https://doi.org/10.25559/SITITO.020.202402.287-298
УДК 004.82

Оригинальная статья

Методы структуризации знаний в онлайн-энциклопедиях

И. М. Кобозева[1*], Н. В. Лукашевич[1], М. Б. Панич[1], Ф. А. Садковский[1,2], Е. Н. Раевский[1]
[1] ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова», г. Москва, Российская Федерация
Адрес: 119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1
[2] ФГБУН «Институт языкознания Российской академии наук», г. Москва, Российская Федерация
Адрес: 125009, Российская Федерация, г. Москва, Большой Кисловский пер. 1, стр. 1
[*] kobozeva@list.ru

Аннотация

В настоящее время идет активная работа над онлайн-энциклопедиями на русском языке. Создается Большая российская энциклопедия (БРЭ), начато создание русскоязычной энциклопедии Рувики. В МГУ имени М.В. Ломоносова развивается проект «Ковчег знаний», который рассматривается как хранилище знаний различного формата, а также база для подготовки текстов статей для сетевого издания Большой Российской Энциклопедии.

Современные электронные энциклопедии включают сотни тысяч статей, что требует использования специальных средств структуризации знаний для улучшения эффективности поиска информации и навигации пользователя по страницам энциклопедии.

Общепринятым способом навигации между страницами в онлайн-ресурсах являются гиперссылки. В Википедии, самой большой и известной интернет-энциклопедии, имеется развитая система (иерархия) категорий и подкатегорий, которые объединяют страницы Википедии в тематические группы. Похожая категоризация используется в русскоязычном ресурсе типа Википедии – Рувики. В Большой российской энциклопедии структуризация включает несколько десятков областей знаний, список категорий, а также ключевые слова.

В системе «Ковчег знаний» МГУ предполагается использование для структуризации знания онтологий, т.е. формализованных описаний предметных областей в виде системы классов и отношений между ними. Онтологии могут улучшить эффективность поиска информации, как для пользователей, так и для использования в автоматической обработке программными агентами.

В статье рассматриваются особенности каждого способа структуризации знаний. В частности, будут рассмотрены проблемы обширной системы категорий Википедии, создаваемой пользователями, которые отражают типичные сложности при формализации описания знаний. Проблемы использования онтологий для категоризации знаний будут рассмотрены на примере Викиданных.

Ключевые слова: онлайн-энциклопедии, структуризация знаний, представление знаний, онтологии

Финансирование: Исследование выполнено при финансовой поддержке Междисциплинарных научно-образовательных школ Московского университета (грант № 23-Ш05-11).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Методы структуризации знаний в онлайн-энциклопедиях / И. М. Кобозева [и др.] // Современные информационные технологии и ИТ-образование. 2024. Т. 20, № 2. С. 287-298. https://doi.org/10.25559/SITITO.020.202402.287-298

Original article

Methods for Structuring Knowledge in Online Encyclopedias

I. M. Kobozeva[a*], N. V. Loukachevitch[a], M. B. Panich[a], F. A. Sadkovskii[a,b], E. N. Raevsky[a]
[a] Lomonosov Moscow State University, Moscow, Russian Federation
Address: 1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation
[b] Institute of Linguistics of the Russian Academy of Sciences, Moscow, Russian Federation
Address: 1, build. 1, Bolshoy Kislovsky lane, Moscow, 125009, Russian Federation
[*] kobozeva@list.ru

Abstract

Currently, several online encyclopedias are being developed in Russian. The Great Russian Encyclopedia (GRE), the Russian-language encyclopedia Ruwiki are being created. The "Ark of Knowledge" project is being developed at Lomonosov Moscow State University, which is considered as a repository of knowledge in various formats, as well as a base for preparing articles for the Great Russian Encyclopedia Internet-portal. Modern electronic encyclopedias include hundreds of thousands of articles. This requires the use of special means of structuring knowledge to improve the efficiency of information search and user navigation through the pages of the encyclopedia.

Hyperlinks are a generally accepted method of navigation between pages in online resources. Wikipedia, the largest and most famous Internet encyclopedia, has a developed system (hierarchy) of categories and subcategories, which unite Wikipedia pages into thematic groups. Similar categorization is used in a Russian Wikipedia-like resource – Ruwiki. The Great Russian Encyclopedia uses the system of facet classification and keywords.

The Lomonosov Moscow State University "Ark of Knowledge" system assumes the use of ontologies for structuring knowledge, i.e. formalized descriptions of subject areas in the form of a system of classes and relationships between them. Ontologies are supposed to improve the efficiency of information access, both for users and for automatic processing by software agents.

The article considers the features of each approach of knowledge structuring. In particular, the problems of the extensive Wikipedia category system created by users, which reflect typical difficulties in formalizing knowledge description, will be presented. The problems of using ontologies for knowledge categorization will be considered using Wikidata as an example.

Keywords: online-encyclopedias, knowledge structuring, knowledge representation, ontologies

Funding: The study was carried out with the financial support of the Interdisciplinary Scientific and Educational Schools of Moscow University (grant No. 23-Ш05-11).

Conflict of interests: The authors declares no conflict of interest.

For citation: Kobozeva I.M., Loukachevitch N.V., Panich M.B., Sadkovskii F.A., Raevsky E.N. Methods for Structuring Knowledge in Online Encyclopedias. *Modern Information Technologies and IT-Education.* 2024;20(2):287-298. https://doi.org/10.25559/SITITO.020.202402.287-298

1. Введение

В настоящее время идет активная работа над онлайн-энциклопедиями на русском языке. Создается Большая российская энциклопедия (БРЭ)[1], начато создание русскоязычной энциклопедии Рувики[2]. В МГУ имени М.В. Ломоносова развивается проект «Ковчег знаний»[3] [1, 2], который рассматривается как хранилище знаний различного формата, а также база для подготовки текстов статей для сетевого издания Большой Российской Энциклопедии.

Современные электронные энциклопедии включают сотни тысяч статей, что требует использования специальных средств структуризации знаний для улучшения эффективности поиска информации и навигации пользователя по страницам энциклопедии. Общепринятым способом навигации между страницами в онлайн-ресурсах являются гиперссылки. В Википедии, самой большой и известной интернет-энциклопедии имеется развитая система (иерархия) категорий и подкатегорий [3], которые объединяют страницы Википедии в тематические группы. Похожая категоризация используется в русскоязычном ресурсе типа Википедии – Рувики. В Большой российской энциклопедии используется фасетная система областей знаний и категорий, а также ключевые слова.

В проекте Wikidata (Викиданные), связанным с Википедией, знания представляются в формализованном виде семантической сети понятий и отношений между ними [4-6]. В системе «Ковчег знаний» МГУ предполагается использование для структуризации знания онтологий, т.е. формализованных описаний предметных областей в виде системы классов и отношений между ними [2]. Онтологии могут улучшить эффективность поиска информации, как для пользователей, так и для использования в автоматической обработке программными агентами, в том числе для обучения систем искусственного интеллекта [7-10].

