СПРАВОЧНО-ИНФОРМАЦИОННЫЙ РАЗДЕЛ

УДК 81’322:004.9

А.Б. Антопольский

Лингвистические связанные открытые данные: состояние и перспективы

Представлен проект лингвистических связанных открытых данных (Linguistic Linked Open Data – LLOD), реализуемый международной рабочей группой на платформе Семантической сети. Излагаются принципы проекта, перечисляются принятые в нем стандарты и методики, рассматриваются главные ресурсы, загруженные в облако проекта в настоящее время. Обсуждаются основные понятия и приводится классификация языковых ресурсов, представленных в проекте. История проекта описывается через изложение материалов семинара участников LLOD в течение 2012-2020 гг. Приводятся проекты по развитию LLOD и преобразованию в него различных языковых ресурсов, которые выполняют исследовательские коллективы по всему миру. Делается вывод, что этот проект является наиболее перспективным направлением по интеграции и коллаборации в области языковых технологий и ресурсов.

Ключевые слова: *лингвистические ресурсы, языковые ресурсы, языковые технологии, лингвистические связанные открытые данные, облако LLOD, семинар по связанным данным, коллаборации*

DOI: 10.36535/0548-0027-2021-08-4

ВВЕДЕНИЕ

Быстрое развитие языковых технологий, языковых и лингвистических ресурсов (ЛР), создаваемых как в научно-образовательных целях, так и для различных прикладных задач, потребовало от лингвистического сообщества выработать стратегию по интеграции ЛР, их унификации, обеспечению совместимости и возможности повторного использования, а также по организации международной коллаборации при создании и поддержке лингвистических ресурсов.

Наиболее известным направлением для решения этих задач является проект *Лингвистические связанные открытые данные* (Linguistic Linked Open Data – LLOD. – URL: https://linguistic-lod.org/llod-cloud) как неотъемлемая часть Семантической сети.

Платформа Семантической сети разрабатывается многочисленными исследовательскими группами из разных стран под общим руководством Консорциума Всемирной паутины (W3C – The World Wide Web Consortium. – URL: https://www.w3.org), председателем которого является создатель Интернета Тим Бернерс-Ли.

Реализацию проекта LLOD осуществляет Международная рабочая группа по открытой лингвистике (OWLG – The Open Linguistics Working Group – URL: https://ru.scribd.com/ document/126882831/The-Open-Linguistics-Working-Group). Современное состояние проекта представлено в коллективном труде Ф. Джимиано и его соавторов [1].

Проект LLOD представляет собой международную коллаборацию, цель которой – создание, поддержание и развитие облака LLOD, а также необходимых для этого стандартов, методик и инструментов.

ОПИСАНИЕ ПРОЕКТА LLOD

Принципы открытых связанных данных

LLOD создается в соответствии с принципами открытых связанных данных, сформулированными в рекомендациях W3C:

- данные должны быть открытыми и лицензированы с использованием таких лицензий, как Creative Commons;
- элементы в наборе данных должны быть однозначно идентифицированы с помощью унифицированного идентификатора ресурса (URI – Uniform Resource Identifier);
- URI должен давать возможность пользователям получать доступ к дополнительной информации с помощью веб-браузеров;

- ресурс LLOD должен возвращать пользователю результаты с использованием веб-стандартов, таких как Resource Description Framework (RDF);
- должны быть включены ссылки на другие ресурсы, чтобы помочь пользователям их открывать и создавать условия для точности семантики запросов и других транзакций.

Реализация LLOD обеспечивает следующие основные преимущества:

- *представление*: связанные графы – это более гибкий формат представления лингвистических данных;
- *совместимость*: общие модели RDF могут быть легко интегрированы;
- *федеративность*: данные из нескольких источников можно легко объединять;
- *открытость*: инструменты для RDF и связанных данных должны быть свободно доступны по лицензиям с открытым исходным кодом;
- *выразительность*: существующие словари помогают представлять все необходимые ЛР;
- *семантика*: ссылки должны точно определять понятия;
- *динамичность*: веб-данные можно постоянно улучшать.

Стандарты и форматы проекта LLOD

Помимо непосредственного создания облака LLOD, сообщество OWLG стимулирует разработку стандартов в отношении словарей, метаданных и других аспектов лингвистических ресурсов. Согласно обзору Ф. Джимиано [1] к ним относятся:

- OntoLex-Lemon – стандарт для моделирования лексических ресурсов (машиночитаемые словари, многоязычная терминология, лексикализация онтологий);
- модели лингвистических аннотаций (в корпусах или для NLP – Natural Processing Language):
  - Web Annotation – стандарт W3C для аннотации веб-ресурсов (текстовых или иных),
  - формат обмена NIF (NLP Interchange Format) – стандарт сообщества для грамматической аннотации текста,
  - CoNLL-RDF – словарь на основе NIF для представления RDF корпусов в традиционных форматах (CoNLL),
  - POWLA – словарь для общих лингвистических структур данных, которые можно использовать для дополнения NIF, CoNLL-RDF или Web Annotation;
- стандарты для категорий лингвистических данных:
  - Онтологии лингвистической аннотации (OLiA – Ontologies of Linguistic Annotation),
  - LEXINFO для грамматических и других функций в лексических ресурсах;
- стандарты для идентификация языка:
  - в виде строк с языковыми тегами IETF BCP 47 (https://tools.ietf.org/html/bcp47),
  - ISO 639-3 (URI, предоставленными lexvo.org),
  - Glottolog (URI для языковых разновидностей, не охваченных ISO 639);
- стандарты для метаданных:
  - Dublin Core,
  - словарь каталогов данных (DCAT – Data Catalog Vocabulary) – стандарт W3C для каталогов данных, опубликованных в Интернете,
  - METASHARE-OWL – словарь для метаданных языковых ресурсов.

