

Граф знаний научного института в онтологии семантической библиотеки

О.М. Атаева¹, В.А. Серебряков¹, Н.П.Тучкова¹

¹ФИЦ ИУ РАН, Москва, ул. Вавилова, 40

Аннотация. Рассматривается проблема представления научных результатов в цифровой среде. Новый взгляд на пространство знаний научного института составляет естественный этап развития WEB-технологий. Заложенная в предыдущих исследованиях структура данных, позволяет организовать поиск и навигацию по ним с помощью графа знаний, как версию семантической библиотеки LibMeta. Граф знаний дает более полное и качественное представление о пространстве знаний, зачастую снимает *когнитивную* нагрузку в восприятии сложных структур и связей данных.

Ключевые слова: прикладная онтология, граф знаний, источники данных, разработка онтологии, научные результаты в цифровой среде

Knowledge graph of a scientific institute in the semantic library ontology

O.M.Ataeva¹[0000-0003-0367-5575], V.A.Serebryakov²[0000-0003-1423-621X],
N.P.Tuchkova³[0000-0001-5357-9640]

^{1,2,3}FRS «Computer Sciences and Control», Russian Academy of Sciences,
Vavilov str., 40, Moscow, 119333, Russia

¹oli.ataeva@gmail.com, ²serebrvas@gmail.com, ³natalia_tuchkova@mail.ru

Abstract. The problem of presenting scientific results in a digital environment is considered. A new look at the knowledge space of a scientific institute constitutes a natural stage in the development of WEB technologies. The data structure inherent in previous studies allows you to organize search and navigation through them using a knowledge graph, as a local version of semantic library LibMeta. The knowledge graph gives a more complete and high-quality idea of the knowledge space, often removing the cognitive load in the perception of complex structures and data connections

Keywords: applied ontology, knowledge graph, data sources, ontology development, scientific results in the digital environment

1. Введение

Структура научного института практически многократно описана, рассмотрим ее в современном представлении в виде онтологии и графа знаний. По прошествии многих лет со времени первых цифровых библиотек научных организаций изменился взгляд на представление структуры данных, произошел переход от иерархического подхода к сетевому. Изменения связаны с развитием онтологического проектирования, учетом семантических связей тематик, публикаций и персон. Именно данные о персонах, публикациях, темах исследований составляют основу представления научного института в виде графа знаний, а административное подчинение, которое ранее играло основополагающую роль в описании структуры, может быть представлено, как альтернативный граф на тех же данных.

Исследования опираются на разработку, «Научный Институт РАН», которая была реализована в Вычислительном центре им. А.А.Дородницына РАН [1, 2]. Эта работа дала старт различным изысканиям в области представления научной и административной деятельности академических и отраслевых институтов, на платформе интегрированной системы Единого Научного Информационного Пространства РАН (ЕНИП) реализована и функционирует более 20 лет портал Академии наук РАН (<https://www.ras.ru/>).

Мотивацией новых исследований послужило развитие интегрированного подхода на основе онтологического проектирования и навигации с помощью графа знаний в цифровой семантической библиотеке LibMeta(<https://www.libmeta.ru/>) [3, 4]. Также необходимо отразить новый взгляд на тематику представления знаний в связи с возросшим интересом к этой проблеме [5, 6, 7] и необходимостью научных междисциплинарных коммуникаций [8].

2. Модель, данные, навигация

2.1. Модель

Описание данных и модели подробно изложены в монографиях [1, 2] и многочисленных библиографических ссылках этих публикаций. В основе описания данных ЕНИП лежит онтология и язык OWL (<https://www.w3.org/OWL/>). Основными элементами онтологии OWL ЕНИП являются описания классов, их свойств, отношений между классами и представителями (индивидов) классов (их свойств и отношений). Для этих описаний OWL последовательно использует бинарные отношения своего словаря, а также словарей RDF и RDFS (<https://www.w3.org/TR/owl-semantics/>).

Центральным понятием информационного ресурса организации в ЕНИП выступает *персона*. Исходя из свойств *персоны* в научной

организации, строится модель, структура данных и связей, функциональность (Рис. 1).



Рис. 1. Диаграмма зависимости основных схем в ЕНИП.

Свойствами персоны, в частности, являются публикации, которые помещены в библиотеку ЕНИП ВЦ РАН с соответствующим библиотечным сервисом. Публикации обрабатываются, индексируются, komponуются в указатель трудов института (Рис. 2), составляют одно их свойств *персоны*.



Рис. 2. Указатель публикаций авторов ВЦ РАН на портале OMN РАН.

