ISSN 2413–9858

УДК 004.822:908

СЕРВИС ЭКСПОРТА И ОБНОВЛЕНИЯ СВЕДЕНИЙ ОБ ОБЪЕКТАХ КУЛЬТУРНОГО НАСЛЕДИЯ КАК ЭЛЕМЕНТ СЕМАНТИЧЕСКОЙ ПАУТИНЫ

А.В. Соловьев

CULTURAL HERITAGE EXPORT AND UPDATE BOT AS AN ELEMENT OF SEMANTIC WEB

A.V. Soloviev

Аннотация. В статье описана реализация на волонтёрской площадке Викигида комплекса программных средств для экспорта и обновления сведений об объектах культурного наследия (ОКН) России в базу данных Викиданные. Описанные программные средства позволяют решить проблему согласованности и непротиворечивости сведений об ОКН. Описан алгоритм преобразования сведений об ОКН в семантические аннотации. Таким образом, работа является этапом интеграции этих сведений в семантическую паутину проектов Викимедиа.
Ключевые слова: *вики; семантическая паутина; объект культурного наследия.*

Abstract. The article describes the implementation of a set of software tools for exporting and updating information about cultural heritage monuments of Russia to the Wikidata database. The described software tools are based on the Wikivoyage volunteer platform and allow to keep cultural heritage data in consistent and reliable state. An algorithm for converting cultural heritage information into semantic annotations is described. Thus, this work is the first stage of integrating this information into the semantic web of Wikimedia projects.
Key words: *wiki; semantic web; cultural heritage.*

Введение

В 1994 году Уордом Каннингеном была предложена концепция «вики» – класса систем управления контентом сайтов, в которых предусмотрено изменение содержимого самими посетителями сайта. При этом на стороне редакторов не используются какие-либо специальные средства, кроме самого браузера. Для оформления страниц используется особый язык разметки (вики-разметка), отличающийся простотой и лёгкостью форматирования структурных элементов и гиперссылок в содержимом страниц. Такие системы управления контентом предусматривают появление изменений сразу после их внесения и в то же время ведут детальный учёт всех изменений, что даёт возможность сравнения редакций и восстановления ранних версий страницы. Как правило, проекты на основе вики-движков поддерживаются достаточным количеством активных редакторов, что позволяет успешно поддерживать содержимое этих сайтов в целостном и непротиворечивом виде, несмотря на отдельные случаи вандализма со стороны некоторых пользователей, поскольку механизм вики даёт возможность легко откатить страницу к любой ранней версии, то есть реализует принцип «исправлять легче, чем портить» [1].

В настоящий момент существует более 30 систем управления контентом такого типа. Среди наиболее распространённых можно упомянуть MediaWiki, DocuWiki, Foswiki, DrupalWiki, TWiki и другие [2]. В частности, одним из наиболее популярных проектов на вики-движке MediaWiki [3] является Википедия и родственные ей проекты фонда Викимедиа (Викисклад, Викиновости, Викигид, Викиданные и другие). В данной работе речь идёт о проекте Викигид – открытом мультиязычном вики-проекте по созданию свободных туристических путеводителей и описаний достопримечательностей. Один из подпроектов

Викигида – списки объектов культурного наследия (ОКН) России [4]. В настоящий момент это крупнейший в России каталог объектов, признанных ОКН, выявленных ОКН или обладающих признаками ОКН. Каталог ОКН Викигида постоянно пополняется благодаря энтузиазму местных краеведов, обнаруживающих ценные объекты задолго до того, как они будут поставлены на официальную охрану, а вики-движок MediaWiki является естественным инструментом для наполнения и редактирования этого каталога.

Для группирования тематически объединённых страниц в MediaWiki используется механизм категоризации: страницам проекта может быть назначен один или несколько атрибутов «категория» с определённым значением. В частности, страницы каталога ОКН объединены в категорию «Списки культурного наследия России». Сам каталог разбит на страницы по географическому принципу: сначала по субъектам федерации, потом по районам субъекта и далее, при необходимости, по населённым пунктам. Если какой-то населённый пункт особенно богат на ОКН, то ему может быть посвящено несколько страниц списков.

Унифицированное оформление описаний ОКН достигается путём применения вики-шаблонов. Вики-шаблон – это специальная страница, которая содержит параметризированное описание вики-разметки определённого контента. Со страниц в основном пространстве сайта можно обратиться к шаблону, передав ему необходимые параметры, тогда вики-движок вставит в это место страницы содержимое страницы-шаблона с подставленными параметрами. В терминах языков программирования о вики-шаблоне можно говорить как об описанной пользователем функции, которой передаются параметры и которая возвращает какой-то текст в результате обработки этих параметров. В движке MediaWiki предусмотрено использование языка описания сценариев Lua для манипуляции кодом вики-разметки в шаблонах и на основных страницах сайта [3]. Описание ОКН в Викигиде оформляется при помощи шаблона {{monument}}. Первоначально код шаблона был целиком написан на языке вики-разметки, но в 2018 году активные редакторы Викигида переписали его на Lua, что позволило сократить время генерации страниц списков.