В статье мы опишем особенности каждого способа структуризации знаний. В частности, будут рассмотрены проблемы обширной системы категорий Википедии, создаваемой пользователями. Проблемы использования онтологий для категоризации знаний будут показаны на примере Викиданных.

2. Методы структуризации текстовой информации и знаний

Одним из самых давних методов структуризации информации являются категории, организованные в классификационные системы. Классификационные системы для категоризации текстов могут быть фасетными и иерархическими. Согласно ГОСТу 7.74-96, иерархическая классификационная структура — это классификационная структура, основанная на отношениях подчинения. Рубрикатор – это классификационная таблица иерархической классификации, содержащая полный перечень включенных в систему классов и предназначенная для систематизации информационных фондов, массивов и изданий, а также для поиска в них[4]. Таким образом, цель создания рубрикаторов — обеспечить логически цельную систему категорий, обеспечивающую максимально полное покрытие имеющихся в системе ресурсов, документов [11, 12]. Для структуризации знаний в библиотеках обычно используются рубрикаторы, такие как УДК (Универсальная десятичная классификация), ББК (Библиотечно-библиографическая классификация), ГРНТИ (Государственный рубрикатор научно-технической информации).

Фасетная классификация — это классификационная структура, основанная на делении классифицируемого множества по нескольким классификационным признакам одновременно [8]. Так, новостное сообщение может классифицироваться как по основной теме, так и по региону, в котором произошло событие данной новости. Используются также и смешанные формы классификационных систем.

В настоящее время в качестве средств представления знаний о мире часто используются онтологии. Онтология — это система описания знаний о мире, включающая единицы определенных типов: понятия (классы) онтологии, их экземпляры, атрибуты единиц онтологии, а также отношения между ними. Такое определение онтологий объединяет под собой большое количество форм структуризации знаний. Фактически классификационные системы, рубрикаторы являются простейшей формой онтологий, так называемые легкие онтологии (*lightweight ontologies*) [13-15], которые имеют ограниченное число типов отношений между понятиями, не имеют атрибутов и экземпляров. Для применения онтологий в автоматических программных системах существенным является использование более выразительных форм представления знаний, включающих свойства отношений, аксиомы над понятиями и отношениями, правила для осуществления логического вывода для решения целевых задач.

Одним из самых востребованных свойств (аксиом) отношений в онтологиях является свойство транзитивности, которым обладают некоторые известные типы отношений. Использование свойства транзитивности, в частности, позволяет более компактно хранить и представлять знания. Исследованию транзитивности отношений посвящено много работ. Известно, что отношение *класс-подкласс*, которое представляет собой отношение между более широким множеством сущностей и его подмножеством (например, *собаки* – это класс, а *пудели* – это подкласс) является транзитивным, а отношение *экземпляр-класс* (*Москва* – экземпляр, *город* – класс) не является транзитивным [16, 17]. Отношение *часть-целое* часто счита-

[1] Научно-образовательный портал «Большая российская энциклопедия» : офиц. сайт [Электронный ресурс] // АНО БРЭ, 2024. URL: https://bigenc.ru (дата обращения: 21.05.2024).

[2] Интернет-энциклопедия «Рувики» : офиц. сайт [Электронный ресурс] // АНО «Интернет-Энциклопедия «Рувики»», 2024. URL: https://ru.ruwiki.ru/wiki (дата обращения: 21.05.2024).

[3] Ковчег знаний МГУ : офиц. сайт [Электронный ресурс] // МГУ имени М.В.Ломоносова, 2024. URL: https://arc.msu.ru (дата обращения: 21.05.2024).

[4] ГОСТ 7.74-96 Система стандартов по информации, библиотечному и издательскому делу Информационно-поисковые языки. Термины и определения: межгосударственный стандарт: издание официальное: утвержден и введен в действие Постановлением Государственного комитета Российской Федерации по стандартизации, метрологии и сертификации от 27 ноября 1996 г. № 651: введен впервые: дата введения 1997-07-01. М.: ИПК Издательство стандартов, 1997.

ется транзитивным, однако проявляет отклонения от транзитивности в конкретных предметных областях [18, 19]. Неточное установление отношений может приводить к неправильному логическому выводу и, вследствие этого, ошибочным ответам автоматических систем.

3. Подходы к структуризации информации в онлайн энциклопедиях

3.1 Категоризация статей в Большой Российской Энциклопедии

Большая российская энциклопедия (БРЭ) — фундаментальная энциклопедия на русском языке, создаваемая экспертным сообществом российских ученых. Она была издана в 36 бумажных томах, включающих более 80 тысяч статей. В настоящее время в соответствии с решениями Правительства Российской Федерации начато создание Большой российской энциклопедии в цифровой форме. БРЭ в настоящий момент существует в двух версиях: старой (https://old.bigenc.ru) и новой (https://bigenc.ru, до 2022 г. – портал «Знания»). Старая версия представляет собой собрание оцифрованных статей, входящих в бумажные тома. В новой версии старые статьи проходят повторное редактирование, чтобы соответствовать последнему слову науки, а также добавляются новые статьи.

В БРЭ для структуризации статей используется каталог [20], который представляет собой трехкомпонентную фасетную систему классификации энциклопедических статей. В состав каталога БРЭ входят: 1) области знаний («Геология», «Искусство» и др.); 2) категории типов объектов (Астрономы, Горные хребты, Восстания и т.д.); 3) теги (ключевые слова). Данный каталог в 2024 году содержит около 40 областей знаний, более 200 категорий типов объектов и теги, которые можно приписать к статьям без ограничений.

На портале БРЭ реализован поиск по перечисленным компонентам каталога. Каждая статья приписана к некоторой категории типа объекта, которая дает возможность перемещаться по порталу и находить статьи той же категории. При этом навигация по областям знаний по тому же принципу недоступна: приписываемая статье область знаний не является ссылкой. В целом, представленная фасетная система дает дополнительные возможности для нахождения в энциклопедии нужной информации. Вместе с тем в системе каталога БРЭ есть особенности, которые усложняют индексацию статей элементами каталога и могут затруднить поиск информации в энциклопедии.

Система областей знаний содержит одновременно более широкие области знаний и более узкие («Искусство», «Театр»). Среди областей знания имеется область «Игры», которая по смыслу больше соответствует категориям. Области знаний и категории объектов никак не связаны друг с другом. В системе областей знаний нет таких областей, как «Экология» или «Филология» (есть «Лингвистика»), но одновременно существуют категории «География», «Социально-экономическая география» и «Страны, регионы, города». При этом во всей системе категоризации БРЭ ни разу не упоминается термин *компьютер*. Видимо, состав областей знаний и категорий определяется прежде всего наполнением БРЭ статьями по соответствующей области, а не по заранее предопределенным принципам. Среди тегов имеется достаточно много таких, которые относятся только к одной статье, что увеличивает количество тегов и не дает дополнительные возможности для поиска информации.

