ДОКУМЕНТАЛЬНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ

УДК 7/9:004.774.25

А.Б. Антопольский

Связанные открытые данные в цифровой гуманитаристике (обзор публикаций)

Предлагается обзор применения технологий связанных открытых данных в зарубежных проектах в сфере цифровой гуманитаристики. Выделяется несколько направлений: преобразование в LOD цифровых коллекций по культуре и искусству; интеграция разнородных данных, связанных с событием; библиографические ресурсы; языковые информационные ресурсы; музыкальные и музыковедческие данные. Особый, наиболее ценный вариант проектов по связанным данным – это технологические разработки и инфраструктурные проекты, создающие основу для национальных систем цифровой гуманитаристики. Подчеркивается роль онтологий в реализации проектов связанных данных.

Ключевые слова: *связанные открытые данные, цифровая гуманитаристика, цифровые коллекции, интеграция данных, библиографические ресурсы, языковые ресурсы, музыкальные данные*

DOI: 10.36535/0548-0019-2022-05-4

ВВЕДЕНИЕ

Технология связанных открытых данных (Linked Open Data – LOD) на платформе Семантической сети, с начала XXI в. стала ведущим направлением для представления научных данных, их интеграции и совместимости, а также коллабораций в этой области.

В области общественных (социальных и гуманитарных) наук технология связанных открытых данных также является наиболее перспективным направлением для интеграции информационных ресурсов. Она рассмотрена автором настоящей статьи в монографии [1].

В настоящей статье содержится обзор и анализ применения этой технологии в зарубежных проектах в сфере цифровой гуманитаристики (DH), подготовленный на основе исследования инфосферы DH, результаты которого опубликованы в работе [2].

Среди проектов в области DH, использующих технологию LOD, можно выделить следующие:

- инфраструктурные и технологические проекты на основе LOD;
- преобразование в LOD цифровых коллекций по культуре и искусству;
- интеграция разнородных данных, связанных с событием;
- библиотечно-библиографические ресурсы в LOD;
- языковые информационные ресурсы в LOD;
- представление в LOD музыкальных и музыковедческих данных.

ИНФРАСТРУКТУРНЫЕ И ТЕХНОЛОГИЧЕСКИЕ ПРОЕКТЫ

В настоящей статье не предполагается дать исчерпывающий анализ инфраструктурных и технологических решений, основанных на связанных открытых данных. Мы приведем лишь некоторые примеры, которые были реально использованы в нескольких проектах DH.

Примером инфраструктурного национального проекта связанных открытых данных для цифровых гуманитарных наук является проект LODI4DH [3], реализуемый в Финляндии.

LODI4DH – это совместная инициатива факультета компьютерных наук Университета Аалто и Центра цифровых гуманитарных наук Хельсинкского университета по созданию централизованных национальных сервисов связанных данных для открытой науки.

LODI4DH основан на большой сети для совместной работы и программном обеспечении, созданном в ходе выполнения длинной череды проектов с 2002 г., в результате которых было создано несколько используемых прототипов инфраструктуры, таких как онтология ONKI служба онтологии Finto в Нацио-

нальной библиотеке Финляндии. Эта служба развернула ONKI и занималась ее дальнейшим развитием на основе SKOS и на платформе Linked Data Finland LDF.fi.

ONKI/Finto и LDF.fi уже имеют широкую пользовательскую базу, что свидетельствует о востребованности инфраструктуры LODI4DH. Приложения на их основе также прошли путь от академических исследований до реального использования. Так, серия смысловых порталов Sampo имеет миллионы пользователей в Интернете. Многие музеи Финляндии, например, Городской музей Эспоо, Консорциум 8 музеев AKSELI и новая национальная система каталогизации KOOKOS, используют онтологии ONKI/Finto. В дополнение к финским проектам существует несколько совместных исследовательских проектов с международными университетами, такими как Оксфорд, Стэнфорд, Колорадо и Пенсильвания, в которых использовались финские службы связанных данных для DH. LODI4DH направлена прежде всего на исследовательскую инфраструктуру DH, но лежащие в ее основе технологии связанных данных и Семантической сети могут использоваться и в других областях исследований, что существенно расширяет потенциал этой инфраструктуры.

Так, данные от сотрудничающих организаций объединяются в общие открытые общедоступные онтологии для: 1) исторических мест и карт, 2) исторических лиц, 3) событий, 4) ключевых понятий и 5) времени. Эти основные онтологии, предоставляемые в виде веб-сервисов, применяются в качестве «семантического клея» при связывании и слиянии данных.

Еще один инфраструктурный проект связанных открытых данных LINCS [4] предназначен для канадских культурных исследований. Этот проект обеспечивает преобразование больших наборов данных во взаимосвязанную машинно-обрабатываемую сеть ресурсов. Исследователям нужна более умная семантическая сеть, которая встраивает смысл в машиночитаемые ссылки, чтобы прояснить различные взаимосвязи понятий и объектов. Технологии связанных открытых данных делают сеть умнее, структурируя и интегрируя данные. LINCS использует эти технологии для увязки канадских исследований и данных о наследии из всего Интернета, преобразуя, соединяя, улучшая и делая доступными ранее гетерогенные и разрозненные наборы данных. Такая увязка обеспечивает пути к новым идеям через сетевое производство знаний как внутри Канады, так и за ее пределами.