Научные учреждения в целом заинтересованы в получении и предоставлении доступа к данным о научных достижениях, научной деятельности сотрудников и организаций. Эта информация представляет интерес для конечных пользователей информационной системы организации. Она поз-

воляет сотрудникам получить информацию о смежных работах в других коллективах. Организации могут предоставлять свои данные в слабоструктурированном виде, например, как текстовые или XML-документы. Формат таких документов, как правило, отражает аспекты конкретной предметной области и сложившихся потребностей организации. Кроме того, многие организации поддерживают собственные сайты, заполненные статическими страницами, и необходимо предоставить им простой инструментарий, который позволил бы как управлять сайтами через веб-интерфейс, так и агрегировать их информацию в объединенное научное пространство.

При разработке семантических библиотек особое внимание уделяют модели данных содержимого библиотеки, в центре которой в LibMeta – *публикация*, научный труд. При этом контент цифровых библиотек может быть описан различными форматами и представлен различными способами. Библиотека, определяемая с помощью системы LibMeta, рассматривается как хранилище структурированных разнообразных данных с возможностью их интеграции с другими источниками данных и предполагает возможность специфицирования своего контента за счет описания предметной области. В качестве средства формализации выступает *онтология контента* семантической библиотеки.

Семантическая цифровая библиотека – это цифровая библиотека, элементы контента которой связаны иерархическими и ассоциативными отношениями в соответствии с онтологией предметной области.

Онтология цифровой семантической библиотеки – формальное описание множества данных (типы элементов данных, связи) предметной области. Для определения онтологий используется язык OWL. Онтология на OWL – это RDF граф. Описание на OWL – это описание *структуры* данных, а не самих данных. Каждый экземпляр данных – это экземпляр элемента онтологии. В библиотеке LibMeta тезаурус представляется как онтология. Средства редактирования LibMeta – это средства редактирования онтологии.

Решая задачу проектирования онтологии, приходим к необходимости использования метаданных разного уровня:

- метаданные как универсальные понятия структуры данных цифровой библиотеки;
- метаданные как часть описания объектов прикладной области или подмножества прикладной области;
- метаданные прикладной области как таковые.

В такой онтологии на *верхнем уровне используются понятия*, по сути своей относящиеся к высокоуровневым онтологиям и не связанные со спецификой какой-то конкретной предметной области. На *втором уровне используются понятия, описывающие предметную область*, при этом являющиеся экземплярами классов, определенных на первом уровне, но

при этом используемые как определения классов для описания данных третьего уровня уже в конкретной предметной области. Определение предметной области задается тезаурусом, который содержит основные термины этой предметной области, связанные иерархическими и горизонтальными связями между собой. Содержимое библиотеки задается типами ресурсов, описание которых задает множество допустимых объектов, возможно объединенных в разнообразные коллекции, составляющие вместе с тезаурусом ее контент.

В онтологии цифровой библиотеки LibMeta можно выделить «системную» часть, так или иначе описывающую структуру самой библиотеки, и «пользовательскую», описывающую структуру данных предметной области, загруженных в библиотеку. Тезаурус является в каком-то смысле верхней частью пользовательской онтологии, полностью и явно видимой пользователю.

2.2. Данные

Для интеграции данных в информационную систему научного института необходимо выделять в них структуру, приводить к агрегатным моделям и схемам данных и связывать между собой и с имеющимися в среде данными. К примеру, организация может предоставлять информацию о научных публикациях, отчётах, о результатах научной деятельности за некоторый промежуток времени. Желательно, чтобы информация была автоматизированным образом проанализирована и структурирована (Рис.3) с выделением авторов публикации, издательства, и других атрибутов – названия, ISBN и пр.

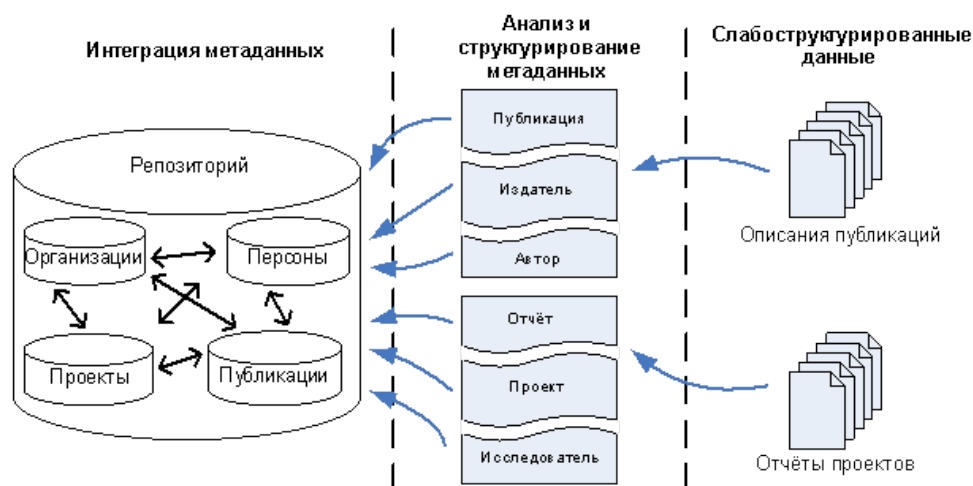


Рис. 3. Схема структуризации слабоструктурированных данных.