Для реализации связанных данных требуется приложение RDF или соответствующие стандарты – это рекомендации W3C: SPARQL, Turtle, JSON-LD, RDF-XML, RDFa и т. д. Однако в языковых технологиях и лингвистических ресурсах в настоящее время более популярны другие формализмы, поэтому регулярно возникает вопрос о включении в облако LLOD тех или иных ЛР. Для нескольких таких формализмов существуют стандартизированные W3C механизмы преобразования (например, для XML, CSV или реляционных баз данных), и такие данные могут быть интегрированы при условии, что соответствующее отображение предоставляется вместе с исходными данными

Избранные ресурсы проекта LLOD

Облако LLOD развивается достаточно быстро и в нем уже размещены сотни лингвистических ресурсов. Далее приводится список десяти ЛР с наибольшим количеством связей, представленных в LLOD по состоянию на октябрь 2018 г. (в порядке количества связанных наборов данных):

- Онтология лингвистической аннотации (OLiA, связанная с 74 наборами данных) –предоставляет справочную терминологию для лингвистических аннотаций и грамматических метаданных;
- WordNet (связан с 51 набором данных) – лексическая БД для английского языка и сводная база для разработки аналогичных БД для других языков в нескольких редакциях (редакция Princeton связана с 36 наборами данных; редакция W3C – с 8 наборами данных; версия VU – с 7 наборами данных);
- DBpedia (связан с 50 наборами данных) – многоязычная база знаний, основанная на Википедии;
- Онтология LEXINFO (связана с 36 наборами данных) – предоставляет справочную терминологию для лексических ресурсов;
- BabelNet (связана с 33 наборами данных) – многоязычная лексикализованная семантическая сеть, основанная на агрегировании различных других ресурсов, в первую очередь WordNet и Wikipedia;
- LEXVO (связана с 26 наборами данных) – предоставляет идентификаторы языка и другие данные, связанные с языком, обеспечивает представление RDF ISO 639-3 трехбуквенных кодов для идентификаторов языков и информации об этих языках;
- Реестр категорий данных ISO 12620 (ISOcat; версия RDF, связанная с 10 наборами данных) – предоставляет частично структурированный репозиторий для различной терминологии, связанной с ЛР. Хостинг ISOcat находится в проекте DOBES, в Языковом архиве Института психолингвистики им. Макса Планка, но в настоящее время осуществляется переход на CLARIN;
- UBY (RDF версия, связан с 9 наборами данных) – лексическая сеть для английского языка, собранная из различных лексических ресурсов;

- Glottolog (связан с 7 наборами данных) – предоставляет детализированные идентификаторы для языков с низким уровнем ресурсов, в частности, многие из них не охвачены lexvo.org;
- Викисловарь – Ссылки на DBpedia (wiktionary.dbpedia.org, связан с 7 наборами данных) – лексикализация концепций DBpedia на основе викисловаря

Приложения LLOD для решения проблем научных исследований.

- Во всех областях прикладной лингвистики, компьютерной филологии и обработки естественного языка лингвистическая аннотация и лингвистическая разметка представляют собой центральные элементы анализа. Однако прогрессу в этой области препятствуют проблемы совместимости, в первую очередь – различия в словарях и схемах аннотаций, применяемых для разных ресурсов и инструментов. Связанные данные для интеграции ЛР облегчают повторное использование общих словарей.
- В корпусной лингвистике перекрывающаяся разметка – это известная проблема для обычных форматов XML. Модели данных на основе графов были предложены с конца 1990-х гг. Они традиционно представлены в виде множества взаимосвязанных файлов XML, которые плохо поддерживаются стандартной технологией. Моделирование таких сложных аннотаций, как связанные данные, представляет собой формализм, семантически эквивалентный XML, но устраняет необходимость в специальной технологии и, вместо этого, опирается на существующую экосистему RDF.
- Многоязычные проблемы, включая связывание лингвистических ресурсов, таких как WordNet, как это выполнено в Межъязыковом указателе WordNet, и взаимосвязанные разнородные ресурсы, такие как WordNet и Wikipedia, как это было сделано в BabelNet.
- Информационное обеспечение стандартизации лингвистических ресурсов.

LLOD применяется в разработке передовых методов связывания лексических данных в Интернете, лучших методов создания аннотаций (например, с использованием стандарта Web Annotation) и лучшей практики моделирования и совместного использования текстовых ресурсов с перекрывающейся разметкой.

Классификация лингвистических данных

При формировании проекта LLOD существенным был вопрос о релевантности включаемых в него понятий. Для решения этой задачи разработаны критерии «лингвистической релевантности». Следствием этого подхода стала классификация лингвистически релевантных наборов данных. Международная рабочая группа по открытой лингвистике (OWLG) разработала следующую классификацию лингвистических ресурсов, размещаемых в LLOD:

- корпус:

лингвистически проанализированный набор языковых данных;

- лексиконы:

лексико-концептуальные данные:

  - лексические ресурсы: лексиконы и словари,
  - БД терминов: терминология, тезаурусы;
- метаданные:
  - метаданные ЛР (включая цифровые и печатные ЛР),
  - категории лингвистических данных (метаданные, включая лингвистические категории, языковые идентификаторы),
  - типологические базы данных (метаданные об отдельных языках, лингвистических особенностях этих языков);
- другое (множество ресурсов, которые (пока) не классифицированы).