Как часть этого проекта осуществляется преобразование метаданных в открытые связанные данные о проектах в сфере культуры [5].

Среди технологических проектов, направленных на развитие LOD, можно указать на метод кодирования связанных данных с помощью JSON (JSON-LD). Соответствующая рекомендация разработана рабочей группой консорциума World Wide Web [6]. Одна из целей JSON-LD заключалась в том, чтобы потребовать от разработчиков как можно меньше усилий для преобразования существующего JSON в JSON-LD. В настоящее время этот стандарт поддерживается Рабочей группой JSON-LD [7].

Проект LodLive [8], разработанный группой итальянских специалистов, демонстрирует использование стандартов связанных данных (RDF, SPARQL) для просмотра ресурсов RDF. Приложение направлено на распространение принципов связанных данных с использованием простого и дружественного интерфейса с многоразовыми методами. Основной принцип, лежащий в основе LodLive, заключается в том, чтобы доказать, что ресурсы, публикуемые в соответствии со стандартом W3C SPARQL, могут быть легко доступны и понятны. Проект предполагает, что подход LodLive может стимулировать государственные администрации и крупных владельцев данных добавлять свои ресурсы в LOD и делиться ими. Можно начинать просмотр, запросив конечную точку для определенного ресурса, или начать с одного из приведенных примеров URI.

Веб-инструмент LodLive разработан для демонстрации стандартов связанных данных, применяемых к просмотру ресурсов RDF с помощью простого интерфейса. LodLive состоит из подключаемого модуля jQuery (lodlive-core.js), карты конфигурации JSON (lodlive-profile.js), HTML-страницы, нескольких изображений (спрайтов) и некоторых других общедоступных плагинов jQuery.

Интересный проект разработан кафедрой цифровых гуманитарных наук Университета Этвёша Лоранда (Венгрия) [9]. Он называется ELTEdata и направлен на организацию просопографических, библиографических и других данных в семантическую сеть и их публикацию. ELTEdata следует структуре Викиданных, связанной с Викиданными семантическими утверждениями и сущностями. Таким образом, ELTEdata может быть интерпретирована как часть Викиданных, хотя и полностью независима от этого ресурса.

Элементы ELTEdata имеют уникальный идентификатор, и каждый семантический оператор может быть описан как пара *свойство–значение*. Язык семантических запросов SPARQL обеспечивает сложный поиск и визуализацию на карте или временной шкале. Эта функция позволяет структурировать большой набор данных и кажется полезной при описании сетей. Слияние с внешними базами данных играет важную роль в формировании семантической библиографии, поэтому актуально сопоставлять некоторые свойства семантических операторов с их вариантами, содержащимися в интерфейсах структурированных метаданных.

К технологическим проектам в области DH можно отнести и проект LIDER *Связанные данные как средство кросс-медиа и многоязычной аналитики контента* [10], определяющий архитектуру, которая создается для анализа многоязычного и мультимедийного контента. Эталонная архитектура LIDER основана на открытых стандартах, существующих и будущих платформах и обеспечивает:

- эталонную модель, определяющую задачи, в которых лингвистические связанные данные могут поддерживать контент-аналитику, и обеспечивающую стандартную декомпозицию этих задач на элементы, которые совместно могли бы решать эти задачи вместе с потоками данных между элементами;

- каталог архитектурных шаблонов, описывающий типы элементов, которые могут быть использованы в вышеупомянутых задачах, типы взаимосвязей между этими элементами, и ограничения на то, как они могут быть использованы.

В проекте описываются различные источники информации для этой эталонной архитектуры, проблемы и препятствия, которые решены лишь частично. Собрана и упорядочена обширная коллекция задач и архитектурных шаблонов NLP в качестве основы для разработки согласованной справочной архитектуры, которая ориентирует первых пользователей данных на основе ссылок из целевых групп лидеров промышленных заинтересованных сторон. Кроме того, эталонная архитектура LIDER включена в контекст деятельности в области связанных данных.

КОЛЛЕКЦИИ КУЛЬТУРНОГО НАСЛЕДИЯ

Центральным направлением цифровой гуманитаристики, вероятно, следует признать создание и обработку разнообразных цифровых коллекций артефактов, относящихся к культурному наследию. Таких проектов в ходе проведенного обследования было выявлено достаточно много. Среди них было несколько, использующих технологии семантической сети и LOD.