Такой подход в организации списков ОКН несёт в себе ряд присущих традиционным вики недостатков:

- проблема обеспечения согласованности содержимого (не гарантируется отсутствие дублирующих элементов, повторяющихся идентификаторов, противоречий в описаниях ОКН);
- сложность формирования поисковых запросов (например, найти все ОКН заданного архитектора или определённого периода постройки) и др.

Использование вики-шаблона позволяет немного структурировать хранимую информацию об ОКН, что облегчает формализацию машинной обработки списков в части повторного использования знаний вики.

Для решения такого рода проблем вводится концепция семантических вики, то есть таких систем управления контентом, которые хранят структурированные данные либо отдельно, либо прямо в тексте вики-разметки. Это позволяет реализовать семантические аннотации (*утверждения*, или *суждения*), когда связь между объектами статей не просто декларируется в виде ссылки, но и указывается характер этой связи [5]. Например: «Михайловский замок – *построен по проекту* – В. И. Баженова». Субъект «Михайловский замок» связан с объектом «В. И. Баженов» при помощи аннотации (предиката) «построен по проекту». Сами предикаты также могут выступать субъектом и объектом какого-либо утверждения: «Построен по проекту – *обратное свойство* – является автором (проектировщиком)». Такого рода аннотации дают возможность изменять представление содержимого страницы в зависимости от контекста и осуществлять семантическую навигацию. Например, построить иерархию объектов по аннотируемому признаку (*все архитекторы*) и перемещаться по ней (*постройки выбранного архитектора*). Это также приводит к возможности формирования многокритериальных поисковых запросов на

формальном языке, так называемому семантическому поиску. Структурированные данные в семантических вики формируют некоторую систему фактов. Благодаря наличию аннотаций из этой системы фактов можно извлекать неявные (скрытые) данные, то есть на основе заранее заданных логических формализмов строить логические выводы. Например, из приведённых суждений:

- «Михайловский замок – *построен по проекту* – В. И. Баженова» и
- «Построен по проекту – *обратное свойство* – является автором (проектировщиком)», – можно сделать логический вывод:
- «В. И. Баженов – *является автором (проектировщиком)* – Михайловского замка».

В качестве стандарта модели представления данных обычно используется RDF (Resource Description Framework – среда описания ресурса) [6]. Эта концепция предусматривает представление данных в виде *утверждений о ресурсах*, имеющих вид «субъект – *предикат* – объект», что в определённом смысле соответствует понятию суждения в математической логике. Для хранения данных по модели RDF существует несколько распространённых форматов представления на основе JSON, XML и др.

Для обработки RDF-данных используются различные языки запросов, например SPARQL (SPARQL Protocol and RDF Query Language) [7]. Использование формального языка для запроса к базе знаний позволяет получать ответ в машиночитаемом виде.

Семантические вики, поддерживающие хранение утверждений в виде RDF и программный интерфейс для выполнения SPARQL-запросов, обычно называются семантической паутиной.

Фонд Викимедиа планировал развивать свою вики-энциклопедию как семантическую вики. Для этого движок MediaWiki был дополнен расширением, позволяющим вводить и обрабатывать структурированную информацию, – Sematic MediaWiki. Однако вместо этого был запущен проект Викиданные как централизованное хранилище структурированных данных для всех остальных проектов Викимедиа [8]. Предметы статей в вики-проектах привязаны к *сущностям* Викиданных, а для каждой сущности хранится набор RDF-совместимых *утверждений*. Каждая сущность идентифицирует некоторый субъект. Утверждения, касающиеся этого субъекта, содержат сопоставления свойств (предикатов) и объектов, формируя, таким образом, суждения. Для уточнения значения утверждений и предоставления дополнительной информации о контексте данного утверждения могут использоваться *квалификаторы*, имеющие также форму утверждения (но субъектом в этом случае является не сущность в целом, а значение квалифицируемого утверждения). Например, свойство (предикат) «координаты» может иметь единственное значение для «точечного» объекта, а для протяжённого объекта – несколько значений, как минимум, координаты с квалификатором «применимо к части – начало» и координаты с квалификатором «применимо к части – конец». Также утверждения могут снабжаться ссылками на контент, подтверждающий утверждение. Викиданные имеют точку входа для SPARQL-запросов. Таким образом, любой вики-проект Викимедиа обретает черты семантической паутины.