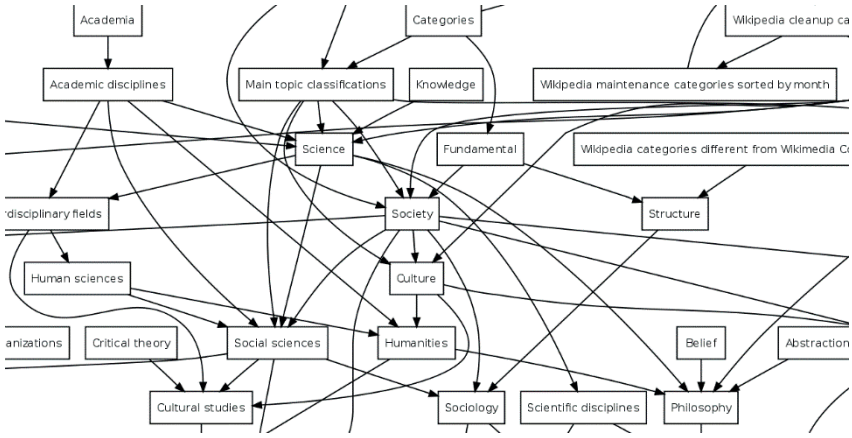
3.2 Категоризация статей в Википедии

В Википедии, самой большой онлайн энциклопедии, используется иерархическая система категорий, которая дает перемещаться в поиске информации от более широких категорий к более узким и наоборот. Правила создания и именования категорий прописаны в специальной инструкции[5]. Категории подразделяются на тематические (*topic categories*) и категории-множества (*set categories*). Тематические категории соответствуют отдельному понятию или конкретной сущности и именуются существительным или именной группой в единственном числе, например, *Закон*, *Франция*, и др. Категории-множества содержат перечни объектов одного типа, например, *Писатели*, *Города в Польше*, и должны иметь имена во множественном числе. Как представляется, категории-множества Википедии соответствуют категориям БРЭ. Тематические категории и категории-множества могут иметь похожие названия, различающиеся лишь грамматическим числом. Например, в тематической категории Опера собраны различные понятия, относящиеся к этому виду искусства, а в категории Оперы собраны статьи, посвященные конкретным операм.

Каждая страница Википедии должна принадлежать как минимум к одной категории, за исключением служебных страниц (страниц обсуждений, перенаправлений и др.). Каждая категоризированная страница должна быть помещена во все наиболее конкретные категории, к которым она логически принадлежит. Это означает, что если страница принадлежит к подкатегории *C* (или подкатегории подкатегории C и т. д.), то она не должна помещаться непосредственно в категорию *C*, а в наиболее подходящую подкатегорию.

В результате система категорий в Википедии представляет собой иерархическую структуру в виде дерева. При этом подкатегории могут иметь несколько родительских категорий (Рис. 1). Имеется одна категория верхнего уровня, называемая *Category:Contents*. Согласно правилам, результирующие цепочки категорий от более общих к более частным не должны образовать циклы. При добавлении подкатегории в категорию рекомендуется убедиться, что все потенциальные подкатегории данной подкатегории также относятся к данной родительской категории, т.е. предполагается, что вложение подкатегорий в родительские категории должно быть транзитивно: если *A* является подкатегорией *B*, и *B* является подкатегорией *C*, то *A* является подкатегорией *C*.

[5] Naming conventions : офиц. сайт [Электронный ресурс] // Wikipedia, 2024. URL: https://en.wikipedia.org/wiki/Wikipedia:Categorization#Naming_conventions (дата обращения: 21.05.2024).



Р и с. 1. Визуализация фрагмента системы категорий Википедии[6]

F i g. 1. Partial view of Wikipedia's category system from 2007. Arrows point from category to subcategory[6]

Для улучшения качества системы категорий имеется также инструкция по типовым случаям, когда нужно избегать образования новых категорий[7] (так называемые случаи чрезмерной категоризации – *overcategorization*). В частности, указывается, что категории должны создаваться для существенных, определяющих (*defining*) характеристик объектов, а не для периферийных, например, категории по высказанным мнениям, вкусам в еде, количеству татуировок и т.п.

Таким образом, система категорий Википедии обеспечивает полное покрытие статей подходящими категориями. Количество категорий достигает нескольких сот тысяч. Иерархия категорий дает значимые возможности для навигации по энциклопедии для поиска нужной информации. В то же время большая свобода в создании категорий и установлении связей между категориями иногда приводит к существенным нарушениям провозглашенной иерархической структуры категорий, несмотря на имеющиеся инструкции. В частности, в системе категорий имеются многочисленные циклы [21, 22] а также нарушения транзитивности подчинения категорий [23].

Нарушение транзитивности проявляется, например, в задаче сбора терминов или статей по некоторой тематике или для некоторой предметной области по системе категорий Википедии [19-21]. При такой постановке задачи выбираются категории Википедии, представляющие область в целом (например, химия или общественное питание), а затем собираются подкатегории в нижележащей иерархии категорий. Нарушение транзитивности приводит к тому, что в процессе такой процедуры собирается большое количество нерелевантных тематике категорий и статей. Так, авторы работы [21] указывают, что по тематике *Химия* (*Chemistry*) для английского языка было собрано более 188 тысяч подкатегорий, среди них такие как *Лексическая семантика*, кроме того было обнаружено 247 циклов. Программное обеспечение Википедии не отслеживает появление циклов, а также нарушения транзитивности в системе категорий.

Проблемы с транзитивностью часто связаны с фундаментальными проблемами создания больших онтологий, поскольку в основе таких онтологий обычно лежит большая иерархия (таксономия) понятий. Эти проблемы будут рассмотрены в следующем разделе.

3.3 Проблемы поддержки качества большой иерархической системы понятий на примере системы категорий Википедии

Отношения подчинения между категориями в Википедии объединяют разные типы семантических отношений. В частности, среди отношений между подкатегориями и категориями часто встречаются такие типы отношений [23]:

- Отношение *класс-подкласс*: населенные пункты России → города России, системное программное обеспечение → операционная система;
- Отношение *класс-экземпляр*: города Англии → Лондон, операционные системы → UNIX;
- Отношение *часть-целое*: Москва → достопримечательности Москвы, голова → мозг;
- Разнообразные типы ассоциативных отношений: наука → объект науки (ботаника → растения), агент→ контрагент (Растения → Гербициды), сырье → продукт (виноград → изюм) и др.

Отметим, что само по себе смешение разных типов отношений в иерархии категорий может привести к проблемам с транзитивностью. Кроме того, имеются типовые неточности описаний, которые приводят к нарушению транзитивности в системе категорий. Рассмотрим наиболее распространенные ошибки на конкретных примерах появления в категориях не-

[6] Wikipedia:Categorization : офиц. сайт [Электронный ресурс] // Wikipedia, 2024. URL: https://en.wikipedia.org/wiki/Wikipedia:Categorization (дата обращения: 21.05.2024).

[7] Wikipedia:Overcategorization : офиц. сайт [Электронный ресурс] // Wikipedia, 2024. URL: https://en.wikipedia.org/wiki/Wikipedia:Overcategorization# (дата обращения: 21.05.2024).

релевантных подкатегорий или статей, к чему приводит нарушение транзитивности отношений подчинения:

Классификация понятия по лексико-синтаксической структуре, а не по смыслу:

В следующем примере Электрический стул отнесен к категориям Мебель и Потребительские товары, но все-таки по смыслу – это орудие казни, напоминающее стул, и лучше вообще не относить данный предмет к категории Стулья, поскольку сходство со стулом менее существенно, чем основное назначение.