В этой классификации терминологические базы находятся на грани лингвистической релевантности, поскольку они обычно создаются для целей, отличных от языковых технологий или лингвистических исследований.

Открытые данные и их доступность

Лингвистические связанные открытые данные должны соответствовать лицензиям. Однако для генерации облака LLOD это требование выполняется не всегда, поэтому техническими критериями являются доступность через Интернет и наличие метаданных. В OWLG неоднократно обсуждалось, могут ли быть включены в облако LLOD нелицензированные некоммерческие (академические) ресурсы. В 2015 г. было достигнуто согласие их включить, но с последующим введением более строгих требований вместе с ростом облака LLOD. По состоянию на январь 2020 г. было доступно 86 лицензированных ресурсов LLOD, из них 82 приняли открытые лицензии, 4 – некоммерческие лицензии.

Семинары по лингвистическим связанным открытым данным

С момента создания в 2012 г. семинары «Связанные данные в лингвистике» – Linked Data in Linguistics (LDL) стали основным форумом для представления, обсуждения и распространения технологий, словарей, ресурсов и опыта связанных открытых данных применительно к лингвистическим ресурсам. Труды 2-го, 4-го и 7-го семинаров доступны по адресу [2], 5-го семинара – по адресу [3], а 6-го – по адресу [4].

Семинары LDL, организованные OWLG, способствуют обсуждению, распространению и установлению стандартов сообщества, в первую очередь модели OntoLex-Lemon для лексических ресурсов, а также стандартов для других типов языковых ресурсов, которые все еще находятся в стадии разработки

Большинство семинаров носит тематический характер. Например, на 2-м семинаре (LDL-2013) обсуждались темы:

- Примеры создания, ведения и публикации коллекций лингвистических данных, связанных с другими ресурсами
- Моделирование лингвистических данных и метаданных с помощью OWL и/или RDF;
- Онтологии для коллекций лингвистических данных и метаданных
- Применение других онтологий или связанных данных из любой субдисциплины лингвистики

- Описания наборов данных, в идеале следующие принципам связанных данных
- Правовые и социальные аспекты лингвистических связанных открытых данных.

На 4-м семинаре (LDL-2015) обсуждались применение парадигмы связанных открытых данных к ЛР в различных областях лингвистики – это обработка естественного языка, управление знаниями и информационные технологии, а также принципы, исследования и лучшие практики представления, публикации и связывания моноязычных и многоязычных коллекций лингвистических данных, включая корпуса, грамматики, словари, словосочетания, память переводов, предметно-ориентированные онтологии и т. д.

На 5-м семинаре (LDL-2016) в центре обсуждения были вопросы применения парадигмы связанных открытых данных к лингвистическим ресурсам, поскольку это становится важным шагом, чтобы сделать лингвистические данные легко и единообразно запрашиваемыми, совместимыми и использующими открытые стандарты, такие как протокол HTTP и модель данных RDF.

Было показано, что связанные данные имеют большое значение для управления ЛР в сети, однако эта практика все еще далека от общепринятого стандарта. Поэтому важно, чтобы продолжалась разработка и внедрение технологий связанных данных среди создателей ЛР. В частности, способность связанных данных повышать качество, совместимость и доступность данных в сети побудила OWLG сосредоточить внимание на управлении, улучшении и использовании ЛР в сети в качестве ключевого направления семинара.

6-й семинар (LDL-2018), проведенный совместно с Конференцией по языковым ресурсам и оценке Conference on Language Resources and Evaluation (LREC 2018), был посвящен формированию лингвистической науки о данных, т. е. исследовательским методикам и приложениям, основанным на технологиях LLOD и интеграции лингвистических ресурсов для исследований, автоматической обработки естественного языка и цифровой гуманитаристики.

Наконец, на последнем, 7-м семинаре (LDL-2020) также совместно с LREC 2020, рассматривались вопросы создания инструментов и инфраструктуры для LLOD.

В последние годы наблюдается рост интереса к применению технологий Семантической сети к лингвистическим ресурсам и их размещению в виде связанных данных в сети. На сегодняшний день большое количество ЛР либо преобразовано, либо создано изначально как связанные данные на основе моделей, специально разработанных для представления лингвистического контента. Ведущие лингвистические центры разрабатывают различные модели взаимодействия с LLOD.

Однако, несмотря на то, что критическая масса LLOD уже существует, по-прежнему наблюдается острая потребность в надежной экосистеме инструментов, которые работают с LLOD. Недавно начатые исследовательские сети и европейские проекты, такие как Nexus Linguarum, ELEXIS и Prêt-à-LLOD, направлены на создание устойчивых инфраструктур для лингвистических ресурсов с использованием LLOD в качестве одной из основных технологий. Проекту Prêt-à-LLOD был посвящен центральный доклад на 7-м семинаре LDL-2020 [5], который мы кратко изложим далее.