В качестве инфраструктуры для технической реализации задач, связанных с автоматическим или автоматизированным анализом, синтезом или иной обработкой данных в вики-проектах, фонд Викимедиа поддерживает WCS (Wikimedia Could Services – облачные службы Викимедиа). В наиболее общем виде WCS является IaaS (инфраструктурой как услуга) и предоставляет облачные виртуальные машины Cloud VPS (cloud virtual private server), реализованные на основе платформы виртуализации OpenStack. Один из проектов, работающий в рамках концепции Cloud VPS, – это Toolforge. Toolforge представляет собой разновидность виртуального хостинга (то есть PaaS – платформу как услуга). Toolforge реализует дополнительные веб-серверы, СУБД и другие системы хранения, а также распределённую систему исполнения заданий, реализованную на Kubernets (ранее для этой цели использовался Oracle Grid Engine) [9].

Для неинтерактивного взаимодействия с вики движок MediaWiki предоставляет программный интерфейс Action API типа RESTful веб-сервиса [3]. С помощью этого интерфейса можно выполнять такие действия как аутентификация в вики, создание и редактирование страниц, их синтаксический разбор и извлечение различных метаданных.

Задача интеграции сведений об ОКН в семантическую паутину проектов Викимедиа представляется актуальной. Ранее сведения об ОКН хранились в Викигиде в виде текстов шаблонов. В терминах семантической вики объекты культурного наследия из списков Викигида должны быть представлены сущностями Викиданных. В ходе данной работы была предложена схема систематического экспорта сведений об ОКН в реляционную базу данных и далее их конвертирование в сущности Викиданных с возможностью периодического обновления. Для реализации этой схемы необходимо разработать ряд программных средств, позволяющих структурировать первичные текстовые данные для помещения их в реляционную базу данных, производить различные проверки целостности и непротиворечивости данных и в результате строить утверждения (семантические аннотации) для Викиданных.

Практическая часть

Для реализации поставленной задачи предлагается следующий алгоритм. Разрабатываемое программное средство при помощи MediaWiki Action API запрашивает список страниц в категории «Списки культурного наследия России». Для каждой страницы запрашивается номер ревизии. Если номер ревизии изменился, текстовый список объектов на данной странице подвергается синтаксическому разбору и помещается в реляционную базу данных. На основе этой структурированной информации формируется система фактов, которая экспортируется в виде утверждений для соответствующих сущностей Викиданных. Описываемое программное средство разрабатывается на языке PHP и предусматривает периодический запуск на хостинге Toolforge при помощи Kubernets.

Страницы из категории «Списки культурного наследия России» содержат данные об объектах культурного наследия (ОКН) в виде вики-шаблона {{monument}}. Параметры этого шаблона перечислены в таблице 1.

Таблица 1 – Параметры шаблона {{monument}}

Обозначение	***Пояснение***
1	***2***
name	Наименование объекта.
type	Тип памятника: памятник архитектуры (architecture), памятник истории (history), археологический памятник (archeology) или памятник монументального искусства (monument).
lat и long	Географические координаты: широта и долгота в градусах. Северным широтам и восточной долготе соответствуют положительные значения, южным широтам и западной долготе – отрицательные.
precise	Флаг точности координат – эвристический параметр, обозначающий субъективную точность или приблизительность географических координат.
knid	10-значный код объекта, использующийся для идентификации ОКН в списках Викигида. Первоначальный смысл такого идентификатора – номер из реестра информационно-вычислительного центра Министерства культуры РФ kulturnoe-nasledie.ru, который существовал до 2015 года.
complex	Код комплекса – 10-значный код объекта, описывающего как целое ансамбль памятников, к которому принадлежит данный объект.
knid-new	Код ЕГРОКН – 15-значный код в Едином государственном реестре объектов культурного наследия (с 2015 года).
wdid	Сущность Викиданных для данного ОКН.
region	ISO-код региона (субъекта РФ).

Продолжение таблицы 1

1	*2*
district	Название района.
municipality	Название города, посёлка, села или деревни.
munid	Сущность Викиданных, соответствующая городу, посёлку, селу или деревне.
address	Адрес.
year	Год постройки (возникновения).
author	Автор (архитектор, скульптор, инженер, …).
status	Флаг утраты (status=destroyed) – если памятник утрачен, разрушен и т.п.
dismissed	Флаг снятия с охраны задаётся в виде кода документа о снятии с охраны.
protection	Категория охраны: Ф – федеральная, Р – региональная, М – местная, В – выявленный объект, Н – не охраняется (обладает признаками ОКН).
style	Архитектурный стиль (для памятников архитектуры): конструктивизм, авангард, модерн.
image	Изображение, иллюстрирующее ОКН – имя файла на Викискладе.
wiki	Название статьи в Русской Википедии
commons	Название категории Викисклада.
sobory	Идентификатор объекта в Народном каталоге православной архитектуры sobory.ru.
temples	Идентификатор объекта в проекте «Храмы России» temples.ru.
link	Ссылка на внешний ресурс с описанием.
linkextra	Ссылка на дополнительный внешний ресурс с описанием.
description	Дополнительные сведения в произвольной форме.
document и document2	Код документа (документов) о постановке на охрану.