Потребительские товары → Предметы быта → Мебель → Мебельные изделия → Мебель для сидения (лежания) → Стулья → Электрический стул

Использование одной и той же категории в разных смыслах:

В следующем примере архив научных публикаций Sci-hub относится к категории Учреждения культуры. Здесь проблема связана с совмещением разных значений слова *библиотека* (библиотека как учреждение и электронная библиотека) в одной категории Библиотеки:

Учреждения культуры → Библиотеки → Электронные библиотеки → Sci-hub.

В следующем примере мусульманское религиозное сооружение Kaaba оказывается в англоязычной системе категорий в категории Геометрия, что также связано с совмещением двух разных значений в слове Cub (куб как форма и куб как предмет), а также отнесение к категории по несущественному признаку (в русскоязычной системе категорий такой проблемы нет):

Geometry → Geometric Objects → Geometric shapes → Elementary shapes → Cubes → Cubic buildings → Kaaba

Смешение понятия и знака

Категория House_of_Habsburg (Дом Габсбургов) в английской системе категорий является подкатегорией категории Words_and_phrases (Слова и словосочетания), что связано с проблемным переходом от категорий, связанных с семьями людей к фамилиям этих людей.

Words_and_phrases → … → Surnames_of_Swiss_origin → Swiss_families → Swiss_noble_families →House_of_Habsburg

В другом примере той же неточности токсины относятся к категории Язык в англоязычных категориях, поскольку токсины трактуются как вещество и как слово

Language → Terminology → Biology_terminology → Toxin

Еще раз подчеркнем, что такие неточности описаний являются типичными при поддержке больших онтологических ресурсов [16, 17], и, возможно, лишь усиливаются коллаборативным подходом к пополнению категорий Википедии и отсутствием соответствующего программного контроля.

3.4 Использование онтологий для представления знаний в онлайн-энциклопедиях

Взаимодействие онлайн энциклопедии и формальной онтологии можно видеть на примере двух связанных между собой ресурсов Википедии и Викиданных. Проект Викиданные стартовал в 2012 году и позиционируется как «Википедия для данных», т.е. исходно в Викиданные были извлечены из Википедии структурированные данные разных типов: числовые данные, даты, разные типы отношений между сущностями [3]. В настоящее время Викиданные представляют собой самую большую открытую базу знаний. 13 марта 2024 года проект «Викиданные» объявил о внесении в базу 2,100,000,000-ой сущности.

Викиданные – это совокупность сущностей. Каждая сущность может быть объектом (*item*) или отношением (свойство, *property*) и имеет отдельную страницу в ресурсе. Для всех содержательных (неслужебных) страниц Википедии имеется соответствующая страница (*item*) в Викиданных. Дополнительно пользователи могут добавить новые страницы для сущностей в Викиданных.

Каждый объект (*item*) в общем случае имеет уникальный идентификатор (начинается с буквы *Q*), основное название (*label*), описание, и совокупности вариативных текстовых именований на различных языках. Таким образом, база Викиданных является многоязычной. Свойство (*Property*) в Викиданных обозначает отношения между объектами и/или значениями. Свойство также имеет уникальный идентификатор (начинается с буквы *P*). Например, свойство «основан кем» (*founded by*) имеет идентификатор *P112*. Свойства также имеют основное название, описание и варианты его упоминания на разных языках.

Утверждения об отношениях между сущностями описаны как триплеты, например (*Рим, население,* 2777979). Утверждения также могут содержать квалификаторы, например, указание на время, к которому относится утверждение, источник информации и др.

Хотя в Викиданных было значительно больше усилий уделено аккуратному представлению структурированных данных, тем не менее исследователи обнаруживают достаточно много неточностей и ошибок в описаниях [5, 24]. В частности, есть классы, которые трудно отличить друг от друга, например, *geographical location* (Q2221906), *location* (Q115095765), *geographic region* (Q82794), *physical location* (Q17334923), *geographical area* (Q3622002)). Классы и экземпляры путаются, также смешиваются отношения *экземпляр-класс* и *класс-подкласс*. В частности, в работе [5] исследуется история изменений отношений в Викиданных. Было обнаружено, что имеется около 500 тысяч отношений, в которых эти два отношения были изменены друг на друга, т.е. сообщество авторов Викиданных имеет проблемы с четким различением этих базовых отношений описания внешнего мира.

Кроме того, таксономия Викиданных содержит циклы: более чем 47 пар классов являются подклассами друг друга, например *method* (Q1799072) и t*echnique* (Q2695280), имеется 15 циклов длины 3 и больше, например, *axiom* (Q17736), *first principle* (Q536351), *principle* (Q211364)).

Таким образом, в ресурсе Викиданные мы также видим проблемы с поддержкой качественных описаний, схожие с проблемами системы категорий Википедии. В работе авторов Викиданных [25], посвященной десятилетию существования ресурса, указывается, что в Викиданных была принята гибкая схема для представления знаний, что, с одной стороны, привело к возможности описывать разнообразные виды информации, с другой стороны, привело к проблемам неоднородности и несогласованности данных, которые еще предстоит решить. Также отметим, что в настоящее время Викиданные не используются для навигации по статьям Википедии, хотя ссылки между единицами этих ресурсов имеются.

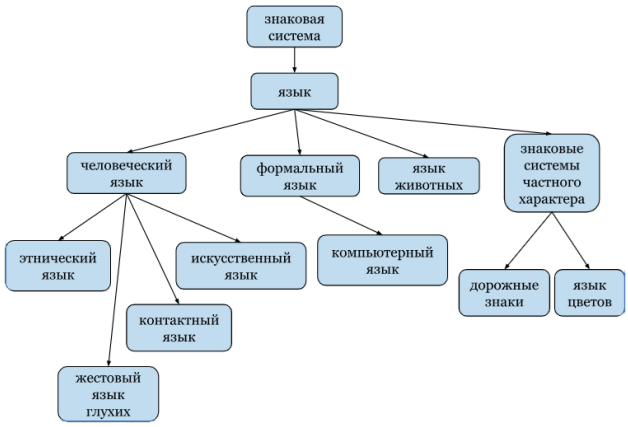
4. Представление понятия ЯЗЫК в разных энциклопедиях

В рамках проекта «Ковчег знаний» [1, 2] изначально был принят метод структуризации научных знаний, аккумулируемых в данном ресурсе, в виде онтологии. Естественным образом возник вопрос: можно ли использовать в этих целях уже имеющиеся иерархии категорий Википедии или онтологию Викиданных? С этой целью мы исследовали, как представлено в них понятие ЯЗЫК, центральное для научной области «Лингвистика», но играющее при этом важную роль в целом ряде других наук, как гуманитарных, так и точных.

