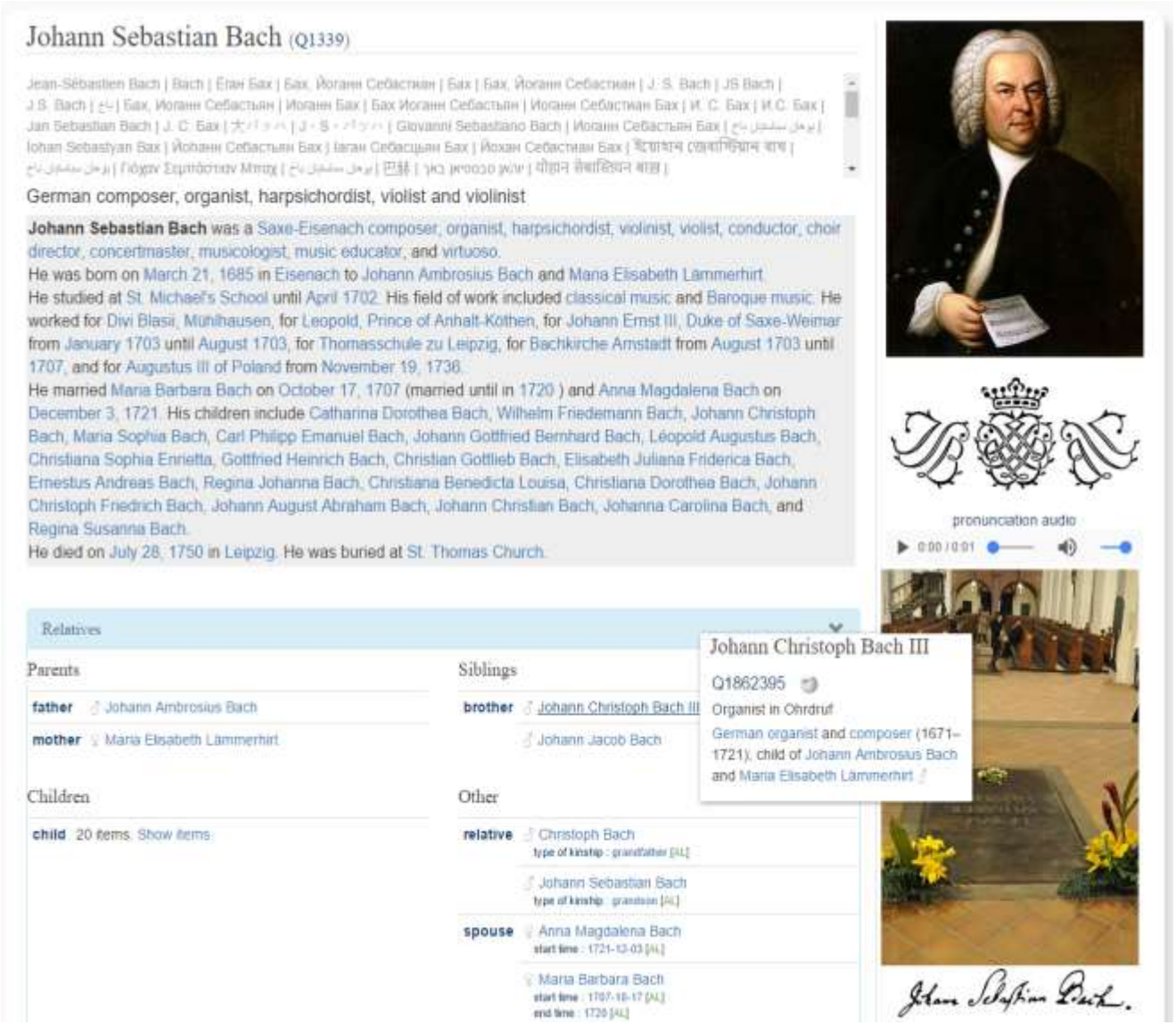

Рис. 5. Викиданные во внешних приложениях: "Резонатор" – программа просмотра данных

2. Сервис запросов к Викиданным

Wikidata Query Service (WDQS) представляет собой пакет программного обеспечения и публичный сервис[8], предназначенный для выполнения SPARQL-запросов, позволяющий запрашивать данные из базы данных Викиданные. Время выполнения каждого запроса ограничено 30 секундами. Это справедливо как для графического интерфейса пользователя (GUI), так и публичной точки доступа SPARQL.