Один из наиболее крупных проектов такого рода – это проект ModRef [11] (моделирование, репозиторий, цифровая культура), который объединяет проекты лаборатории Labex PasP [12] из Университета Парижа в Нантере и обеспечивает взаимодействие несколько организаций таких, как:

- MoDyCo (моделирование, динамика, корпус) [13],
- BDIC (Международная библиотека современной документации) [14],
- MAE (Дом археологии и этнологии) [15],
- ArScAn (Археология и античная наука) [16].

Цель ModRef состоит в том, чтобы провести цифровую экспертизу проектов Labex, а также доказать справедливость концепции LOD. В задачу проекта входит разработка моделирования с использованием ссылок.

Проект ModRef должен стимулировать обсуждение вопросов, связанных с миграцией данных в веб-семантику путем создания и использования «хранилищ троек» (коллекций или хранилищ данных RDF-файлов).

Перемещение данных в хранилища троек включает в себя различные этапы:

- подготовка данных (исследование и структурное описание данных),
- семантическое моделирование и сопоставление данных (или сопоставление и выравнивание),
- создание хранилищ троек – перенос данных в хранилища троек,
- публикация и визуализация хранилища троек,
- интерфейс для выполнения sparql-запросов в хранилища троек.

Основные проблемы – это (1) переход от неструктурированных или полуструктурированных данных (блокнот, тексты, html) к структурированным данным (электронные таблицы, реляционные базы данных, XML-файлы), а затем (2) перемещение этих структурированных данных в семантические данные (RDF-файлы) с целью улучшения обмена и открытия новых знаний.

В качестве нормированной онтологии была выбрана CIDOC-CRM [17], поскольку в настоящее время она является справочной для семантического описания данных в музееграфическом или культурном наследии.

CIDOC-CRM доступна в формате OWL, который предоставил Университет Эрлангена-Нюрнберга [18].

Следует отметить, что на сайте ModRef приводится перечень организаций, которые работают над преобразованием метаданных каталогов своих коллекций культурного наследия в связанные открытые данные, часть из них использует в качестве онтологии CIDOC-CRM:

- Британский музей [19],
- Йельский центр британского искусства [20],
- Arches [21] – результат сотрудничества между Институтом сохранения Гетти (GCI) и Всемирным фондом памятников (WMF) по недвижимому культурному наследию,
- Портал Biblissima [22] предоставляет унифицированный доступ к набору цифровых данных о средневековых рукописях, инкунабулах и раннепечатных книгах, выпущенных партнерами консорциума Biblissima. На сайте указано, что предполагается представление всех данных проекта в формате LOD. В настоящее время в этом формате доступны авторитетные файлы системы.

Упомянутые на сайте ModRef известные проекты DH вместо онтологии CIDOC-CRM используют другие системы метаданных, в том числе:

- Онлайн-энциклопедию DBPedia [23] – системы метаданных dbpedia, foaf, umbel, schema.org, dublin core, geo,
- Сервис для хранения, документирования и публикации данных Nakala [24] – системы метаданных foaf, skos, dublin core, vcard,
- Платформа исторической информации Symogih [24] – системы метаданных symogih, example.org, geo.

Потребность в создании связанных открытых данных была рассмотрена в проекте PHAROS Международного консорциума фотоархивов [26]. Это первый шаг на пути к разработке реальной цифровой инфраструктуры фотографических архивов произведений искусства в Европе и Соединенных Штатах Америки. Консорциум организовал сотрудничество между учреждениями, ответственными за четырнадцать фотоархивов с тем, чтобы создать общую платформу для исследований изображений и метаданных.

В качестве примера работы с фотобанками опишем проект преобразования метаданных фотоархива Зери в LOD [27], который включает 290 тыс. фотографий памятников и произведений искусства. Каталогизация хранилища Зери была проведена на основе двух итальянских стандартов метаданных: Scheda F для фотографий и Scheda OA для произведений искусства.

Первый выпуск набора данных Zeri Photo Archive RDF представляет собой значительное подмножество данных, уже доступных на веб-сайте каталога Zeri и для поиска через портал Europeana – это в основном произведения искусства Нового времени (XV–XVI вв.): около 19.000 – сами произведения и более 30 тыс. фотографий, описанных с помощью примерно 11 млн троек RDF.

ИНТЕГРАЦИЯ ДАННЫХ, СВЯЗАННЫХ С СОБЫТИЕМ

Наиболее интересное с точки зрения цифровой гуманитаристики применение технологии LOD заключается в методике интеграции данных и метаданных, посвященных историческим или культурным событиям и объектам культурного наследия, в разных системах хранения, прежде всего в библиотеках, архивах, музеях. Нам уже приходилось делать обзор методов интеграции таких данных [28]. Однако в последние годы в мире было реализовано много новых проектов, интегрирующих разнородные данные на основе технологии LOD. Здесь мы опишем более подробно два таких проекта из различных сфер цифровой гуманитаристики, и кратко – еще несколько.