РАЗВИТИЕ ПРОЕКТА LLOD

Проект CLLD – кросс-лингвистические связанные данные (Cross-Linguistic Linked Data. – URL: https://clld.org/) координирует более десятка лингвистических баз данных, охватывающих языки мира, и проводится в отделении лингвистической и культурной эволюции Института истории человечества им. Макса Планка в Йене (MPI-EVA, Тюрингия, Германия).

Цель проекта – разработать и сопровождать методы совместимости лингвистических данных с использованием принципов LLOD в качестве механизма интеграции для распределенных ресурсов.

Этот подход позволяет с минимальными затратами публиковать отдельные лингвистические ресурсы, такие как Всемирный атлас языковых структур (WALS – World Atlas of Language Structures) или Всемирная база данных заимствованных слов (WOLD – The World Loanword Database), сохраняя бренды этих проектов и в то же время обеспечивая унифицированный пользовательский интерфейс для всех ЛР.

CLLD включает лингвистические ресурсы, уже скомпилированные в MPI-EVA и в других местах. Это привело к созданию программной среды, которую можно использовать для разработки отредактированных коллекций баз данных, представленных лингвистами всего мира.

Описание методологии CLLD можно найти в работе [6]. Список баз данных, реализованных как приложения CLLD и опубликованных на платформе CLLD, доступен по ссылке Datasets https://clld.org/datasets.html. «*Dictionaria*» (https://dictionaria.clld.org/) – электронный журнал словарей редко изучаемых языков, который работает на платформе CLLD, уже опубликовал 10 словарей.

Для целей однозначной привязки лингвистических данных к языкам и каждой разновидности проект CLLD включает «Glottolog» – каталог всех языков, семейств и диалектов с исчерпывающей справочной информацией.

Формат CLDF – кросс-лингвистических данных (Cross-Linguistic Data Formats https://cldf.clld.org/). Возможно, наиболее важным результатом проекта CLLD была спецификация формата CLDF, который содержит рекомендации по хранению наборов лингвистических данных в виде взаимосвязанных текстовых файлов, упрощая долгосрочное архивирование и доступ к наборам данных через такие репозитории, как Zenodo – стандартизированный формат подачи заявок для журналов, например, «Dictionaria», а также упрощенное создание приложений CLLD из схем элементов, специфичных для модуля CLDF.

Использование наборов данных CLDF в качестве «входных» для приложений CLLD решает одну из самых больших проблем публикации данных в веб-приложении: как обрабатывать несколько версий данных? С помощью CLDF наборы данных могут быть версионными, и несколько версий могут быть

опубликованы в репозитории, в то время как веб-приложение переведено в доступный для просмотра интерфейс последней версии.

Стандартизированные форматы данных могут стать основой не только для инструментов, но и в качестве учебного материала для исторической лингвистики и лингвистической типологии.

Основными типами кросс-лингвистических данных являются любые табличные данные, которые обычно анализируются с использованием количественных (автоматизированных) методов или становятся доступными с помощью программных средств, таких как фреймворк CLLD, например: списки слов (или более сложные лексические данные), структурированные наборы данных, простые словари.

Данные должны быть редактируемыми «от руки» и поддающимися чтению и записи с помощью программного обеспечения, а также быть закодированы в виде текстовых файлов UTF-8.

Если на сущности можно ссылаться, например, на языки через их код в Glottolog, то это следует делать, а не дублировать информацию, такую как названия языков.

Автоматическое повторное использование формата требует, чтобы он определял не только структуру, но и семантику хранимых данных. Конечно, новые типы данных не могут быть немедленно совместимы с независимо разработанными инструментами, поэтому стандарт CLDF должен предоставлять механизмы, позволяющие типам данных развивать хорошо понятную семантику, будучи синтаксически совместимым с самого начала.

CLDF построен на модели W3C для табличных данных и метаданных в Интернете и словаре метаданных для табличных данных. Эта модель – в силу того, что она является диалектом JSON-LD, – идеально подходит для объединения с онтологией, с этой целью следует указывать синтаксис и семантику формата сериализации данных. CLDF структурирует кросс-лингвистические данные, чтобы сделать возможным автоматическое повторное использование.

Одной из основных целей спецификации CLDF – это разграничение данных и инструментов. Использование формата на основе CSV (Comma-Separated Values – текстовый формат для представления табличных данных) упрощает использование этих данных в процедуре их преобразования в среде UNIX. В то время как форматы для обмена лингвистическими данными существуют уже некоторое время, например, SFM для видео или стандартный формат, используемый Toolbox, новые разработки в области исследований языкового разнообразия мотивировали интерес к стандартизации табличных данных в Интернете с особым акцентом на CSV.

Опыт проекта CLLD показал, что на основе одной и той же базовой модели может быть построено множество различных кросс-лингвистических баз данных. Проект предложил идею очень простого формата CSV для обмена кросс-лингвистическими данными. Доступность была главной целью проектирования с самого начала, поэтому рассматриваемые форматы будут развиваться, начиная с максимально простых.

Модель OntoLex-Lemon (https://www.w3.org/2016/05/ontolex/) **–** возникшая в результате работы Группы сообщества W3C, первоначально была разработана с целью обеспечить полное лингвистическое обоснование онтологий. Это означает, что выражения естественного языка, используемые в метках, определениях или комментариях элементов онтологии, следует снабжать подробным лингвистическим описанием.