При интеграции данных может выясниться, что в хранилище уже имеется некоторая информация об авторах публикации, например, другие статьи, или информация об участии в конференциях и пр., желательно,

чтобы было обеспечено автоматизированное сопоставление и связывание соответствующих ресурсов, чтобы исключить дублирование.

В первоначальной реализации проекта ЕНИП не рассматривались такие структуры данных, как *онтология предметной области* и не ставилась задача определения принадлежности публикации к некоторой предметной области исследований.

На приведенной на диаграмме Рис. 4 (а, б) отражен состав классов библиографической специализации ЕНИП и указаны отношения между ними. Подробная информация о каждом из классов, его семантике и вариантах использования, составе свойств приведена в работах [1,2].

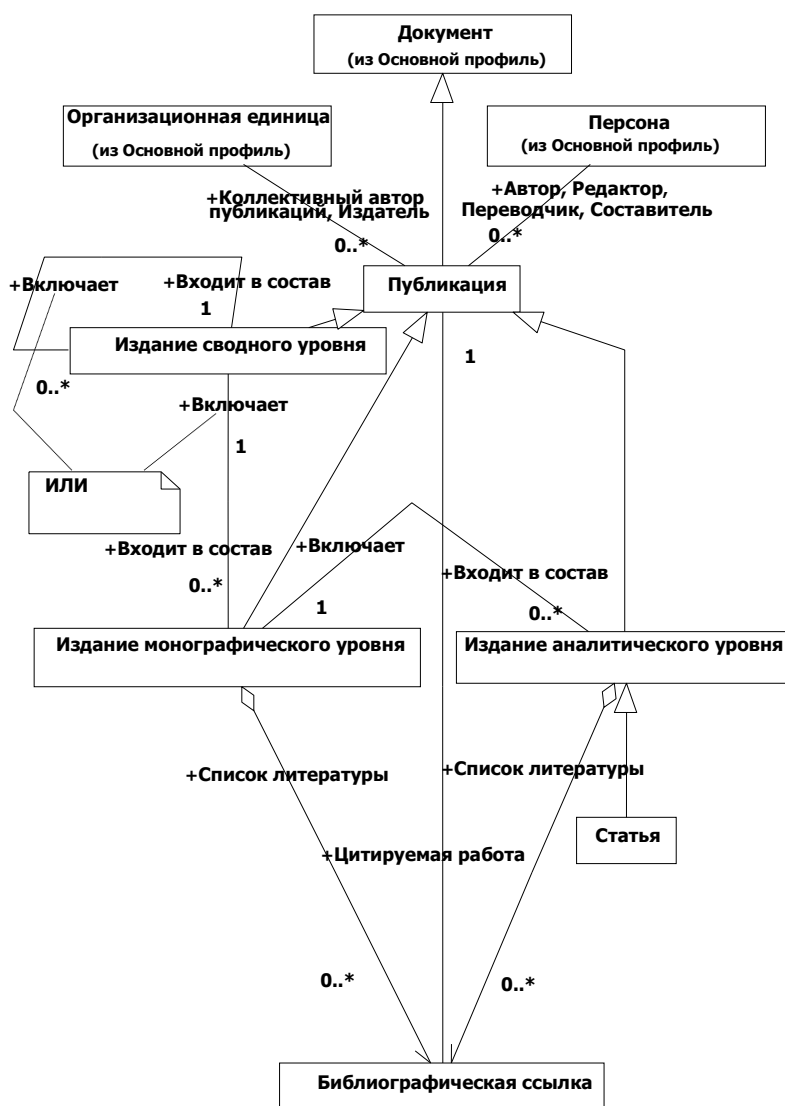


Рис. 4а. Классы библиографической специализации (первая часть).

В целях обеспечить поддержку различных уровней детализации информации о публикациях, необходимых различным приложениям, библиографическая специализация разделена на базовую и расширенную подсхемы, а также выделяется академическая подсхема, отражающая специфику научных публикаций.