Параметры *document*, *document2* и *dismissed* при визуализации шаблона {{monument}} отображаются как сноски. Для унификации содержания этих сносок в Викигиде обозначения документов о постановке на охрану (или снятии с охраны) хранятся в пространстве шаблонов отдельно для каждого региона на страницах "Шаблон:Monument-documents/xxx", где xxx – ISO-код региона. На страницах списков подключается шаблон {{monument-documents}} с параметром – кодом региона, который генерирует подходящий раздел примечаний с расшифровкой использованных обозначений.

Кроме того, и в «старом» реестре с 10-значными номерами, и в ЕГРОКН, есть дублирующиеся записи. Для их учёта создан ряд страниц списков с шаблонами {{monument-duplicate}} и {{monument-duplicate-egrokn}}. В параметрах этих шаблонов указываются номер-дубликат (*knid* или *knid-new*), номер, использованный в Викигиде, (*knid-list* или *knid-new-list*) и краткое описание объекта (*name*, *district*, *municipality*, *address*).

На языке PHP разработан сценарий, который с заданной периодичностью (1 раз в день) выполняет синтаксический разбор страниц списков и помещает первичные данные во временное хранилище – реляционную базу данных MariaDB.

На первом этапе сценарий через Action API запрашивает у движка список страниц, входящих в категорию «Списки культурного наследия России»:

$BASE_URL?action=query&format=json&list=categorymembers&cmlimit=10&cmtitle=*$CATEGORY*

Движок возвращает фрагменты списка порциями по *cmlimit* элементов. Для перебора всего списка, этот запрос повторяется с возвращённым сервером параметром *cmcontinue*. Для каждой страницы возвращается численный идентификатор (*pageid*) и название (*title*). Кроме того, для каждой страницы запрашивается её содержимое в виде вики-разметки (*wikitext*), идентификатор ревизии (*revid*) и свойства (*properties*) (среди которых есть идентификатор сущности Викиданных – *wikibase_item*):

$BASE_URL?action=parse&format=json&pageid=$PAGEID&prop=wikitext|revid|properties

ISSN 2413–9858

Полученная информация сохраняется в таблице *pages* со следующими полями:

- *pageid* (целое число, ключ) – идентификатор страницы в вики;
- *title* (строка) – название страницы в вики;
- *revid* (целое число) – номер последней ревизии страницы;
- *wdid* (целое число) – идентификатор сущности страницы (без префикса „Q“);
- а также два булевых флага *dirty* и *blacklisted*, описанных ниже.

При необходимости строка в этой таблице обновляется (например, сущность *wdid* для страницы ранее не была задана, а при новом обходе обнаружена). При обнаружении новой страницы информация о ней в таблицу добавляется. Если возвращённая движком ревизия страницы не совпадает с сохранённой в базе данных, значит, на странице произошли изменения и она подлежит синтаксическому разбору.

Два дополнительных поля: *dirty* и *blacklisted*, – необходимы для управления вторым модулем описываемого программного средства. Флаг *dirty* помечает страницу, которая должна быть обработана в фазе обновления сущностей Викиданных вне зависимости от того, менялась ли страница или нет. Флаг *blacklisted* помечает страницу, которая не подлежит обработке в фазе обновления сущностей Викиданных, даже если она была недавно изменена. Эти флаги выставляются внешними средствами.

Для выполнения синтаксического разбора используются регулярные выражения. Со страницы выделяются фрагменты текста с шаблонами {{monument}}, {{monument-duplicate}} и {{monument-duplicate-egrokn}}. Каждый фрагмент формирует объект, свойства которого соответствуют параметрам шаблона. Для каждого типа шаблонов создана своя таблица в базе данных. В базе данных объекты хранятся в виде JSON-строк, при этом важные для индексации поля дублируются в отдельных колонках. Перед занесением в таблицы информации с обновлённой страницы из базы данных удаляются старые записи, ассоциированные с данной страницей (по полю *pageid*).

Экземпляры объектов на основе шаблона {{monument}} сохраняются в таблице *monuments* со следующими полями:

- *knid* (целое число, ключ) – 10-значный идентификатор ОКН;
- *pageid* (целое число) – идентификатор страницы, с которой взято описание ОКН;
- *json* (текст) – представление объекта в JSON-формате;
- *knid_new* (строка) – идентификатор ЕГРОКН (обычно 15-значное число, но встречаются и специальные значения);
- *complex* (целое число) – 10-значный идентификатор головного элемента комплекса (если применимо);
- *wdid* (целое число) – идентификатор сущности объекта (без префикса „Q“).