Первый пример из области социальной психологии. Проект направлен на представление информации по *Стэнфордскому тюремному эксперименту (SPE)* в виде LOD [29] и начинается с отбора по критериям релевантности и неоднородности исходных пунктов (документов), отражающих содержание события (SPE) с различных точек зрения. Для каждого пункта предоставляются основные метаданные, краткое описание и справочная ссылка.

На основе этих пунктов были идентифицированы сущности SPE и отношения между ними. Полученная сложная сеть сущностей затем была очерчена с помощью концептуальной карты (хотя карта является лишь предварительным эскизом, она дает четкий обзор и общее понимание сценария проекта), на основе которой строится модель E/R (сущность/отношение) для SPE. Далее выполняется анализ метаданных и их выравнивание, т.е. приведение к фиксированному списку используемых систем метаданных. Инструментом выравнивания послужил стандарт Dublin Core.

Для каждого объекта разработчики этой модели попытались ответить на вопросы WHO?, WHERE?, WHEN? и WHAT?, определив свойства этих объектов и оформив их в виде стандартных терминов (дескрипторов).

Теоретическая модель подробно предоставляет список свойств, установленных для каждого элемента, и связи между всеми сущностями, включенными в проект, определяя свойства на некоторых авторитетных онлайн-ресурсах таких как, Wikidata, DBpedia, TMDb, Last.FM, Getty Vocabularies, WorldCat и т. д. Эта модель позволила уточнить и расширить модель E/R.

Переходя от теоретической модели к концептуальной модели, разработчики определили наиболее подходящие онтологии, чтобы детально представить элементы проекта. Для подбора онтологий были использованы некоторые инструменты, в частности схема классификации Seeing Standards [30] и портал Linked Open Vocabularies (LOV) [31]. Были выбраны такие наиболее распространенные онтологии:

- *FOAF* (Friend Of A Friend) – онтология, описывающая людей, их деятельность и их отношения с другими людьми и объектами;
- *Schema.org* – стандарт семантической разметки данных в сети, объявленный летом 2011 г. поисковыми системами Google, Bing и Yahoo!;
- *Дублинское ядро.* Термины перечисляют текущий набор словаря Dublin Core, т. е. 15 основных элементов плюс все квалифицированные термины. Он может использоваться для нескольких целей – от простого описания ресурсов до объединения словарей метаданных различных стандартов метаданных и до обеспечения совместимости словарей метаданных в связанном облаке данных и реализациях семантической паутины;
- *CIDOC-CRM*, в настоящее время ISO/CD21127, является основной онтологией, цель которой – обеспечение обмена и интеграции между разнородными источниками информации о культурном наследии, архивами и библиотеками;
- *BIBO* – библиографическая онтология предоставляет основные понятия и правила для описания цитат и библиографических ссылок (т.е. цитат, книг, статей и т.д.);
- *Онтология соглашений* предназначена для моделирования социальных контрактов, которые включают лицензии, законы, контракты, стандарты и метаданные определений;
- *Музыкальная онтология* предоставляет словарь для публикации и связывания широкого спектра данных, связанных с музыкой, в Интернете.

Для всех сущностей в этой модели были построены пространства имен и определены уникальные идентификаторы. Это позволило сформировать RDF-тройки, которые размещены на сайте. Конечная цель проекта – формирование графа знаний события.

Другой пример – интеграция разнородных объектов и сущностей, связанных со сложным историческим событием – «*Фестивалем музыки и искусств Вудсток*» [32]. Главная цель этого проекта – попытка добиться взаимосвязанной системы информации о событии, отталкиваясь от уже существующих в сети ресурсов. Схема выполнения проекта похожа на уже описанную. Сначала были выбраны 10 объектов из разных библиотек, архивов, музеев, посвященных Вудстоку. Для этих объектов на основе естественного языка с помощью интуитивно понятной модели E/R были выделены основные сущности и отношения. Затем разработчики проекта проанализировали стандарты метаданных, которые использовались при описании объектов. Сущности и отношения сравнивались со стандартами метаданных, чтобы выявить наиболее важные аспекты в описании данных.

Когда получили более четкое представление о предметной области, было проведено моделирование данных на абстрактном уровне: построена теоретическая модель, обеспечивающая общее естественно-язычное описание информации о сущностях и отношениях, задействованных в проекте.

Это был промежуточный этап, ведущий к созданию концептуальной модели, способной формализовать описание особенностей данных за счет использования уже существующих онтологических формальных языков. Благодаря различным онтологиям удалось представить и описать данные как формализованные понятия, пригодные для выражения RDF-высказывания в виде триплетов, представленных субъектами, предикатами и объектами. Полученные RDF-высказывания были сериализованы через Turtle, что позволило представить данные в виде триплета URI, выраженного – где это возможно – через онтологические словари.

Были созданы такие взаимосвязи между полученными данными, данными органов власти и других организаций, участвующих в проекте, а также другими ресурсами из онлайн-репозиториев или веб-страниц.

Наконец, все результаты проекта были представлены в виде графа знаний, изображающего контекстуализированную информационную сеть данных, связанных с фестивалем Вудстока, что позволило выявить скрытые отношения, составляющие реальную сущность чего-либо.