Онтологии являются важным компонентом семантической сети, но современные языки онтологий, такие как OWL и RDF(S), не поддерживают их обогащение лингвистической информацией, в частности о том, как объекты онтологии, т. е. свойства, классы, индивиды и т. д., могут быть реализованы на естественном языке. Модель OntoLex-Lemon направлена на то, чтобы закрыть этот пробел, предоставив словарь, который позволит онтологиям обогащаться информацией о том, как описанные в них элементы словаря реализуются лингвистически, в частности в естественных языках.

OWL и RDF(S) полагаются на свойство *RDFS:lable* для фиксации связи между словарным элементом и его (предпочтительной) лексикализацией в данном языке. Эта лексикализация обеспечивает лексический якорь, который делает класс, свойство, индивидуум и т. д. понятными для пользователя – человека. Простая метка для лингвистического обоснования, доступная в OWL и RDF(S), далеко не способна нести необходимую лингвистическую и лексическую информацию, в которой нуждаются приложения NLP, работающие с конкретной онтологией.

Цель проекта OntoLex-Lemon – обеспечить обогащенное лингвистическое обоснование онтологий, что включает в себя представление морфологических и синтаксических свойств лексических единиц, а также синтаксически-семантический интерфейс, т. е. отношение этих лексических единиц к онтологии или лексике.

Основной организующей единицей для этих лингвистических описаний является лексический класс, который обеспечивает представление морфологических паттернов для каждой записи (многозначное выражение, слово или аффикс).

Связь лексического входа с онтологической сущностью маркируется свойством денотата или опосредуется классами лексического смысла или лексико-концепта. OntoLex-Lemon включает в себя явный способ кодирования концептуальных иерархий, опирающийся на стандарт SKOS (Simple Knowledge Organization System. – URL: https://www.w3.org/TR/skos-primer). Лексические записи могут быть связаны с такими концептами SKOS, которые представляют собой синсеты WordNet. Эта структура распараллеливает отношение между лексическими записями и онтологическими источниками.

Помимо своей первоначальной области применения, модель OntoLex-Lemon стала де-факто стандартом в области цифровой лексикографии и используется, например, в европейском инфраструктурном проекте ELEXIS (European Lexico-Graphic Infrastructure. – URL: http://www.elex.is/).

Расширение сферы применения модели нашло отражение в разработке новых модулей для лексики OntoLex-Lemon. Это касается, например, лексикографии [7], спецификаций для морфологии [8], а также частотной и корпусной информации [9].

Морфологическая модель особенно важна для кросс-лингвистической применимости OntoLex-Lemon, поскольку она направлена на поддержку языков с большим количеством внутренних чередований.

В настоящее время обсуждаются технические характеристики фонологических процессов и морфологические и синтаксические комбинаторные ограничения, такие как ограничения на компаундирование и деривацию.

Интеграция российских тезаурусов в формате LLOD. Идею интеграции нескольких тезаурусов русского языка в семантическую сеть в облаке открытых лингвистических связанных данных реализовал Д.А. Усталов [10].

В этом проекте интеграции рассмотрены различные тезаурусы русского языка и установлено, что существуют четыре известных электронных тезауруса русского языка, находящиеся в доступе по открытым лицензиям: 1) RuThes-lite, 2) Русский викисловарь, 3) Универсальный сетевой язык и 4) YARN (Yet Another RussNet).

RuThes-lite – это подмножество лексической онтологии RuThes (http://www.labinform.ru/pub/ruthes/index.htm), доступное на условиях лицензии CC BY-NC-SA в виде квазиструктурированных HTML-страниц в Интернете, представляющих примерно 26 тыс. концептов и 100 тыс. межконцептных отношений.

Универсальный сетевой язык (UNL. – URL: http://www.unlweb.net/unlweb/) – проект, возглавляемый Организацией Объединенных Наций, посвященный разработке компьютерного языка, который воспроизводит функции естественных языков. Русская версия его семантической сети, – UNLDC распространяется по лицензии CC BY-SA. Она содержит примерно 62 тыс. универсальных слов (UWS) и 90 тыс. ссылок между ними.

Русский Викисловарь описан на https://ru.wiktionary.org/wiki/Заглавная_страница. Аутентичный формат страниц викисловаря – это квазиструктурированный синтаксис вики. Кроме того, существует Wikokit – проект, который анализирует русские и английские викисловари и выводит их в машиночитаемую форму реляционной базы данных, доступной на условиях лицензии CC BY-SA.

Тезаурус YARN включает в себя лексикон и синсеты русского Викисловаря. Поэтому в проекте интеграции участвуют только три ресурса: RuThes-lite, UNLDC и YARN. Интегрированный тезаурус получил название Russian Thesauri as Linked Open Data (RTLOD).

Итоговые компьютерные лексикографические ресурсы, образующие RTLOD, представляются при помощи следующих RDF-словарей: SKOS – для понятий; OntoLex-Lemon – для лексических входов, значений, определений и примеров употребления; LexInfo – для записи морфосинтаксических помет, а также RDFS, OWL и Dublin Core – для описания онтологии.