4.1 ЯЗЫК в системе категорий и в статье «Язык» Википедии

Иерархия категорий Википедии в интересующей нас области обнаружила свою непригодность уже на самых верхних уровнях. Согласно этой иерархии ЯЗЫК оказывается подкатегорией ЗНАКОВЫХ СИСТЕМ, что верно, но ЗНАКОВЫЕ системы оказываются подкатегорией СЕМАНТИКИ, а это уже в корне неверно, так как семантика — это не род знаковых систем, а один из разделов, т.е. частей ЛИНГВИСТИКИ. Что касается подкатегорий категории ЯЗЫК, то среди них на одном иерархическом уровне оказались и ЯЗЫКИ, и, например, ДОКУМЕНТАЛЬНЫЕ ФИЛЬМЫ О ЯЗЫКЕ, которые очевидным образом не являются видами одного и того же рода сущностей.

Поскольку одним из основных методов создания онтологий предметных областей является анализ текстов, содержащих знания об этой области (энциклопедий, справочников, учебников и т.п.), который пока не автоматизирован, нами была дополнительно проведена проверка того, можно ли извлечь адекватный фрагмент онтологии лингвистики путем концептуального анализа содержания статьи «Язык» в Википедии. Полученный нами результат отражает схема на Рис. 2:



Р и с. 2. Онтология понятия ЯЗЫК по статье «Язык» в Википедии
F i g. 2. The ontology of the concept LANGUAGE according to the article "Language" in Wikipedia

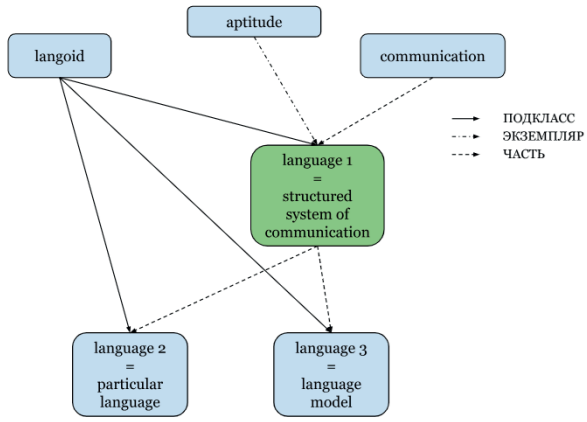
Источник: здесь и далее в статье все рисунки составлены авторами.
Source: Hereinafter in this article all figures were drawn up by the authors.

Извлеченная из указанной статьи онтология не отражает адекватно структуру знаний о языке: категория (= понятие) ЗНАКОВАЯ СИСТЕМА становится лишней, так как по объему полностью совпадает с категорией ЯЗЫК (в широком понимании). Здесь мы сталкиваемся с явлением многозначности терминов. Термин *язык* в языке науки используется для выражения разных понятий (= категорий). Основное выражаемое им понятие — ЕСТЕСТВЕННЫЙ, или ЧЕЛОВЕЧЕСКИЙ ЯЗЫК, но термин *язык* используется и для обозначения других знаковых систем, в том числе таких, как язык математики, язык цветов или язык марок. При создании онтологии необходимо строго различать слова и стоящие за ними понятия. Категория ЯЗЫК в том широком понимании, которое упоминается в статье, в онтологии не нужна. Достаточно категории ЗНАКОВАЯ СИСТЕМА. Кроме того, как мы увидим дальше, в онтологии ЕСТЕСТВЕННОГО ЯЗЫКА как объекта лингвистики не учтены некоторые важнейшие для его понимания категории.

Таким образом, проведенная проверка показала, что при построении онтологии Ковчега нельзя безоговорочно полагаться ни на категориальную иерархию Википедии, ни на содержание ее статей. Рассмотрим далее, как структурированы знания о языке в онтологии Викиданных.

4.2 ЯЗЫК в зеркале Викиданных

Ядро фрагмента онтологии, структурирующего знания о языке, представлено на Рис. 3.



Р и с. 3. Онтология ЯЗЫКА согласно Викиданным
F i g. 3. The ontology of LANGUAGE according to Wikidata

Здесь картина уже ближе к представлениям современной лингвистики. Во-первых, эксплицитно различены три разных понятия (= категории), выражаемые словом *language* ‘язык’. Во-вторых, вводятся различные отношения между понятиями: класс-подкласс, класс-экземпляр и часть-целое и многие другие. Так, понятие ЯЗЫК 1, определяется как “структурированная система передачи информации”, и тем самым эквивалентно понятию “коммуникативная система”. При этом оно представлено как экземпляр СПОСОБНОСТИ ‘aptitude’ (что отражает концепцию естественного человеческого языка в парадигме генеративизма, доминирующей в западной лингвисти-

ке[8]) и как часть процесса КОММУНИКАЦИИ ‘communication’. Однако в целом родовидовая иерархия, образуемая категориями, находящимися в отношении класс-подкласс, не выдерживает критики. Так, если ЯЗЫК 2 (конкретный, этнический язык) можно считать подклассом (= подкатегорией) ЯЗЫКА 1 (языка вообще, абстракции от всех возможных языков), то такие понятия, как ЯЗЫКИ КНР или КЛАССИЧЕСКИЙ ЕГИПЕТСКИЙ должны были бы фигурировать соответственно среди подклассов или экземпляров категории ЯЗЫК 2, определяемой как “конкретная система передачи информации”. Но, как показывает схема на Рис. 4 мы обнаруживаем их среди подклассов категории ЯЗЫК 1, наряду с такими категориями лингвистики, как МАКАРОНИЗМ, НЕПАРЛАМЕНТСКИЕ ВЫРАЖЕНИЯ или МОЛОДЁЖНЫЙ СЛЭНГ, которые не являются непосредственными подклассами ни ЯЗЫКА 1, ни ЯЗЫКА 2, поскольку выражают понятия не о языке вообще или конкретных классах языков, а о частях языков, выделяемых по разным основаниям.

Р и с. 4. Подкатегории категории ЯЗЫК I в Викиданных
F i g. 4. Subcategories of the LANGUAGE I category in Wikidata

Не меньше вопросов вызывают и партонимические отношения (целое-часть) категории ЯЗЫК 1 с другими категориями в Викиданных. Например, рядоположными частями ЯЗЫКА 1 оказываются и ЯЗЫК 3 (= МОДЕЛЬ ЯЗЫКА), и СОГЛАСНЫЕ ЗВУКИ. Это все равно, что сказать, что частями системы электроснабжения страны являются модель этой системы и выключатели в домах.

Понятие ЯЗЫК 3 определяется в Викиданных как абстрактная система правил и конвенций, независимая от конкретных говорящих, находящаяся в отношении часть-целое к понятию ЯЗЫК 1. В описании свойств ЯЗЫКА 3 фигурирует его противопоставленность понятию, обозначаемому при помощи французского термина PAROLE. Это важнейшее понятие лингвистической теории, парное к ЯЗЫКУ 3 как абстрактной системе, в отечественной лингвистике обозначается термином *речь*. Однако, как таковая, самостоятельная категория PAROLE в онтологию ЯЗЫКА 1 Викиданных не входит.

Таким образом, и онтология Викиданных в общем случае не может целиком и полностью заимствоваться в качестве средства структурирования научного знания. При желании использовать этот электронный ресурс придется ее корректировать, иногда весьма существенно.