Кратко опишем еще несколько проектов такого рода.

WarSamp [33] – это первая крупномасштабная система для обслуживания и публикации LOD о Второй мировой войне: 1) инициирует и способствует крупномасштабной публикации этих данных из распределенных разнородных хранилищ и 2) демонстрирует и предлагает их использование в приложениях и исследованиях DH. Граф знаний содержит более 9 млн высказываний (троек) между элементами данных, включая, например, полный набор из более чем 95 тыс. записей о смертях финских солдат во время Второй мировой войны. В виде RDF-высказываний представлены 160 тыс. сделанных во время войны подлинных фотографий, в том числе 32 тыс. фото исторических мест на исторических картах, 23 тыс. военных дневников воинских частей и 3400 мемуарных статей, написанных ветеранами после войны. Информация в WarSampo поступает из нескольких финских организаций и источников,

WarSampo состоит из двух отдельных компонентов: 1) служба данных WarSampo для машин и 2) семантический портал WarSampo с различными приложениями для пользователей.

Цель проекта *Битва при Ватерлоо в формате LOD* [34] – создание абстрактной модели LOD для описания данных, связанных с битвой при Ватерлоо. Модель должна соотнести событие с личностью «Наполеон Бонапарт», местом «Ватерлоо», датой «1815» и концепцией «Поражение». Последовательность действий: выбор 10 предметов, включающих архивные документы, публикации и артефакты, описывающие идею «Битва при Ватерлоо»; согласование между используемыми учреждениями культурного наследия различными стандартами метаданных, относящимися к людям, месту, дате и концепции. Цель – разработка модели, способной описывать выбранные предметы, и отвечать на вопросы:

- Как описать людей?
- Какая информация о местах?
- В каком формате понятие времени?
- Каково основное содержание объектов?

Модель должна применять существующие формальные системы и онтологии такие, как FOAF, RDFS, SKOS, DC, EAC-CPF.

История экспедиции «Кон-Тики» в формате LOD [35]. При реализации этого проекта была сделана подборка из десяти предметов, связанных с историей Кон-Тики. Для каждого предмета указано название, тип, связанный источник и краткое описание. Сценарий был представлен через модель E/R, содержащую выбранные элементы, сущности и наиболее релевантные отношения между ними. Затем были определены стандарты метаданных, принятые для описания этих предметов.

Чтобы обеспечить функциональную совместимость между стандартами, были определены соответствия между свойствами метаданных. Теоретическая модель – это сценарий, описывающий элементы, связанные метаданные и отношения между ними в абстрактном виде. Модель формально представляет рассматриваемое событие, используя существующие онтологии: графическое изображение было создано, чтобы показать результирующую модель. Выбранные элементы описаны в соответствии с концептуальной моделью с помощью набора таблиц, каждая из которых представляет предметы, предикаты и объекты, подходящие для описания особенностей предметов и основных событий. Сущности и элементы, репрезентативные для данной области, были записаны в виде высказываний RDF, а затем с использованием сериализации Turtle, представлены и объединены. В результате получено графическое изображение знаний о данном событии.

Еще один проект – представление в виде LOD метаданных объектов разных типов (из архивов, музеев, библиотек) по различным аспектам фильма "*Сладкая жизнь*" [36]. По случаю шестидесятилетия выхода этого фильма была реализована модель LOD, ориентированная на интеграцию данных о производстве и распространении этого культового фильма режиссера Федерико Феллини. Данные, когда они изолированы, имеют ограниченный потенциал, их ценность возрастает, когда один или несколько наборов данных, подготовленных и опубликованных независимо и различными субъектами, предлагают возможность интеграции и контакта между ними с целью создания нового общего знания.

Проект История политического диссидента в формате LOD [37] исследует такое событие, как задержание Патрика Заки и его содержание в египетской тюрьме. Цель проекта – моделирование LOD, устанавливающих концепции, предметы, людей, учреждения и места, связанные с задержанием Патрика Заки, чтобы отобразить семантически значимую среду наиболее интегрированным образом.

Было рассмотрено 4 понятийные области:

- эмоции художников, которые создавали визуальные произведения искусства и музыку об этом событии;
- контекстуальная информация об окружающей среде, связанной с Патриком в прошлом и настоящем, об учебном заведении, здании, в котором его задержали;
- условия содержания в тюрьмах Египта на основе докладов Хьюман Райтс Вотч и подкаста итальянского радио RAI;
- юридические аспекты этого события по резолюции, принятой Европейским парламентом, и петиции, подписанной учеными в защиту Заки.

В проекте были изучены возможные технологии, используемые в семантической сети. Упомянем еще два проекта данного класса:

- интеграция ресурсов библиотек, архивов и музеев, связанных с гибелью «Титаника», с использованием LOD [38];

- исследование гендерной проблематики в искусствоведении на основе ARTchives и Викиданных, с использованием технологии LOD [39].