2.1. Набор данных. Сервис запросов к Викиданным работает на множестве данных из wikidata.org, представленных в RDF. Предоставляется возможность скачать еженедельную копию всех данных[9], представленных в Викиданные.

2.2. Семантическая тройка – "Предмет – Предикат – Объект" известна как тройка или как утверждение о данных. Утверждение "Небо имеет голубой цвет", состоит из предмета "небо", предиката "имеет цвет" и объекта "голубой". Тройка также используется в качестве формы основной синтаксической схемы запросов в WDQS. Допускается расширенное использование троек, в том числе с использованием троек в качестве объектов или предметов других троек.

[8] http://query.wikidata.org/

[9] https://dumps.wikimedia.org/wikidatawiki/entities/

2.3. Графический интерфейс пользователя. Домашняя страница GUI позволяет редактировать и передавать SPARQL-запросы механизму выполнения запросов, результаты которых отображаются в виде таблицы HTML. Каждый запрос имеет уникальный URL, который может быть закладкой для последующего использования. Переход к этому URL вносит запрос в окно редактирования, но без его выполнения (для его выполнения нужно нажать кнопку "Выполнить").

Можно также генерировать короткий URL для текущего запроса через сервис укорачивания URL, выбрав опцию *Short URL to result* на ссылке *Link* справа. Также по этой ссылке имеется еще две полезные опции: *SPARQL-endpoint* (точка доступа SPARQL), по которой можно получить результирующий XML-файл текущего запроса и *Embed result* (встроить результат), когда по полученному коду результат текущего запроса можно непосредственно вставлять в вики-разметку вновь создаваемых или редактируемых страниц приложений.

Кнопка "Добавить префиксы" формирует заголовок, содержащий стандартные префиксы для SPARQL-запросов. Полный список полезных префиксов указан в документации о формате RDF[10]. Наиболее распространенные префиксы работают в автоматическом режиме.

GUI также имеет простой механизм более детального анализа сущности, который может быть активирован, нажав на символ  перед сущностью в результирующей таблице (рис. 6). Щелчок на идентификаторе Q-ID сущности приводит к обращению на страницу самой сущности в wikidata.org.

При выполнении запроса WDQS в GUI можно выбрать вид представления его результатов, указав в начале запроса комментарий: #defaultView:viewName. Результаты запросов могут быть представлены в виде таблицы, карты, сетки изображений, временной шкалы, графа, линейной диаграммы, гистограммы, точечной диаграммы.

[10] RDF format documentation

Рис. 6. Просмотр свойств, установленных для сущности с идентификатором Q2906022

2.4. Точка доступа SPARQL (API). SPARQL запросы могут быть переданы непосредственно в точку доступа SPARQL запросом GET к https://query.wikidata.org/sparql ?query=SPARQL (POST и другие методы запросов запрещены). Результат возвращается в виде XML по умолчанию, или как JSON, если установлен либо параметр запроса format=json, либо заголовок Accept: application/sparql-results+json. Формат JSON является стандартным[11]. В настоящее время точкой доступа SPARQL поддерживаются следующие форматы вывода результата запросов: XML, JSON, TSV, CSV, бинарный RDF.

2.5. Автономный сервис. Поскольку представляемый сервис запросов – программа с открытым исходным кодом, его можно запустить на сервере любого пользователя, используя инструкции[12].

2.6. Примеры выполнения SPARQL-запросов в WDQS. В документации[13] приводится множество примеров

[11] SPARQL 1.1 Query Results JSON Format

[12] https://www.mediawiki.org/wiki/Wikidata_query_service/User_Manual#Standalone_service

[13] https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples

SPARQL-запросов со ссылками на их выполнение в представляемом сервисе WDQS.

Мы рассказали, как создавать и отображать запросы к Викиданным в сервисе WDQS. Далее будем встраивать запросы и их результаты в разметку страниц создаваемых энциклопедических статей. Для этого можно воспользоваться одним из расширений программной системы MediaWiki, на основе которой строится наш энциклопедический сайт – ExternalData (https://www.mediawiki.org/wiki/Extension:External_Data).

Расширение ExternalData позволяет использовать и отображать значения, извлеченные из различных источников: внешних URL-адресов, локальных вики-страниц и локальных файлов.

3. Процедура создания страницы-списка

Рассмотрим процедуру создания страницы-списка "Музыканты, родившиеся в Украине" на примере одноименного запроса (http://tinyurl.com/zmz8g27) к Викиданным.