4.3 Построение онтология ЯЗЫКА на основе энциклопедических статей

Как уже было сказано, в БРЭ есть категории, но они не выстроены в родовидовые иерархии, зато БРЭ — это кладезь статей, написанных экспертами соответствующих научных областей. На их основе можно как корректировать онтологию Викиданных, так и строить онтологию предметной области с нуля. Так, на основе статьи чл.-корр. РАН Н.Д. Арутюновой “Язык” в БРЭ можно построить онтологию, которая будет гораздо более полно и точно отражать структуру лингвистического знания, чем та, которую можно извлечь из Википедии или найти в Викиданных.

Еще более точную и объемную картину можно найти в статье “Язык” чл.-корр. РАН А.Е. Кибрика в Лингвистическом энциклопедическом словаре (ЛЭС)[9], воспроизведенной в учебнике[10] для студентов, обучающихся по специальности “Фундаментальная и прикладная лингвистика”. Процитируем в минимально сокращенном виде два первых абзаца из этой статьи:

“Язык — основной объект изучения языкознания. Под языком прежде всего имеется в виду естественный человеческий язык (в оппозиции к искусственным языкам и языку животных), возникновение и существование которого неразрывно связано с возникновением и существованием человека — homo sapiens.

Термин “язык” имеет по крайней мере два взаимосвязанных значения: язык вообще, язык как определенный класс знаковых систем; 2) конкретный, так называемый этнический. ... язык ... Язык в первом значении — это абстрактное представление о едином человеческом языке, средоточии универсальных свойств всех конкретных языков. Конкретные языки — это многочисленные реализации свойств языка вообще”.

Из первого ясно, что понятие ЕСТЕСТВЕННЫЙ ЯЗЫК связано с понятие ЧЕЛОВЕК отношением типа ПРЕРЕКВИЗИТ, т.е. необходимое условие возникновения и существования. Это отношение создает необходимый мостик между онтологией лингвистики и онтологиями целого ряда других наук о человеке.

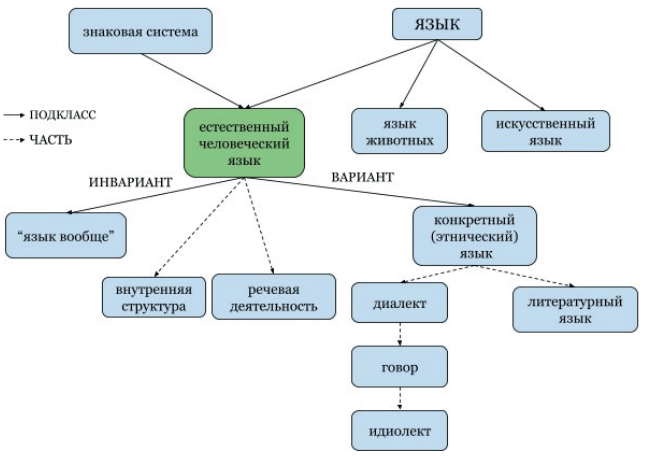
Из второго следует что в онтологии лингвистики необходимы два разных понятия ЯЗЫК — язык вообще, Язык с большой буквы, средоточие общих принципов устройства всех человеческих языков, и язык этнический, т.е. конкретный естественный язык.

Верхние уровни онтологии языка, извлекаемой из этой статьи, представлены в виде схемы на Рис. 5.

[8] Бейлин Дж. Краткая история генеративной грамматики / Под ред. А. А. Кибрик, И. М. Кобозева, И. А. Секерина // Современная американская лингвистика: Фундаментальные направления. М.: URSS. 2016. С. 13-57.

[9] Лингвистический энциклопедический словарь / Гл. ред. В. Н. Ярцева. М.: Советская энциклопедия, 1990. 683 с.

[10] Кибрик А. Е. Язык // Введение в науку о языке / А. А. Кибрик [и др.] ; под ред. О. В. Федоровой и С. Г. Татевосова. М.: Буки Веди, 2019. С. 15-23.



Р и с. 5. Верхние уровни онтологии понятийной области ЯЗЫК в ЛЭС
F i g. 5. The upper levels of the ontology of the conceptual domain of LANGUAGE in the Linguistic Encyclopedic Dictionary

Как мы видим, в этой сети понятий находит свое место и противопоставление языка вообще этническому языку, которые связаны с понятием Естественный язык (ЕЯ) отношением Инвариант и Вариант соответственно, и неразрывная связь в понятии ЕЯ двух его сторон: языка как системы (внутренней структуры) и языка как речевой деятельности.

5. Заключение

В данной статье были рассмотрены подходы к структуризации знаний в онлайн-энциклопедиях: Большой Российской Энциклопедии (БРЭ), Википедии, «энциклопедии данных» Wikidata (Викиданные), проекте «Ковчег знаний» МГУ имени М.В. Ломоносова. В БРЭ используются система фасетной классификации. В Википедии создана большая иерархическая система категорий. Викиданные – это самая большая открытая база знаний на основе онтологий, единицы Викиданных имеют ссылки на статьи Википедии, но отношения в Викиданных не используются для навигации и поиска информации в энциклопедии. В проекте "Ковчег знаний" обсуждается сочетание статей энциклопедии для людей и формальной системы для представления знаний (онтологии). Представленный анализ показал, что структуризация онлайн-энциклопедии (для людей) с помощью полномасштабной формальной онтологической системы (для программных агентов) требует аккуратного позиционирования и планирования. Профессионалы предметной области могут тщательно рассмотреть некоторые базовые понятия (например, Язык), но выстраивать и выверять полную иерархию и структуру отношений предметной области должны инженеры по знаниям (онтологи), поскольку при построении иерархии понятий «наивные онтологи» (включая профессионалов предметной области) могут совершать ошибки, портящие предполагаемую транзитивность иерархической структуры. Кроме того, нужны автоматические инструменты для контроля качества описаний.