Экземпляры объектов на основе шаблона {{monument-duplicates}} сохраняются в таблице *duplicates*, а экземпляры объектов на основе шаблона {{monument-duplicate-egrokn}} сохраняются в таблице *dup_egrokn*. Обе таблицы имеют сходный формат: *knid* (или *knid_new*), *knid_list* (или *knid_new_list*) – дублирующий идентификатор объекта и идентификатор, использованный в списках; *json* – представление записи о дубликате в JSON-формате; *pageid* — идентификатор вики-страницы, с которой взято описание дубликата.

Затем сценарий обращается к странице:

https://ru.wikivoyage.org/wiki/Special:PrefixIndex/Template:Monument-documents/,

чтобы получить список всех страниц-шаблонов с перечнями документов о постановке на охрану. Одна такая страница описывает документы одного региона. Элементы перечня состоят из идентификатора документа, который используется в качестве значения параметра *document* или *document2* в шаблоне {{monument}}, и вики-текста соответствующей сноски. Обычно в вики-тексте присутствует ссылка на загруженный файл либо внешний ресурс.

ISSN 2413–9858

Документы сами по себе должны быть описаны как сущности Викиданных. Для некоторых документов такие сущности созданы, о чём имеется отметка на странице описания загруженного файла. Однако для многих документов (тем более на внешних ресурсах) сущностей нет. А некоторые документы, кроме того, переведены в текстовый вид и размещены в Викитеке – свободной библиотеке фонда Викимедиа. Пока что задача создания сущностей для документов не решена и остаётся актуальной.

В результате обработки этих шаблонов формируется таблица *documents* со следующими полями:

- *id* (строка, ключ) – идентификатор документа;
- *region* (строка) – ISO-код региона, к которому относится документ (пусто – для федеральных);
- *text* (текст) – наименование и описание документа;
- *wdid* (целое число) – идентификатор сущности документа (без префикса „Q“).

На хостинге Toolforge реализован веб-сервис, который на основе данных в этих таблицах формирует страницу-карточку, описывающую объект с заданным 10-значным номером.

Второй этап обработки – это перевод табличных данных в утверждения. Этот этап оформлен в виде независимого программного модуля, который может запускаться отдельно. В качестве параметров командной строки, этот сценарий принимает идентификаторы страниц (*pageid*) для обработки. При выполнении первого этапа формируется список таких страниц. Это страницы с обновлённой ревизией (кроме тех, у которых выставлен флаг *blacklisted*) и страницы с выставленным флагом *dirty*.

Сценарий обновления сущностей Викиданных не обращается непосредственно к страницам Викигида, а извлекает уже структурированную информацию из описанных таблиц реляционной базы данных.

Новая сущность Викиданных для ОКН создаётся при выполнении следующих условий: для объекта заданы координаты и объект не отмечен как снятый с охраны или утраченный. Эти условия продиктованы следующими соображениями. В первую очередь востребована информация о реально существующих объектах, которые не утратили свою культурную ценность. С другой стороны, отсутствие координат зачастую означает неполноту и возможную противоречивость сведений об объекте. Исключением из этого правила являются объекты, описывающие ансамбль (комплекс) как целое, если для них задана категория Викисклада. Такие сущности необходимы для правильной привязки категории Викисклада и для выстраивания правильной иерархии по предикату P361 (является частью).

Для вновь создаваемой сущности или сущности, которая ранее уже была привязана к ОКН, из базы данных экспортируется следующая информация.

Вновь создаваемой сущности задаётся метка на русском языке, соответствующая имени ОКН (*name*). Также в метку в круглых скобках включается название населённого пункта, чтобы снизить эффект неоднозначных наименований. Если ОКН сопоставлена категория Викисклада, то имя этой категории используется как метка на английском языке (поскольку именно такой подход принят в именовании категорий Викисклада). Для уже существующих сущностей метки не изменяются. Для сущности создаются или обновляются утверждения с предикатами, перечисленными в таблице 2.

Кроме того, сущность может быть привязана к страницам в разных проектах Викимедиа. В списках Вигигида указываются ссылки *commonscat* (категория Викисклада) и *wiki* (страница в Русской Википедии). В сущности Викиданных такие ссылки обозначаются commonswiki и ruwiki и если они не заданы, они обновляются по сведениям из Викигида. Если же такие ссылки есть и их значение отличается, то изменение не производится, но в журнале программы об этом факте формируется сообщение. Это может означать, что искомая сущность уже существует, но она не указана в списках Викигида. Потенциально

такая ситуация может означать наличие дублирующей сущности и требует ручного вмешательства.