Шаг 1. Формулировка запроса на языке SPARQL (рис. 7).

```
SELECT DISTINCT ?label ?subj ?placeLabel (year(?dateOfBirth) as ?yearOfBirth)
(coalesce(year(?dateOfDeath), '_') as ?yearOfDeath) ?pic  WHERE {
  ?subj wdt:P106 wd:Q639669 .
  ?subj wdt:P19 ?place .
  ?subj wdt:P569 ?dateOfBirth.
  ?subj wdt:P570 ?dateOfDeath.
  ?subj wdt:P18 ?pic.
  ?place wdt:P17 wd:Q212 .
  ?subj rdfs:label ?label filter (lang(?label) = "ru")
 SERVICE wikibase:label {
 bd:serviceParam wikibase:language "ru" .
 }
} ORDER BY ASC(?label)
```

Рис. 7. Запрос "Музыканты, родившиеся в Украине" на языке SPARQL

Шаг 2. Проверка и отладка запроса в сервисе WDQS (рис. 8).

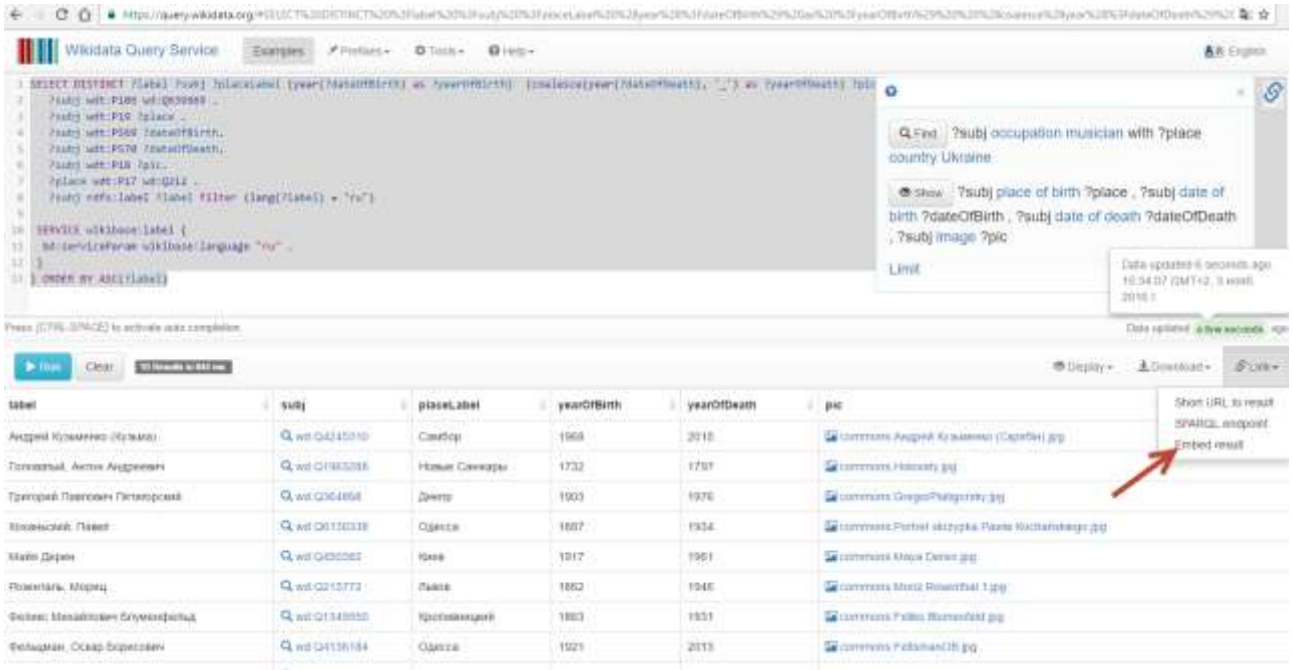

Рис. 8. Выполнение запроса "Музыканты, родившиеся в Украине" в сервисе WDQS

Шаг 3. URL Decoder/Encoder. Далее отлаженный текст запроса нужно закодировать. Существует как минимум два варианта выполнения этой операции. В окне результата WDQS, во всплывающем списке Link выбрать опцию Embed result и в появившемся окне выбираем закодированный код данного запроса. Либо, используя сторонний инструмент, например (http://meyerweb.com/eric/tools/dencoder/).