Список использованных источников

[1] Цифровой ковчег знаний / В. В. Горячко, А. С. Бубнов, Е. В. Раевский, А. Л. Семенов // Доклады Российской академии наук. Математика, информатика, процессы управления. 2022. Т. 508, № 1. С. 128-133. https://doi.org/10.31857/S2686954322070098
[2] Универсальная энциклопедическая платформа работы со знанием / А. Л. Семенов [и др.] // Современные информационные технологии и ИТ-образование. 2023. Т. 19, № 3. С. 696-703. https://doi.org/10.25559/SITITO.019.202303.696-703
[3] Thornton K., McDonald D. W. Tagging Wikipedia: collaboratively creating a category system // Proceedings of the 2012 ACM International Conference on Supporting Group Work (GROUP '12). New York, NY, USA: Association for Computing Machinery, 2012. P. 219-228. https://doi.org/10.1145/2389176.2389210
[4] Vrandečić D., Krötzsch M. Wikidata: a free collaborative knowledgebase // Communications of the ACM. 2014. T. 57, No. 10. P. 78-85. https://doi.org/10.1145/2629489
[5] A study of the quality of Wikidata / K. Shenoy [et al.] // Journal of Web Semantics. 2022. Vol. 72. Article number: 100679. https://doi.org/10.1016/j.websem.2021.100679
[6] Wikidata: A large-scale collaborative ontological medical database / H. Turki [et al.] // Journal of biomedical informatics. 2019. Vol. 99. Article number: 103292. https://doi.org/10.1016/j.jbi.2019.103292
[7] Faber P., León-Araúz P. From specialized knowledge frames to linguistically based ontologies // Applied Ontology. 2024. Vol. 19, No. 3. P. 1-23. https://doi.org/10.3233/AO-230033
[8] Hohenecker P., Lukasiewicz T. Ontology reasoning with deep neural networks // Journal of Artificial Intelligence Research. 2020. Vol. 68. P. 503-540. https://doi.org/10.1613/jair.1.11661
[9] A comprehensive review of ontologies in the hydrology towards guiding next generation artificial intelligence applications / Ö. Baydaroğlu [et al.] // Journal of Environmental Informatics. 2023. Vol. 42, No. 2. P. 90-107. https://doi.org/10.3808/jei.202300500
[10] Methods and tools for ontology building, learning and integration–application in the synat project / A. Wróblewska [et al.] // Intelligent tools for building a scientific information platform. 2012. P. 121-151. https://doi.org/10.1007/978-3-642-24809-2_9
[11] Лукашевич Н. В. Тезаурусы в задачах информационного поиска. М.: Изд-во Московского университета, 2011. 512 с. EDN: RBBMVR
[12] Bates M. J. How to Use Controlled Vocabularies More Effectively in Online Searching // Online. 1988. Vol. 12, No. 6. P. 45-56.

[13] Davies J. Lightweight Ontologies // Theory and Applications of Ontology: Computer Applications ; ed. by R. Poli, M. Healy, A. Kameas. Dordrecht: Springer, 2010. P. 197-229. https://doi.org/10.1007/978-90-481-8847-5_9
[14] SOSA: A lightweight ontology for sensors, observations, samples, and actuators / K. Janowicz [et al.] // Journal of Web Semantics. 2019. Vol. 56. P. 1-10. https://doi.org/10.1016/j.websem.2018.06.003
[15] A lightweight ontology for landmarks to assist rescue in mountainous areas / A. M. Olteanu-Raimond [et al.] // Advances in Cartography and GIScience of the ICA. 2023. Vol. 4. Article number: 15. https://doi.org/10.5194/ica-adv-4-15-2023
[16] Guarino N. Some ontological principles for designing upper level lexical resources // First International Conference on language resources & evaluation: Granada, Spain, 28-30 May 1998. European Language Resources Association, 1998. P. 527-534. https://doi.org/10.48550/arXiv.cmp-lg/9809002
[17] Loukachevitch N. Establishment of Taxonomic Relationships in Linguistic Ontologies // Knowledge Processing and Data Analysis. KPP KONT 2007. Lecture Notes in Computer Science ; ed. by K. E. Wolff, D. E. Palchunov, N. G. Zagoruiko, U. Andelfinger. Vol. 6581. Berlin, Heidelberg: Springer, 2011. P. 232-242. https://doi.org/10.1007/978-3-642-22140-8_15
[18] Romanenko E., Calvanese D., Guizzardi G. Abstracting ontology-driven conceptual models: Objects, aspects, events, and their parts // International Conference on Research Challenges in Information Science. – Cham: Springer International Publishing, 2022. – С. 372-388. doi: https://doi.org/10.1007/978-3-031-05760-1_22
[19] Лукашевич Н. В. Отношения часть-целое: теория и практика // Нейрокомпьютеры: разработка, применение. 2013. Т. 1. С. 007-012. EDN: PVRZNN
[20] Кравец С. Л. Научно-образовательный энциклопедический портал и участие в нём региональных научных энциклопедий // Вопросы энциклопедистики. 2019. № 2. С. 14-20. EDN: NPZYHS
[21] Vivaldi J., Rodríguez H. Finding Domain Terms using Wikipedia // Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association, 2010. P. 386-393. URL: https://aclanthology.org/L10-1518/ (дата обращения: 21.05.2024).
[22] Evaluation Dataset and Methodology for Extracting Application-Specific Taxonomies from the Wikipedia Knowledge Graph / G. Bordea [et al.] // Proceedings of the Twelfth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2020. P. 2341-2347. URL: https://aclanthology.org/2020.lrec-1.285/ (дата обращения: 21.05.2024).
[23] Kirillovich A., Nevzorova O. Ontological Analysis of the Wikipedia Category System // Proceedings of KEOD Conference. – 2018. – С. 356-364. doi: https://doi.org/10.5220/0006961803580366
[24] YAGO 4.5: A Large and Clean Knowledge Base with a Rich Taxonomy / F. M. Suchanek [et al.] // Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). New York, NY, USA: Association for Computing Machinery, 2024. P. 131-140. https://doi.org/10.1145/3626772.3657876
[25] Vrandečić D., Pintscher L., Krötzsch M. Wikidata: The Making Of // Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion). New York, NY, USA: Association for Computing Machinery, 2023. P. 615-624. https://doi.org/10.1145/3543873.3585579

Поступила 21.05.2024; одобрена после рецензирования 04.06.2024; принята к публикации 22.06.2024.

Об авторах:

Кобозева Ирина Михайловна, профессор кафедры теоретической и прикладной лингвистики филологического факультета, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), доктор филологических наук, **ORCID: https://orcid.org/0000-0003-4214-2878**, kobozeva@list.ru

Лукашевич Наталья Валентиновна, ведущий научный сотрудник Научно-исследовательского вычислительного центра, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), доктор технических наук, **ORCID: https://orcid.org/0000-0002-1883-4121**, louk_nat@mail.ru

Панич Мария Борисовна, магистрант кафедры теоретической и прикладной лингвистики филологического факультета, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), **ORCID: https://orcid.org/0009-0005-7400-5837,** shmak1280@gmail.com

Садковский Фёдор Алексеевич, магистрант кафедры теоретической и прикладной лингвистики филологического факультета, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1); лаборант-исследователь лаборатории изучения и сохранения малых языков, ФГБУН Институт языкознания РАН (125009, Российская Федерация, г. Москва, Большой Кисловский пер. 1, стр. 1), **ORCID: https://orcid.org/0009-0006-4502-9466,** feudor987@mail.ru

Раевский Евгений Наумович, ведущий программист лаборатории программного обеспечения вычислительного практикума факультета вычислительной математики и кибернетики, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), **ORCID: https://orcid.org/0000-0002-0119-6418**, raev@cs.msu.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

References

[1] Goryachko V.V., Bubnov A.S., Rayevskii E.V., Semenov A.L. Digital Ark of Knowledge. *Doklady Mathematics*. 2022;106(Suppl 1):S113-S117. https://doi.org/10.1134/S1064562422060096

[2] Semenov A.L., Raevskij E.N., Bubnov A.S., Grishin I.Yu., Gulyaev A.V., Kobozeva I.M. Universal Encyclopedic Platform for Working with Knowledge. *Modern Information Technologies and IT-Education*. 2023;19(3):696-703. (In Russ., abstract in Eng.) https://doi.org/10.25559/SITITO.019.202303.696-703