Таблица 2 – Формируемые утверждения для сущностей Викиданных

Предикат	***Название***	***Способ формирования утверждения***
1	***2***	***3***
P17	Страна	Должно указывать на сущность Q159 (Россия). Если утверждение P17 с таким объектом отсутствует, оно будет добавлено. В утверждении будет дополнительно указан квалификатор P580 (дата начала). Для ОКН с номерами на 82xxx и 92xxx дата начала действия утверждения будет установлена на 18.03.2014, для остальных объектов – на 25.12.1991.
P18	Изображение	Добавляется, если у сущности его до сих пор нет, а в списке Викигида изображение задано.
P1483	kulturnoe-nasledie.ru ID	Должно указывать на единственное значение, совпадающее с *knid* ОКН (10-значный код ОКН). Если значение отличается или отсутствует, оно будет создано или обновлено.
P2817	Перечислен в списке ОКН	Должно указывать на единственное значение – сущность Викиданных исходной страницы Викигида. Если значение отличается или отсутствует, оно будет создано или обновлено.
P2186	Wiki Loves Monuments ID	Должно указывать на значение с кодом памятника на конкурсе «Вики любит памятники». Для российских ОКН в качестве такого кода используется значение *knid* с префиксом "RU". Если значение отличается или отсутствует, то оно будет создано или обновлено.
P361	Является частью	Должно ссылаться на сущность, описывающую весь ансамбль. Если ОКН является частью какого-то ансамбля и такое утверждение отсутствует, оно будет добавлено. Если утверждение уже есть, оно не меняется. Запрет на обновление обусловлен тем, что сущности могут быть вручную выстроены в более сложную многоуровневую иерархию, чем это предусмотрено концепцией ОКН.
P5381	Регистрационный номер ЕГРОКН	Должны указывать на единственное значение, совпадающее с таким значением в списках Викигида (*knid-new*, *sobory*, *temples*). Если значение отличается или отсутствует, оно будет создано или обновлено.
P8316	Идентификатор в каталоге sobory.ru	
P9343	Идентификатор temples.ru	
P31	Это частный случай понятия	Должно ссылаться на сущность, соответствующую типологии ОКН: для памятника градостроительства и архитектуры – Q2319498 (достопримечательность, памятник архитектуры), для памятника истории – Q1081138 (историческое место), для произведения монументального искусства – Q4989906 (монумент), для археологического памятника – Q839954 (археологический памятник), для исторического поселения – Q3920245 (исторический город в России). Если такое значение отсутствует, оно будет добавлено к существующим. У некоторых существующих сущностей было замечено утверждение P31 со ссылкой на Q8346700 (ОКН России), что неверно, поскольку эта сущность предназначена для утверждений с предикатом P1435, поэтому утверждения "– P31 – Q8346700" будут удалены при обнаружении. Кроме того, исходя из названия ОКН может быть предложено дополнительное утверждение с предикатом P31. Если название содержит слово «церковь» – то Q16970 (христианский храм); если слово «часовня» – то Q1975485 (православная часовня); если слово «здание» или «дом» – то Q41176 (здание); если слово «братская могила» – то Q734271 (братская могила) и Q5003624 (мемориал).

Продолжение таблицы 2

1	*2*	*3*
P1435	Статус культурного наследия	Должно ссылаться на сущность, соответствующую категории охраны: Q105835774 (выявленный ОКН), Q23668083 (ОКН федерального значения), Q105835744 (ОКН регионального значения), Q105835766 (ОКН местного значения), Q105835782 (неохраняемый объект с признаками ОКН). Категория охраны может быть явно не указана в списках Викигида, в таком случае объектом этого утверждения будет сущность Q8346700 (ОКН России). Утверждение P1435 с одним из перечисленных объектов должно быть единственным. Если такое утверждение отличается или отсутствует, оно будет создано или обновлено. Утверждения P1435 с другими объектами не изменяются (например, другим объектом утверждения может быть Q43113623 (часть наследия ЮНЕСКО) и т.п.).
P131	Находится в административно-территориальном образовании (АТО)	Должно ссылаться на сущность *munid*, указанную в списке Викигида. Если значение отличается или отсутствует, оно будет создано или обновлено. Игнорируется (не затрагивается) для исторических поселений (когда в списках Викигида *wdid* совпадает с *munid*), поскольку должно ссылаться на АТО более высокого уровня иерархии. Существующие утверждения с квалификаторами также не затрагиваются, так как обычно это означает протяжённое сооружение, которое находится в нескольких АТО, например, мост или канал.
P625	Координаты	Должно иметь единственное значение, отличающееся от координат в списке Викигида менее, чем на 30 метров (для расчёта расстояния между парой GPS-координат используется модель WGS84), иначе оно будет создано или обновлено. Если у сущности есть несколько утверждений P625, они не проверяются и не затрагиваются, так как обычно это означает протяжённое сооружение.
P2795	Указание о расположении	В структурированном виде адрес ОКН может указываться в утверждениях P6375 (почтовый адрес) или P669 (расположен на улице…). Однако тривиального способа преобразования имеющихся в списках Викигида данных в такую форму не существует. Поэтому, если утверждения с предикатом P6375 или P669 у изменяемой сущности отсутствуют, будет создано или обновлено утверждение с предикатом P2795, в котором в неструктурированном виде приводится содержимое поля *address*.
P571	Дата создания или возникновения	Если в списках Викигида заполнено поле *year* и это значение может быть распознано как год (4-значное число), то при отсутствии у изменяемой сущности утверждения P571 оно будет создано с таким значением. Если утверждение P571 уже существует, оно не затрагивается.
P576	Дата прекращения существования	Если в списках Викигида у ОКН стоит флаг status=destroyed (объект утрачен), то при отсутствии у изменяемой сущности утверждения P576 оно будет создано. В качестве объекта предиката будет использоваться специальное значение «неизвестно». Если утверждение P576 уже существует, оно не затрагивается.
P373	Категория Викисклада	Если в списках Викигида заполнено поле *commonscat*, то при отсутствии у изменяемой сущности утверждения P373 оно будет создано. Если утверждение P373 уже существует, оно не затрагивается.