Приведенные примеры убедительно показывают возможность использования технологии связанных открытых данных – LOD для тематически и функционально различных областей цифровой гуманитаристики – DH, когда ставится задача представить по возможности полную информацию о событии с использованием разнородных источников.

БИБЛИОТЕЧНО-БИБЛИОГРАФИЧЕСКИЕ РЕСУРСЫ В ТЕХНОЛОГИИ СВЯЗАННЫХ ОТКРЫТЫХ ДАННЫХ

Специфическим видом ресурсов, которые также можно отнести к области цифровой гуманитаристики, являются библиотечные данные, в том числе библиотечные каталоги, библиографические БД и авторитетные (нормативные) файлы. Библиотечные данные очень удобно представлять в виде связанных открытых данных, и работы в этом направлении в ведущих библиотеках мира ведутся с начала XXI в.

Специально для представления библиотечных данных в LOD разработаны модель Библиотеки Конгресса BIBFRAME [40] и модель Международной федерации библиотечных ассоциаций [41], которые активно применяются различными библиотеками. Общее состояние внедрения LOD в библиотеки можно оценить по опросу, который проводил OCLC [42]. Цель этого опроса – изучение опыта тех библиотек, где реализованы или реализуются проекты/услуги связанных данных. Результаты этого опроса приводятся на сайте OCLC [43], а также в презентации [44]. OCLC и сам ведет исследования технологии LOD [45]. В рамках этого пилотного проекта OCLC и пять партнерских учреждений изучили методы и целесообразность преобразования метаданных в связанные данные для улучшения возможности обнаружения и управления оцифрованными культурными материалами и их описаниям.

В российской информатике известность получил проект О.Л. Лавреновой и ее коллег по представлению в формате LOD классификации знаний, а именно Библиотечно-библиографической классификации (ББК) [46].

Последовательным пропагандистом применения связанных данных в библиотечном деле является О.Н. Жлобинская. Один из последних принадлежащих ей обзоров зарубежного опыта представлен в презентации [47]. Развернутая модель преобразования в LOD библиографических данных из формата RUSMARC, разработанная этим автором, содержится в работе [48].

ЯЗЫКОВЫЕ ИНФОРМАЦИОННЫЕ РЕСУРСЫ

Наиболее очевидным и адекватным объектом для представления в виде связанных данных являются языковые информационные ресурсы, поскольку аналогичные языковые ресурсы в форме тезаурусов и онтологий активно создавались в течение последних десятилетий, а исследование семантических сетей – это традиционная область лингвистической семантики.

Описание технологии LOD применительно к языковым информационным ресурсам и современное состояние языковых связанных открытых данных (LLOD) содержится в монографии Ф.Джимиано и его соавторов [49]. Некоторые действующие проекты в этой области рассмотрены в нашей работе [50]. Поэтому в настоящей статье мы ограничимся перечислением языковых связанных открытых данных и проектов, вошедших в цитированное выше исследование инфосферы цифровой гуманитаристики:

- лингвистические связанные открытые данные [51];
- формат кросс-лингвистических связанных данных [52];
- преобразование лингвистических данных в связанные открытые данные [53];
- кросс-лингвистические связанные данные [54];
- связанные открытые словари [55];
- связанные данные в лингвистике [56];
- связывание корпусных данных способом удобным для NLP [57];
- информация о языках в формате связанных данных [58];
- словари Getty как связанные открытые данные [59].

Кажется, что единственный российский проект этого направления – это проект Д. Усталова по интеграции тезаурусов русского языка в связанные данные [60].

Предложенный список не является исчерпывающим, но дает достаточно полное представление о масштабах применения технологии LOD к компьютерной лингвистике.

МУЗЫКАЛЬНЫЕ И МУЗЫКОВЕДЧЕСКИЕ СВЯЗАННЫЕ ДАННЫЕ

Разработки в области цифровизации музыкальных данных и тем более представление их в виде связанных данных – это относительно новое направление цифровой гуманитаристики и поэтому не получило такого развития, как перечисленные выше. В нашем исследовании было обнаружено два подобных проекта, краткие сведения о которых мы приводим.

Встраивание графов и преобразование музыкальных данных в связанную форму [61]. Цель настоящего проекта состояла в том, чтобы извлечь необработанные данные о музыкальных записях и преобразовать их в связанную форму, из которой можно извлечь знания.

Авторы проекта указывают, что известные ресурсы DBpedia [62] и Викиданные уже предлагают огромное количество сущностей, связанных с музыкой, но для непопулярных музыкальных записей их свойства часто очень ограничены. Использование гораздо более широкого набора данных, специализированных на музыкальных записях, позволяет делать более интересные и точные прогнозы. В области музыкальных записей это может позволить создать систему, которая не только автоматически предсказывает жанры, характеристики и особенности альбомной записи, но и вероятность ее сходства с учетом средних оценок, выставленных пользователем.