Проект PRET-a-LLOD (https://pret-a-llod.github.io/). Языковые технологии, как правило, полагаются на большие объемы данных, а качественный доступ и использование лингвистических ресурсов позволят реализовать многоязычные решения, которые будут поддерживать формирующийся единый цифровой рынок в Европе. Однако данные редко бывают готовыми к использованию, и специалисты по языковым технологиям тратят более 80% своего времени на очистку, организацию и сбор наборов данных. Снижение этих усилий обещает огромную экономию средств для всех секторов, где требуются языковые технологии. Важная часть процесса извлечения – преобразования – загрузки включает связывание наборов данных с существующими схемами, но лишь немногие специалисты используют преимущества технологий связанных данных для выполнения этой задачи.

Цель проекта – увеличить использование лингвистического ресурса с помощью LLOD для создания готовых многоязычных данных. PRET-a-LLOD стремится достичь этого путем создания новой методики и технологии создания данных, применимых к широкому спектру секторов и приложений, основанных на лингвистических ресурсах, которые могут быть интегрированы с помощью семантических технологий.

В рамках проекта разрабатываются новые инструменты для преобразования и связывания наборов данных, которые будут применяться как к данным, так и к метаданным, чтобы обеспечить доступ к разнородным репозиториям данных. Проект предполагает автоматический анализ лицензий, чтобы определить, как данные могут быть законно использованы и проданы поставщиками ЛР.

В проекте создаются инструменты для объединения языковых сервисов и ресурсов в сложные контейнеры. Это приведет к появлению устойчивых предложений данных и услуг, которые можно будет развернуть на многих платформах, включая еще неизвестные, которые могут быть самоописаны с помощью связанной семантики данных. Этот инструментарий апробируется в четырех пилотных проектах. Он увеличит распространение языковых технологий за счет устранения препятствий на пути их использования и обеспечит экономию средств, которая принесет пользу пользователям как государственного, так и частного секторов.

Основная цель проекта заключается в обеспечении многоязычного междисциплинарного доступа к лингвистическим ресурсам, используемым в многоязычных трансграничных ситуациях. Это достигается за счет предоставления инструментов обнаружения данных, основанных на метаданных, агрегированных из нескольких источников, методологий описания свойств данных и услуг, а также инструментов для вывода возможных значений ресурса, полученного после сложного контейнера. С этим связана разработка трансформационной платформы, которая отображает наборы данных в форматы и схемы, которые могут быть использованы LLOD. Наконец, проект развивает экосистему для поддержки разработки языковых технологий, основанных на связанных открытых данных, – от базовых инструментов, таких как теггеры, до полноценных приложений, таких как

системы машинного перевода, оснащенные семантическими технологиями. Существующие технологии семантического связывания применяются, чтобы обеспечить полуавтоматическую интеграцию услуг.

Устойчивость языковых технологий и ресурсов – это серьезная проблема, поэтому необходимо повысить их устойчивость, предоставляя услуги в виде данных и используя программное обеспечение с открытым исходным кодом.

Создаются также вспомогательные инструменты для измерения и анализа достоверности, ремонтопригодности и лицензирования данных и услуг. Это повышает качество и охват языковых ресурсов и технологий, гарантируя, что услуги легче архивировать и повторно использовать, позволяя им таким образом дольше оставаться доступными.

В проекте реализуются методы обнаружения, преобразования и связывания лингвистических данных так, чтобы они могли быть опубликованы как LLOD.

Обнаружение. PRET-a-LLOD предоставляет гибкую платформу обнаружения и поиска, как лингвистических ресурсов, так и сервисов. Поскольку многие реальные проблемы возможно решить только комбинацией нескольких наборов данных и сервисов, проект разрабатывает новую систему workflow, которая поддерживает цепочку нескольких сервисов с использованием семантических описаний сервисов и контейнеризации.

Полученная в результате платформа обнаружения и поиска состоит из единого и удобного для пользователя портала. Эта платформа построена поверх платформы Linghub, которая теперь импортирована на платформу CKAN (портал открытых данных), обеспечивая устойчивость и масштабируемость. CKAN – это система управления данными (DMS) с открытым исходным кодом для порталов данных (https://ckan.org/).

Трансформация. Существующие лингвистические ресурсы используют различные форматы. Для того чтобы их (повторно) применять, необходимо преодолеть структурные и концептуальные различия. PRET-a-LLOD решает эту проблему с помощью интегрированной методики, которая преобразует языковые ресурсы. Модель преобразования – это OntoLex-Lemon (кратко представленная выше) для лексических данных, поддерживающих представление языковых данных в RDF.

PRET-a-LLOD объединяет множество компонентов для трансформации, обогащения и манипулирования ЛР в рамках гибкой интегрированной платформы RDF-преобразований, получившей название *Fintan* [11].

Помимо преобразования корпусных данных, *Fintan* был расширен для преобразования лексических наборов данных в представления RDF с использованием OntoLex-Lemon. В настоящее время *Fintan* поддерживает 16 разных корпусных форматов.

Выбранные наборы данных, преобразованные в рамках проекта, включают:

- RDF-преобразование полных данных Apertium: 55 двуязычных данных,
- RDF-преобразование базы данных PanLex: 2500 словарей, скомпилированных в 1651 двуязычный словарь (т. е. те, которые содержат более 10 тыс. записей в языковой паре),
- RDF-конверсия других словарных коллекций: 252 двуязычных словаря (FreeDict, XDXF),
- RDF-конверсия инвентаризаций морфем: 110 моноязычных инвентаризаций морфем из UniMorph и 7 крупномасштабных морфологических ресурсов для 7 языков ЕС,
- RDF-преобразование WordNet: три для романских языков и один для немецкого языка,
- Преобразование пяти терминологических ресурсов из TBX в RDF.