Для перечисленных утверждений в качестве подтверждения даётся ссылка на соответствующую страницу-список в Викигиде, кроме утверждений:

- P1435 – в качестве подтверждения даётся документ о постановке на охрану;
- P2817 – даётся без подтверждения (поскольку само утверждение является ссылкой на страницу-список);
- P5381, P8316, P9343 – в качестве подтверждения даётся ссылка на соответствующую сущность: ЕГРОКН (Q7382189), каталог sobory.ru (Q105343287), каталог temples.ru (Q105956103);
- P2186 – в качестве подтверждения даётся ссылка на страницу-карточку объекта на Toolforge.

Корректная работа программы экспорта и обновления сведений об ОКН в сущности Викиданных возможна при условии отсутствия противоречий в исходных данных. Из-за того, что исходные данные основаны на вручную введённых сведениях, гарантировать отсутствие противоречий невозможно.

Написан ряд программных модулей, которые выполняют различного рода проверки. На основе таких проверок составляются «черные списки» сущностей Викиданных и номеров ОКН, которые не подлежат автоматическому обновлению. Кроме того, для целых страниц списков предусмотрен флаг *blacklisted* в таблице *pages*, которым помечается страница, не подлежащая обработке в фазе обновления сущностей Викиданных, даже если она была недавно изменена. Обычно это означает, что странице не присвоена сущность Викиданных (то есть на неё нельзя сослаться в утверждениях) либо на странице не использован шаблон {{monument-documents}}, а значит утверждения P1435 останутся без подтверждения.

Занесение в «чёрный список» отдельных ОКН обычно обусловлено следующими причинами:

- у разных ОКН указан одинаковый идентификатор сущности (*wdid*);
- в утверждениях имеющейся сущности имеются серьёзные противоречия с данными в списке Викигида: указаны различающиеся 10-значные (*knid*) или 15-значные (ЕГРОКН) идентификаторы, координаты объекта в списке и в утверждении P625 отличаются более чем на 100 м.

Если при очередной проверке противоречий обнаруживается, что причина занесения в «черный список» пропала, страница Викигида, на которой упомянут этот объект, в таблице *pages* помечается флагом *dirty*, чтобы она была обработана в фазе обновления сущностей Викиданных вне зависимости от того, менялась ли страница или нет.

Заключение

В результате выполнения работы создан ряд программных средств на языке PHP для платформы Toolforge:

- Модуль синтаксического разбора вики-страниц списков ОКН. Заносимые редакторами-волонтёрами Викигида в контент проекта в виде параметров вики-шаблонов текстовые данных об объектах культурного наследия подвергаются синтаксическому разбору и помещаются в реляционную базу данных.
- Модуль создания сущностей Викиданных и обновления утверждений у существующих сущностей. На основе структурированных данных из реляционной базы данных строятся утверждения (семантические аннотации) для Викиданных в соответствии с описанной схемой.
- Группа сценариев проверки целостности и непротиворечивости информации в списках ОКН. Производятся различные проверки целостности и непротиворечивости данных: уникальность идентификаторов (отсутствие дубликатов), заполненность обязательных полей, отсутствие недействительных ссылок на другие вики-проекты, отсутствие противоречий в утверждениях существующих сущностей и т. п.).

По состоянию на 1 июня 2022 года списки Викигида содержат информацию о 205745 объектах, из них 43614 – памятники археологии. Памятники археологии выделены отдельно, поскольку информация о них официальными органами в публичный доступ не

выкладывается, а значит, в списках Викигида большей частью они будут представлены лишь отрывочными сведениями. Имеется информация о координатах 69846 объектов (34% от общего числа), или в статистике без археологии – 68661 объект (43%).