Джазовые коллекции в LOD – ресурс для изучения джазовых исполнителей, выступлений и сетей [63].

Проект использует семантические веб-технологии для соединения данных о джазе, включая дискографии, сольные транскрипции и информацию об исполнителях. Проект предполагает интеграцию с другими наборами данных и разработку интерфейса для семантического поиска и визуализации данных.

ЗАКЛЮЧЕНИЕ

Из приведенного обзора связанных открытых данных в цифровой гуманитаристике с очевидностью следует, что в мире сложилась методика создания информационного ресурса для анализа предметной области на основе технологии связанных данных. Чаще всего эта технология применяется при формировании графа знаний применительно к событию, персоне, или тематической коллекции. Для таких направлений цифровой гуманитаристики как библиотечные и языковые данные разработаны специфические модели преобразования данных в форму связанных открытых данных 8 – LOD.

Особый, наиболее ценный вариант работ по связанным данным – это технологические разработки и инфраструктурные проекты, создающие основу для национальных систем цифровой гуманитаристики, как это сделано в Финляндии и в Канаде.

Важная особенность проектов по связанным данным заключается в том, что они опираются на уже созданные онтологии и системы метаданных, которые фиксируют сложившийся в мировой науке консенсус по представлению знаний в определенных областях. Поэтому необходимо создание русскоязычной онтологии гуманитарного знания на основе как российских, так и международных источников, что могло бы стать важным компонентом инфраструктуры российской цифровой гуманитаристики.

СПИСОК ЛИТЕРАТУРЫ

1. Антопольский А.Б., Ефременко Д.В. Инфосфера общественных наук России: монография / под ред. В.А. Цветковой. – Москва; Берлин : Директ-Медиа, 2017. – 676 с. DOI 10.23681/468227.
2. Антопольский А.Б. Инфосфера цифровой гуманитаристики: опыт анализа // Информационные ресурсы России. – 2022. – № 1. – С. 30-38.
3. Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH). – URL: https://seco.cs.aalto.fi/projects/lodi4dh/
4. Linked Infrastructure for Networked Cultural Scholarship. – URL: https://lincsproject.ca/
5. CoDHR. – URL: http://codhr.dh.tamu.edu/2018/04/24/linked-infrastructure-for-networked-cultural-scholarship-lincs
6. JSON-LD 1.1 A JSON-based Serialization for Linked Data. Draft Community Group Report 19 April 2019. – URL: https://json-ld.org/spec/latest/json-ld/
7. JSON-LD Working Group. – URL: https://www.w3.org/2018/json-ld-wg/
8. Live on LodLive. – URL: http://en.lodlive.it/
9. ELTEdata-project. – URL: https://eltedata.elte-dh.hu/wiki/Main_Page
10. LIDER: FP Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe FP7-610782D3. – URL: https://docplayer.net/139432478-Lider-fp-linked-data-as-an-enabler-of-cross-media-and-multilingual-content-analytics-for-enterprises-across-europe.html
11. ModRef Project: Modelling, References, Digital Culture. – URL: http://modref-labexpassespresent.huma-num.fr/ModRef/
12. Labex Past in Present: history, heritage, remembrance – Labex Les Passés dans le Présent: histoire, patrimoine, mémoires. – URL: http://passes-present.eu/
13. MoDyCo. – URL: http://www.modyco.fr/fr/
14. Bibliothèque de la Documentation Internationale Contemporaine. – URL: http://www.bdic.fr/
15. Maison de L'Archéologie et de L'Ethnologie. – URL: http://www.mae.u-paris10.fr/
16. Archéologies et Sciences de l'Antiquité (ArScAn). – URL: http://www.arscan.fr/
17. CIDOC-CRM. – URL: http://www.cidoc-crm.org
18. The Erlangen CRM / OWL. – URL: http://www.erlangen-crm.org/
19. The British Museum. Explore the collection. – URL: https://www.britishmuseum.org/collection
20. Yale Center for British Art. – URL: https://britishart.yale.edu/collections/using-collections/technology/linked-open-data
21. The Getti Conservation Institute Arches Project. – URL: http://www.getty.edu/conservation/our_projects/field_projects/arches
22. Biblissima, the Observatory for Medieval and Renaissance Written Cultural Heritage. – URL: https://biblissima.fr
23. DBPedia. – URL: http://www.dbpedia.org/sparql
24. NAKALA. – URL: https://www.nakala.fr/about
25. SyMoGIH project. – URL: http://www.symogih.org/
26. PHAROS. – URL: http://pharosartresearch.org/
27. The Zeri & LODE project. – URL: http://data.fondazionezeri.unibo.it/
28. Антопольский А.Б. Интеграция библиотечных и архивных информационных систем. – [Б.и.] . – 5 с. – URL: https://rucont.ru/efd/158
29. The Stanford Prison Experiment on LOD. – URL: https://spelod.github.io/#erModel
30. Seeing Standards: A Visualization of the Metadata Universe. – URL: http://jennriley.com/metadatamap/
31. Linked Open Vocabularies (LOV). – URL: https://lov.linkeddata.es/dataset/lov
32. Woodstock Music and Art Festival. – URL: https://woodslod.github.io/woodsLOD/
33. WarSampo Finnish World War II on the Semantic Web. – URL: https://www.sotasampo.fi/en/
34. Battle of the WaterLOD. – URL: https://waterlod.github.io/index.html
35. The Kon-Tiki Expedition. A Linked Open Data project. – URL: https://kontikilod.github.io/KonTiki/
36. La Dolce Vita – Linked Open Data. – URL: https://fellini-lod.github.io/contenuto.html
37. Patrick aLOuD. – URL: https://patrickaloud.github.io/
38. The INKING of RMS Titanic. – URL: https://linkingoftitanic.wixsite.com/linkingtitanic
39. Martrioska. – URL: https://martrioska.github.io/martrioska.html