Связывание. Проект разрабатывает полуавтоматизированные механизмы связывания. Это касается как концептуального уровня языковых описаний, так и лексических данных.

В контексте межъязыкового сопоставления концептов уже существующий инструмент сопоставления онтологий CIDER-CL (A system for monolingual and cross-lingual ontology alignment. – URL: https://oeg.fi.upm.es/files/cider-cl/) дополняется современными технологиями, основанными на межъязыковых встраиваниях слов. Лексикализация онтологий направлена на разработку методов, которые могут связать существующие онтологии с лексиконами в более широком масштабе.

Другие работы по связыванию выполняются при поддержке “Naisc”, инструмента, разработанного в Национальном университете Ирландии в Голуэе и используемого в рамках проекта Европейской лексикографической инфраструктуры (ELEXIS – URL:https://elex.is/).

Обнаружение охраняемых ЛР и доступ к ним. В рамках PRET-a-LLOD решалась проблема обнаружения и исполнения лицензионных условий для лингвистических ресурсов, объединенных в сложные контейнеры. Разрабатывались методы автоматизированного исполнения лицензионной политики для операций с ЛР. Эта работа основана на спецификациях Открытого языка цифровых прав (ODRL – Open Digital Rights Language. W3C specification. – URL: https://www.w3.org/TR/odrl-model). Поскольку все эти шаги должны быть интегрированы в рабочий процесс, PRET-a-LLOD разрабатывает основанный на семантической разметке протокол, который должен позволить языковым сервисам легко подключаться к многосерверной рабочей среде.

Практические результаты PRET-a-LLOD включают в себя четыре отраслевые пилотные проекта, которые призваны продемонстрировать актуальность, переносимость и применимость методов к практическим проблемам в индустрии языковых технологий.

Извлечение терминов и сопоставление понятий. В пилотном проекте I для компании Semantic Web было необходимо улучшить процесс извлечения терминов и сопоставления концептов, предлагаемых флагманским продуктом PoolParty. Кроме того, следует заменить некоторые проприетарные лингвистические ресурсы, используемые в настоящее время в Pool Party, на ресурсы с открытым исходным кодом.

Связывание лексических данных для облегчения интеграции лексикографических ресурсов в технологических компаниях. В пилотном проекте II для Изда-

тельства Оксфордского университета разрабатывалась методология связывания лексических данных с языковыми сервисами.

Два варианта этой задачи будут решаться в соответствующих предметных областях, а именно: связывание различных словарей (моно- или двуязычных) на уровне значения (т. е. на уровне смысла) и связывание корпусных данных со словарными смыслами посредством устранения смысловой неоднозначности слов.

Поддержка развития государственных услуг в рамках Открытого правительства как внутри страны, так и за ее пределами. В рамках пилотного проекта III для сервиса Derilinx (Сервис и хранилище для открытых данных – URL:https://derilinx.com) поставлена задача предложить инструменты и интерфейсы для интуитивного и трансграничного доступа к открытым данным с использованием естественного языка, т. е. веб-приложение, предоставляющее ответы на запросы, касающиеся информации о государственных услугах, через информационную панель; это включает анализ пользовательских запросов на естественном языке, преобразование их в формальные запросы к порталу здравоохранения и разработку чат-бота, обеспечивающего устные ответы на запросы.

Многоязычная текстовая аналитика для извлечения реальных данных в фармацевтическом секторе. В пилотном проекте IV для компании Semalytix (https://www.semalytix.com/) разработана система многоязычного обучения, поиска и анализа текстов для фармацевтической промышленности. Извлечение реальных данных требует анализа больших объемов разнородного контента, включая субъективные оценки пациентов и медицинских экспертов, которые обычно доступны в виде неструктурированного текста на нескольких языках. При разработке специфичных для предметной области многоязычных текстовых аналитических приложений необходимы методы, поддерживающие генерацию фактических данных, взаимодействие между ресурсами LLOD и архитектурами глубокого машинного обучения.

СОТРУДНИЧЕСТВО И ВЗАИМОДЕЙСТВИЕ СООБЩЕСТВА LLOD С ЕВРОПЕЙСКИМИ ЛЕКСИКОГРАФИЧЕСКИМИ ПРОЕКТАМИ

Взаимодействие сообщества LLOD с ELEXIS актуально в силу того, что связанные данные играют все большую роль в цифровой лексикографии, а модель OntoLex-Lemon находится в центре PRET-a-LLOD. Связь с ELEXIS важна для PRET-a-LLOD, поскольку все большее сообщество лексикографов использует OntoLex-Lemon и другие технологии LLOD, обеспечивая тем самым устойчивость методов, разработанных в рамках PRET-a-LLOD.

Сотрудничество с ELEXIS связано с повышением совместимости стандартов, например, с установлением мостов между OntoLex-Lemon, как результатом работы Группы сообщества W3C, и руководящими принципами кодирования TEI Lex-0 в рамках развития сообщества TEI. Важно определить, какой дефакто стандарт лучше всего подходит для разных аспектов цифровой лексикографии.