Шаг 4. Получение результата в виде XML-файла. Выбрав опцию SPARQL-endpoint (точка доступа PARQL) на ссылке Link справа в окне результата запроса сервиса WDQS получаем ответ в отдельном XML-файле. Фрагмент полученного файла показан на рис. 9.

Шаг 5. Задание функции #get_web_data расширения ExternalData. Расписываем составляющие функции #get_web_data для данного примера и вносим в вики-разметку создаваемой страницы "Музыканты, родившиеся в Украине" (рис. 10).

```
<?xml version='1.0' encoding='UTF-8'?>
<sparql xmlns='http://www.w3.org/2005/sparql-results#'>
    <head>
        <variable name='label'/>
        <variable name='subj'/>
        <variable name='placeLabel'/>
        <variable name='yearOfBirth'/>
        <variable name='yearOfDeath'/>
        <variable name='pic'/>
    </head>
    <results>
        <result>
            <binding name='label'>
                <literal xml:lang='ru'>Головатый, Антон Андреевич</literal>
            </binding>
            <binding name='subj'>
                <uri>http://www.wikidata.org/entity/Q1963288</uri>
            </binding>
            <binding name='placeLabel'>
                <literal xml:lang='ru'>Новые Санжары</literal>
            </binding>
            <binding name='yearOfBirth'>
                <literal datatype='http://www.w3.org/2001/XMLSchema#integer'>1732</literal>
            </binding>
            <binding name='yearOfDeath'>
                <literal datatype='http://www.w3.org/2001/XMLSchema#integer'>1797</literal>
            </binding>
            <binding name='pic'>
                <uri>http://commons.wikimedia.org/wiki/Special:FilePath/Holovaty.jpg</uri>
            </binding>
        </result>
```

Рис. 9. Фрагмент результата запроса в виде XML-файла

```
{{#get_web_data:url=https://query.wikidata.org
/sparql?query=SELECT%20DISTINCT%20%3Flabel
%20%3Fsubj%20%3FplaceLabel%20(year(%3Fdat
eOfBirth)%20as%20%3FyearOfBirth)%20%20(coal
esce(year(%3FdateOfDeath)%2C%20%27_%27)%
20as%20%3FyearOfDeath)%20%3Fpic%20%20W
HERE%20%7B%0A%20%20%20%3Fsubj%20wdt
%3AP106%20wd%3AQ639669%20.%0A%20%20
%20%3Fsubj%20wdt%3AP19%20%3Fplace%20.%
0A%20%20%20%3Fsubj%20wdt%3AP569%20%3
FdateOfBirth.%0A%20%20%20%3Fsubj%20wdt%
3AP570%20%3FdateOfDeath.%0A%20%20%20%
3Fsubj%20wdt%3AP18%20%3Fpic.%20%0A%20
%20%20%3Fplace%20wdt%3AP17%20wd%3AQ2
12%20.%0A%20%20%20%3Fsubj%20rdfs%3Alab
el%20%3Flabel%20filter%20(lang(%3Flabel)%20
%3D%20%22ru%22)%0A%20%20%20%20%20%
20%20%20%20%0A%20SERVICE%20wikibase%3
Alabel%20%7B%0A%20%20bd%3AserviceParam
%20wikibase%3Alanguage%20%22ru%22%20.%0
A%20%7D%0A%7D%20ORDER%20BY%20ASC(%
3Flabel)limit%20343&format=xml

|format=xml

|use xpath

|data=name=//binding[@name='label']/literal,item
=//binding[@name='subj']/uri,place=//binding[@n
ame='placeLabel']/literal,yearOfBirth=//binding[@n
ame='yearOfBirth']/literal,yearOfDeath=//binding[
@name='yearOfDeath']/literal,pic=//binding[@nam
e='pic']/uri}}
```

Рис. 10. Пример задания функции #get_web_data в вики-разметке

Шаг 6. Отображение таблицы значений, задание функции #for_external_table расширения ExternalData. Добавляем в вики-разметку создаваемой страницы следующий фрагмент, содержащий описание заглавия таблицы результата и имена внешних переменных (рис. 11).

```
{| class="wikitable"
! Имя
! URI
! Место рождения
! Годы жизни
! Фото
{{#for_external_table:<nowiki/>
{{!}}-
{{!}} {{{name}}}
{{!}} {{{item}}}
{{!}} {{{place}}}
{{!}} {{{yearOfBirth}}}-{{{yearOfDeath}}}
{{!}} <img src={{{pic}}} width=200px>
}}
```

Рис. 11. Пример задания функции #for_external_table в вики-разметке

Сохраняем его и переходим в режим просмотра этой страницы[14]. Все данные таблицы получены из Википедии и сопутствующих проектов. Далее показан внешний вид страницы "Музыканты, родившиеся в Украине", рис. 12.