[3] Thornton K., McDonald D.W. Tagging Wikipedia: collaboratively creating a category system. In: Proceedings of the 2012 ACM International Conference on Supporting Group Work (GROUP '12). New York, NY, USA: Association for Computing Machinery; 2012. p. 219-228. https://doi.org/10.1145/2389176.2389210

[4] Vrandečić D., Krötzsch M. Wikidata: a free collaborative knowledgebase. *Communications of the ACM.* 2014;57(10):78-85. https://doi.org/10.1145/2629489

[5] Shenoy K., et al. A study of the quality of Wikidata. *Journal of Web Semantics*. 2022;72:100679. https://doi.org/10.1016/j.websem.2021.100679

[6] Turki H., et al. Wikidata: A large-scale collaborative ontological medical database. *Journal of biomedical informatics*. 2019;99:103292. https://doi.org/10.1016/j.jbi.2019.103292

[7] Faber P., León-Araúz P. From specialized knowledge frames to linguistically based ontologies. *Applied Ontology*. 2024;19(3):1-23. https://doi.org/10.3233/AO-230033

[8] Hohenecker P., Lukasiewicz T. Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research.* 2020;68:503-540. https://doi.org/10.1613/jair.1.11661

[9] Baydaroğlu Ö., et al. A comprehensive review of ontologies in the hydrology towards guiding next generation artificial intelligence applications. *Journal of Environmental Informatics*. 2023;42(2):90-107. https://doi.org/10.3808/jei.202300500

[10] Wróblewska A. et al. Methods and tools for ontology building, learning and integration–application in the synat project. *Intelligent tools for building a scientific information platform.* 2012. p. 121-151. doi: https://doi.org/10.1007/978-3-642-24809-2_9

[11] Loukachevitch N.V. *Tezaurusy v zadachah informacionnogo poiska* [Thesauri in information retrieval tasks]. Moscow: Moscow University Press; 2011. 512 p. (In Russ.) EDN: RBBMVR

[12] Bates M.J. How to Use Controlled Vocabularies More Effectively in Online Searching. *Online*. 1988;12(6):45-56.

[13] Davies J. Lightweight Ontologies. In: Poli R., Healy M., Kameas A. (eds.) Theory and Applications of Ontology: Computer Applications. Dordrecht: Springer; 2010. p. 197-229. https://doi.org/10.1007/978-90-481-8847-5_9

[14] Janowicz K., et al. SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*. 2019;56:1-10. https://doi.org/10.1016/j.websem.2018.06.003

[15] Olteanu-Raimond A.M., et al. A lightweight ontology for landmarks to assist rescue in mountainous areas. *Advances in Cartography and GIScience of the ICA*. 2023;4:15. https://doi.org/10.5194/ica-adv-4-15-2023

[16] Guarino N. Some ontological principles for designing upper level lexical resources. In: First International Conference on language resources & evaluation: Granada, Spain, 28-30 May 1998. European Language Resources Association; 1998. p. 527-534. https://doi.org/10.48550/arXiv.cmp-lg/9809002

[17] Loukachevitch N. Establishment of Taxonomic Relationships in Linguistic Ontologies. In: Wolff K.E., Palchunov D.E., Zagoruiko N.G., Andelfinger U. (eds.) Knowledge Processing and Data Analysis. KPP KONT 2007. *Lecture Notes in Computer Science*. Vol. 6581. Berlin, Heidelberg: Springer; 2011. p. 232-242. https://doi.org/10.1007/978-3-642-22140-8_15

[18] Romanenko E., Calvanese D., Guizzardi G. Abstracting Ontology-Driven Conceptual Models: Objects, Aspects, Events, and Their Parts. In: Guizzardi R., Ralyté J., Franch X. (eds.) Research Challenges in Information Science. RCIS 2022. *Lecture Notes in Business Information Processing*. Vol. 446. Cham: Springer; 2022. p. 372-388. https://doi.org/10.1007/978-3-031-05760-1_22

[19] Loukachevitch N.V. Part-whole relations in theory and practice. *Neurocomputers*. 2013;(1):007-012. EDN: PVRZNN

[20] Kravec S.L. Scientific and educational encyclopedic portal and participation of regional scientific encyclopedias in it. *Voprosy ehnciklopedistiki.* 2019;(2):14-20. (In Russ., abstract in Eng.) EDN: NPZYHS

[21] Vivaldi J., Rodríguez H. Finding Domain Terms using Wikipedia. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association; 2010. p. 386-393. Available at: https://aclanthology.org/L10-1518/ (accessed 21.05.2024).

[22] Bordea G., et al. Evaluation Dataset and Methodology for Extracting Application-Specific Taxonomies from the Wikipedia Knowledge Graph. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2020. p. 2341-2347. Available at: https://aclanthology.org/2020.lrec-1.285/ (accessed 21.05.2024).

[23] Kirillovich A., Nevzorova O. Ontological Analysis of the Wikipedia Category System. In: Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) – KEOD. Seville, Spain: SciTePress; 2018. p. 358-366. https://doi.org/10.5220/0006961803580366

[24] Suchanek F.M., Alam M., Bonald T., Chen L., Paris P.-H., Soria J. YAGO 4.5: A Large and Clean Knowledge Base with a Rich Taxonomy. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24). New York, NY, USA: Association for Computing Machinery; 2024. p. 131-140. https://doi.org/10.1145/3626772.3657876

[25] Vrandečić D., Pintscher L., Krötzsch M. Wikidata: The Making Of. In: Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion). New York, NY, USA: Association for Computing Machinery; 2023. p. 615-624. https://doi.org/10.1145/3543873.3585579

Submitted 21.05.2024; approved after reviewing 04.06.2024; accepted for publication 22.06.2024.

About the authors:

Irina M. Kobozeva, Professor of the Department of Theoretical and Applied Linguistics, Faculty of Philology, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), Dr. Sci. (Philol.), **ORCID: https://orcid.org/0000-0003-4214-2878**, kobozeva@list.ru

Natalia V. Loukachevitch, Leading Researcher of the Research Computing Center, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), Dr. Sci. (Tech.), **ORCID: https://orcid.org/0000-0002-1883-4121**, louk_nat@mail.ru

Maria B. Panich, Master degree student of the Department of Theoretical and Applied Linguistics, Faculty of Philology, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), **ORCID: https://orcid.org/0009-0005-7400-5837,** shmak1280@gmail.com

Fedor A. Sadkovsky, Master degree student of the Department of Theoretical and Applied Linguistics, Faculty of Philology, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation); Researcher of the Laboratory for Study and Preservation of Minority Languages, Institute of Linguistics of the Russian Academy of Sciences (1, build. 1, Bolshoy Kislovsky lane, Moscow, 125009, Russian Federation), **ORCID: https://orcid.org/0009-0006-4502-9466,** feudor987@mail.ru

Evgeny N. Raevsky, Lead Software Developer of the Laboratory of Computing Practice Software, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), **ORCID: https://orcid.org/0000-0002-0119-6418**, raev@cs.msu.ru

All authors have read and approved the final manuscript.