Еще одна существенная связь была установлена с проектом Европейской языковой сети (ELG – European Language Grid. – URL: https://www.european-language-grid.e), который стартовал одновременно с PRET-a-LLOD в январе 2019 г. Сотрудничество с ELG заключаются в том, чтобы использовать услуги LLOD на платформе ELG. Первое и успешное испытание было реализовано для выполнения преобразования из набора TBX в RDF на основе OntoLex-Lemon. Это важное достижение, поскольку оно поддерживает устойчивость результатов проекта PRET-a-LLOD, позволяя развертывать свои данные и сервисы на различных платформах, помимо основной инфраструктуры LLOD.

Наконец, следует упомянуть о влиятельной роли, которую PRET-a-LLOD сыграл в недавно созданной Европейской сети для веб-ориентированной лингвистической науки о данных (NexusLinguarum. – URL: https://nexuslinguarum.eu/) – проекта, направленного на взаимодействие лингвистов и компьютерщиков по всему Европейскому Союзу. В NexusLinguarum технологии LLOD будут играть центральную роль, и результаты PRET-a-LLOD будут необходимы для построения целостной экосистемы многоязычных и семантически совместимых лингвистических данных, которые использует NexusLinguarum.

Текущее состояние проекта PRET-a-LLOD свидетельствует о дальнейшем расширении облачной инфраструктуры LLOD и повышении устойчивости LLOD-совместимых сервисов и наборов данных. Разработчики проекта уверены, что технологии и ресурсы LLOD означают развитие устойчивой экосистемы интерактивных, веб-языковых технологических сервисов и лингвистических ресурсов в соответствии с целями Семантической сети. Благодаря проекту PRET-a-LLOD и связанным с ним инфраструктурным инициативам, эффект, ожидаемый от использования связанных данных, будет достигнут в течение ближайших лет.

Более подробная информация о современных инструментах и ресурсах PRET-a-LLOD представлена на веб-сайте проекта (https://www.pret-a-llod.eu/ software-and-resource-descriptions/).

ЗАКЛЮЧЕНИЕ

Представленные в настоящей статье проекты по развитию лингвистических связанных открытых данных (Linguistic Linked Open Data – LLOD), а это далеко не все проекты по данному направлению, а также деятельность по преобразованию разнообразных лингвистических ресурсов в форматы LLOD, убедительно свидетельствуют, что LLOD, как и платформа семантической сети в целом, стала главным и наиболее перспективным направлением развития языковых технологий и лингвистических ресурсов в том, что касается интеграции и обеспечения совместимости этих ресурсов и возможностей их повторного использования.

Нам кажется очень важным, что принципы LLOD полностью соответствуют общим принципам управления научными данными, известными в мировом научном сообществе как Принципы FAIR, т. е. научные данные должны быть видимы, доступны, совместимы и пригодны для повторного использования.

Следование этим общим принципам гарантирует полноценное участие научных результатов в общемировом научном процессе.

СПИСОК ЛИТЕРАТУРЫ

1. Cimiano Philipp, Chiarcos Christian, McCrae John P., Gracia Jorge. Linguistic Linked Data: Representation, Generation and Applications. – Springer International Publishing, 2020.
2. Workshop on Linked Data in Linguistics (LDL). – URL: https://www.aclweb.org/anthology/venues/ldl/
3. 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked. – URL: https://www.science-community.org/en/node/163285
4. 6th Workshop on Linked Data in Linguistics. – URL: https://elex.is/6th-workshop-on-linked-data-in-linguistics/
5. Recent Developments for the Linguistic Linked Open Data Infrastructure Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), P. 5660–5667 Marseille, 11–16 May 2020. – URL: https://zenodo.org/record/3934626
6. Reconciling Heterogeneous Descriptions of Language Resources / John P. McCrae, Philipp Cimiano CIT-EC, Bielefeld University Bielefeld, Germany,Victor Rodr´ıguez Doncel, Daniel Vila-Suero Jorge Gracia Universidad Politecnica de Madrid, Madrid, Spain @fi.upm.es Luca Matteis, Roberto Navigli University of Rome, La Sapienza, Rome, Italy Andrejs Abele, Gabriela Vulcu Paul Buitelaar Insight Centre, National University of Ireland Galway, Ireland. – URL: https://www.aclweb.org/anthology/W15-4205.pdf
7. Bosque-Gil J., Gracia J. (2019). The OntoLex Lemon lexicography module. Technical report, W3C Community Group Ontology-Lexica. Final Community Group Report.
8. Klimek B., McCrae J.P., Bosque-Gil J., Ionov M., Tauber J.K., Chiarcos C. (2019). Challenges for the representation of morphology in ontology lexicons. In Proceedings of eLex 2019. Electronic lexicography in the 21st century: Smart lexicography.
9. Chiarcos C., Ionov M. The OntoLex Lemon module for frequency, attestation and corpus information. Technical report, W3C Community Group Ontology-Lexica. draft version, Mar 3, 2019.
10. Усталов Д.А. Семантические сети и обработка естественного языка, Открытые системы. СУБД, Открытые системы. – 2017. – № 2. – С. 46-47.
11. Fath C., Chiarcos C., Ebbrecht B., Ionov M. (2020). Fintan – Flexible, Integrated Transformation and Annotation Engineering // In Proceedings of the Twelfth Inter-national Conference on Language Resources and Evaluation – LREC 2020 (Marseille, France, May 11-16, 2020). European Language Resources Association (ELRA).

Материал поступил в редакцию 10.06.21.

Сведения об авторе

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