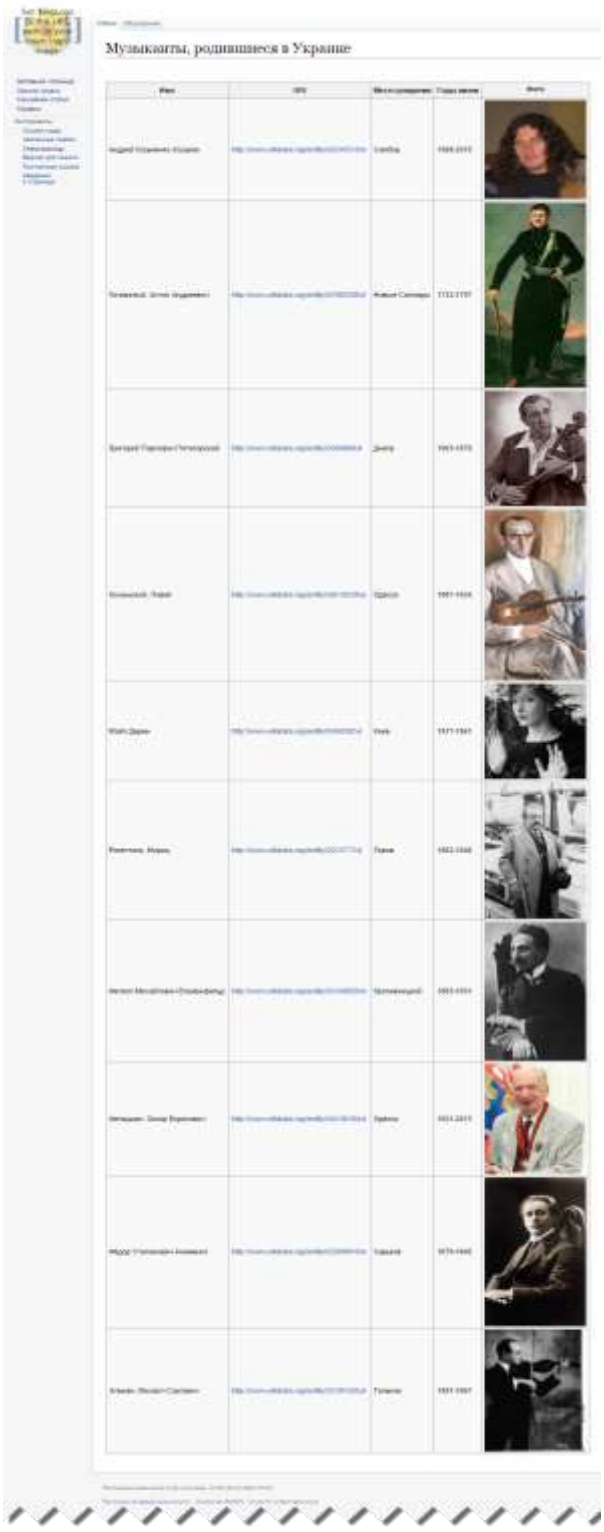

Рис. 12. Просмотр страницы

[14] http://sew.isofts.kiev.ua/index.php/Музыканты,_родившиеся_в_Украине

Аналогичным образом можно создать страницы-списки "Поэты, родившиеся в Украине", "Писатели, родившиеся в Украине" или страницы включающие другие подобные запросы.

Выводы

База Викиданные, её содержание и основное программное обеспечение находятся в стадии постоянного развития, исход которого трудно предвидеть. Учитывая важную роль, которую играет база Викиданные для Википедии, можно быть уверенным в том, что этот проект будет продолжать расти по размеру и качеству. Многие захватывающие возможности использования этих данных еще предстоит исследовать.