40. Bibliographic Framework Initiative. – URL: https://www.loc.gov/bibframe/
41. IFLA Library Reference Model. – URL: https://www.librarianshipstudies.com/2020/04/ifla-library-reference-model-lrm.html
42. Online Computer Library Center. – URL: https://www.oclc.org/en/home.html?Redirect =true
43. Linked Data Survey results 6 – Advice from the implementers. – URL: https://hangingtogether.org/linked-data-survey-results-6-advice-from-the-implementers/
44. Smith-Yoshimura K. Linked data implementations – who, what, why? – URL: https://www.oclc.org/content/dam/research/events/2018/smith-yoshimura-lnked-data-implementations-who-whatwhy-SWIB18.pptx
45. Bahnemann G., Carroll M., Clough P., Einaudi M., Ewing Ch., Mixter J., Roy J., Tomren H., Washburn B., Williams E. Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project. – Dublin, OH: OCLC Research, 2021. – URL: https://doi.org/10.25333/fzcv-0851.
46. LINKED OPEN DATA. Классификационная система организации знаний. – URL: https://lod.rsl.ru/
47. Жлобинская О.Н. Библиотечные связанные данные: анализ зарубежного опыта . – URL: http://www.nilc.ru/text/NMLBD/NMLBD4.pdf
48. Жлобинская О.Н. Представление библиотечных данных в LOD: возможности и перспективы формата RUSMARC. – URL: http://www.rusmarc.ru/publish/%D0%92%D0%BE%D0%B7%D0%BC%D0%BE%D0%B6%D0%BD%D0%BE%D1%81%D1%82%D0%B8%20%D0%B8%20%D0%BF%D0%B5%D1%80%D1%81%D0%BF%D0%B5%D0%BA%D1%82%D0%B8%D0%B2%D1%8B%20RUSMARC_%D0%A0%D0%BE%D1%81%D1%82%D0%BE%D0%B2.pdf
49. Cimiano Ph., Chiarcos Ch.; McCrae J. P.; Gracia J. Linguistic Linked Data: Representation, Generation and Applications. – Springer International Publishing, 2020.
50. Антопольский А.Б. Лингвистические связанные открытые данные: состояние и перспективы // Научно-техническая информация. Сер. 2. – 2021. – № 8. – С. 28-36. DOI: 10.36535/0548-0027-2021-08-4
51. Linguistic Linked Open Data. – URL: http://linguistic-lod.org/
52. Cross-Linguistic Data Formats. – URL: https://cldf.clld.org/
53. The Prêt-à-LLOD Project . – URL: https://pret-a-llod.github.io/
54. CLLD – Cross-Linguistic Linked Data. – URL: https://clld.org/
55. Linked Open Dictionaries. – URL: http://ionov.me/liodi/
56. Workshop on Linked Data in Linguistics (LDL). – URL: https://www.aclweb.org/anthology/venues/ldl/
57. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. – URL: https://www.researchgate.net/publication/318134320_CoNLL-RDF_Linked_Corpora_Done_ in_an_NLP-Friendly_Way
58. Lexvo.org. – URL: http://www.lexvo.org
59. Getty Vocabularies as Linked Open Data. – URL: http://www.getty.edu/research/tools/vocabularies/lod/index.html#definition
60. Усталов Д.А. Тезаурусы русского языка в виде открытых связанных данных. – URL: https://www.dialog-21.ru/media/1103/ustalovda.pdf .
61. Graph embedding and link prediction starting from a non-linked musical dataset. – URL: https://alerosae.github.io/FromRaw2Linked/
62. DBpedia. Global and Unified Access to Knowledge Graphs. – URL: https://www.dbpedia.org/
63. JazzCats (Jazz Collection of Aggregated Triples). – URL: https://jazzcats.cdhr.anu.edu.au/

Материал поступил в редакцию 17.02.22.

Сведения об авторе

АНТОПОЛЬСКИЙ Александр Борисович – доктор технических наук, профессор, главный научный сотрудник ИНИОН РАН, Москва
e-mail: ale5695@yandex.ru