Разработанный сервис экспорта и обновления сведений об ОКН в сущности Викиданных с марта 2021 года запущен в тестовом режиме. Из 2727 страниц списков для обработки разрешены 485 страниц (флаг *blacklisted*=0). При этом в «черный список» занесено 1249 сущностей Викиданных или 1198 ОКН. До начала работы сервиса было известно о примерно 6000 сущностей Викиданных. К первому июня сервисом создана 18891 сущность. Итого в сущности Викиданных из списков Викигида экспортируются сведения о 27840 объектах.

Таким образом сведения об объектах культурного наследия встраиваются в семантическую паутину на базе проектов Викимедиа.

ЛИТЕРАТУРА

1. Leuf B., Cunningham W. The Wiki Way: Quick Collaboration on the Web. Boston : Addison-Wesley, 2001. 435 p.

2. Grzeganek K., Frost I., Gross, D. Spoilt for Choice – Wiki Software for Knowledge Management in Organisations [Электронный ресурс] // Community of Knowledge. 2011. Режим доступа: https://www.community-of-knowledge.de/fileadmin/user_upload/attachments/wikis_for_knowledge_management_in_organisations.pdf (дата обращения: 01.06.2022).

3. MediaWiki – a Wikimedia project. Режим доступа: https://www.mediawiki.org (дата обращения: 01.06.2022).

4. Культурное наследие России – Путеводитель Викигид. Режим доступа: https://ru.wikivoyage.org/wiki/Культурное_наследие_России (дата обращения: 01.06.2022).

5. Krötzsch M., Schaffert S., Vrandečić D. Reasoning in Semantic Wikis // Reasoning Web 2007 : Lecture Notes. Berlin : Springer-Verlag, 2007. Vol. 4636. P. 310-329.

6. RDF – Semantic Web Standards. Режим доступа: https://www.w3.org/RDF/ (дата обращения: 01.06.2022).

7. SPARQL 1.1 Query Language. Режим доступа: https://www.w3.org/TR/sparql11-query/ (дата обращения: 01.06.2022).

8. Peer-production system or collaborative ontology development effort: What is Wikidata? / C. Müller-Birn, B. Karran, J. Lehmann, M. Luczak-Rösch // OpenSym 2015 – Conference on Open Collaboration. San Francisco, 2015. doi: 10.1145/2788993.2789836.

9. Wikitech is the home of technical documentation for Wikimedia Foundation infrastructure and services. Режим доступа: https://wikitech.wikimedia.org/ (дата обращения: 01.06.2022).

REFERENCES

1. Leuf B., Cunningham W. *The Wiki Way: Quick Collaboration on the Web*. Boston : Addison-Wesley, 2001. 435 p.

2. Grzeganek K., Frost I., Gross, D. *Spoilt for Choice – Wiki Software for Knowledge Management in Organisations*. Community of Knowledge. 2011. Available at: https://www.community-of-knowledge.de/fileadmin/user_upload/attachments/wikis_for_knowledge_management_in_organisations.pdf (date accessed: 01.06.2022).

3. *MediaWiki – a Wikimedia project*. Available at: https://www.mediawiki.org (date accessed: 01.06.2022).

4. *Kulturnoye nasledie Rossii – Putevoditel Vikigid* [Cultural heritage of Russia – Travel guide of Wikivoyage]. Available at: https://ru.wikivoyage.org/wiki/Культурное_наследие_России (date accessed: 01.06.2022).

5. Krötzsch M., Schaffert S., Vrandečić D. *Reasoning in Semantic Wikis*. Reasoning Web 2007 : Lecture Notes. Berlin : Springer-Verlag, 2007. Vol. 4636. P. 310-329.

6. *RDF – Semantic Web Standards*. Available at: https://www.w3.org/RDF/ (date accessed: 01.06.2022).

7. *SPARQL 1.1 Query Language*. Availabale at: https://www.w3.org/TR/sparql11-query/ (date accessed: 01.06.2022).

8. Müller-Birn C., Karran B., Lehmann J., Luczak-Rösch M. *Peer-production system or collaborative ontology development effort: What is Wikidata?* OpenSym 2015 – Conference on Open Collaboration. San Francisco, 2015. doi: 10.1145/2788993.2789836.

9. *Wikitech is the home of technical documentation for Wikimedia Foundation infrastructure and services*. Available at: https://wikitech.wikimedia.org/ (date accessed: 01.06.2022).

ИНФОРМАЦИЯ ОБ АВТОРЕ

Соловьев Алексей Владимирович

Петрозаводский государственный университет, г. Петрозаводск, Россия, кандидат физико-математических наук, доцент, доцент кафедры информационно-измерительных систем и физической электроники,

E-mail: avsolov@petrsu.ru

Soloviev Alexei Vladimirovich

Petrozavodsk State University, Petrozavodsk, Russia, Associate Professor of Department of Information Measurement Systems and Physical Electronics, PhD, Assoc. Prof.,

E-mail: avsolov@petrsu.ru