1. *Denny Vrandeciˇ, Markus Krötzsch* Wikidata: A Free Collaborative Knowledgebase. In. Proc. CACM-2014 – Communications of the ACM. October 2014. Vol. 57, N 10. P. 78–85. http://korrekt.org/papers/Wikidata-CACM-2014.pdf
2. *Peter Buneman, James Cheney, Wang-Chiew Tan, Stijn Vansummeren*. Curated databases. In Maurizio Lenzerini and Domenico Lembo, editors. In Proc. 27th Symposium on Principles of Database Systems. PODS'09. P. 1–12. http://citeseerx.ist.psu.edu/viewdoc/download;jses sionid=9A9267DB6C3139BA98E3C309E5DFA81F?doi=10.1.1.168.2515&rep=rep1&type=pdf
3. *Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez and Denny Vrandeciˇ* Introducing Wikidata to the Linked Data Web. In Proc. The Semantic Web – ISWC 2014. Lecture Notes in Computer Science. Vol. 8796. P. 50–65. http://korrekt.org/papers/Wikidata-RDF-export-2014.pdf
4. *Lucie-Aimée Kaffee* Generating Article Placeholders from Wikidata for Wikipedia: Increasing Access to Free and Open Knowledge. HTW Berlin University of Applied Sciences. International Media and Computing Faculty IV. A thesis for the degree of Bachelor of Science. March 4, 2016. 62 p. https://upload.wikimedia.org/wikipedia/commons/9/99/Generating_Article_Placeholders_from_Wikidata_for_Wikipedia_-_Increasing_Access_to_Free_and_Open_Knowledge.pdf
5. *Jakob Voß* Classification of Knowledge Organization Systems with Wikidata. In Proc. 15th European Networked Knowledge Organization Systems Workshop. NKOS 2016. Hannover. September 9, 2016. Vol. 1676. P. 15–22. http://ceur-ws.org/Vol-1676/paper2.pdf

References

1. Denny Vrandeciˇ, Markus Krötzsch Wikidata: A Free Collaborative Knowledgebase. In. Proc. CACM-2014 – Communications of the ACM. October 2014. Vol. 57, N 10. P. 78–85. http://korrekt.org/papers/Wikidata-CACM-2014.pdf
2. Peter Buneman, James Cheney, Wang-Chiew Tan, Stijn Vansummeren. Curated databases. In Maurizio Lenzerini and Domenico Lembo, editors. In Proc. 27th Symposium on Principles of Database Systems. PODS'09. P. 1–12. http://citeseerx.ist.psu.edu/viewdoc/download;jses sionid=9A9267DB6C3139BA98E3C309E5DFA81F?doi=10.1.1.168.2515&rep=rep1&type=pdf
3. Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez and Denny Vrandeciˇ Introducing Wikidata to the Linked Data Web. In Proc. The Semantic Web – ISWC 2014. Lecture Notes in Computer Science. Vol. 8796. P. 50–65. http://korrekt.org/papers/Wikidata-RDF-export-2014.pdf
4. Lucie-Aimée Kaffee Generating Article Placeholders from Wikidata for Wikipedia: Increasing Access to Free and Open Knowledge. HTW Berlin University of Applied Sciences. International Media and Computing Faculty IV. A thesis for the degree of Bachelor of Science. March 4, 2016. 62 p. https://upload.wikimedia.org/wikipedia/commons/9/99/Generating_Article_Placeholders_from_Wikidata_for_Wikipedia_-_Increasing_Access_to_Free_and_Open_Knowledge.pdf
5. Jakob Voß Classification of Knowledge Organization Systems with Wikidata. In Proc.

15th European Networked Knowledge Organization Systems Workshop. NKOS 2016. Hannover. September 9, 2016. Vol. 1676. P. 15–22. http://ceur-ws.org/Vol-1676/paper2.pdf

Получено 10.01.2017

Об авторах:

Проскудина Галина Юрьевна,
научный сотрудник,
Количество научных публикаций в украинских изданиях – 28.
Количество научных публикаций в зарубежных изданиях – 15.
http://orcid.org/0000-0001-9094-1565.

Кудим Кузьма Алексеевич,
младший научный сотрудник,
Количество научных публикаций в украинских изданиях – 12.
Количество научных публикаций в зарубежных изданиях – 7.
http://orcid.org/0000-0001-9483-5495.

Место работы авторов:

Институт программных систем
НАН Украины,
03187, Киев-187,
проспект Академика Глушкова, 40.
Тел.: +38(044)526 6033.

E-mail: gupros@isofts.kiev.ua,
kuzma@isofts.kiev.ua