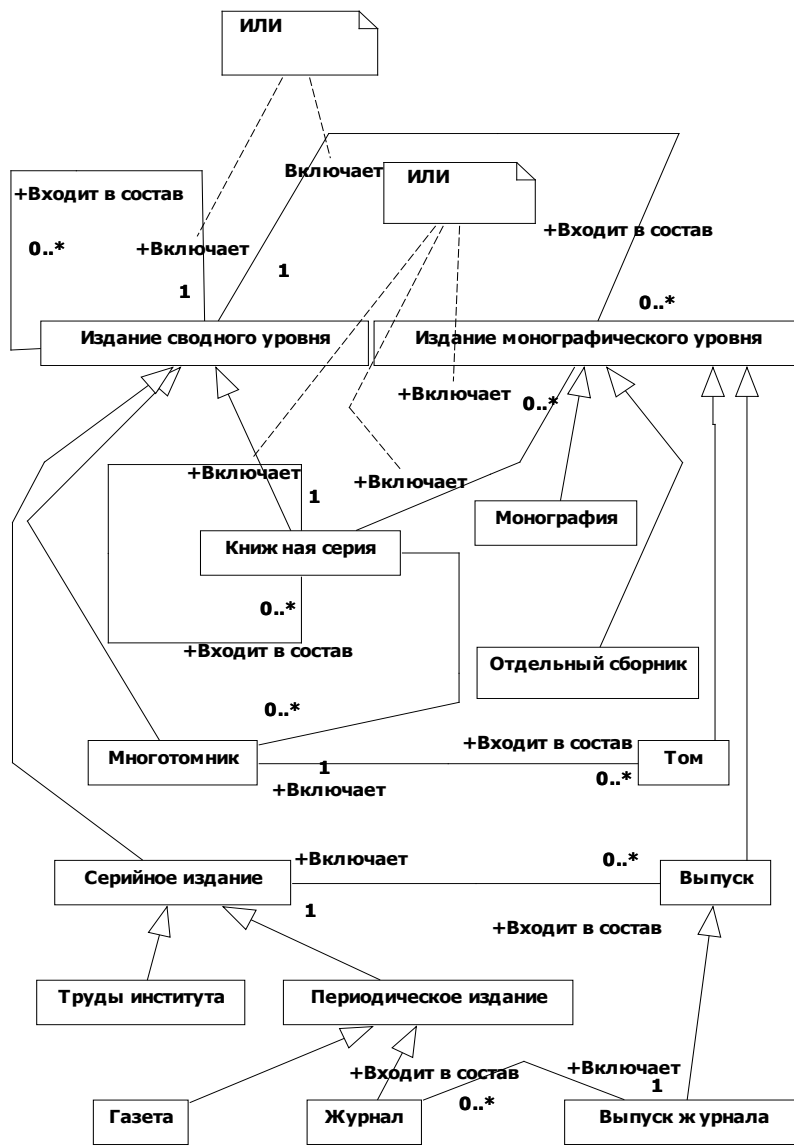


Рис. 4б. Классы библиографической специализации (вторая часть).

К расширенной схеме отнесена дополнительная информация о полном тексте публикации (оглавление, библиографические ссылки), более подробная информация о связанных с публикацией лицах и редколлегии, а также различные «узкоприменимые» описательные характеристики. При этом уже в базовой схеме почти все свойства указаны как необязательные для заполнения, что позволяет даже при наличии неполной информации о публикации описать ее в соответствии со схемами ЕНИП. Уже на базовом уровне требуется структурировать информацию обо всех вышестоящих библиографических уровнях для каждой публикации. Например, для описания ряда статей в журнале, необходимо описать сам журнал как издание сводного уровня, далее описать интересующие выпуски этого журнала как издания монографического уровня, и, наконец, сами статьи как издания аналитического уровня.

Построение *онтологии предметной области в LibMeta* позволяет выделить *метаданные для проектирования конкретных структур данных* научных предметных областей и варианты управления этими данными. Для определения предметной области строится *онтология предметной области*, в рамках которой могут интегрироваться различные источники данных, использоваться различные таксономии понятий и терминов, верифицированные признанными экспертами научной области. Для этого также необходимо структурировать и связать различные *ресурсы*, извлечь из них и контекстуализировать (определить в контексте) *данные, придавая им свойства знаний*. В основу онтологии библиотеки LibMeta были положены энциклопедии, словари, тезаурусы.

2.3. Навигация

Онтология цифровой библиотеки определяет *структуру данных* контента библиотеки. Каждому элементу данных, загруженных в библиотеку, можно сопоставить *вершину онтологии*, определяющую положение элемента данных в онтологии («тип данных» элемента). На основании связей онтологии и связей, определенных на этапе проектирования, можно построить *граф данных*, структура которого определяется онтологией: вершины (статьи) – экземпляры элементов онтологии, связи – связи тезауруса. Это *граф знаний* цифровой библиотеки.

Роль онтологии в процессе проектирования и эксплуатации цифровой библиотеки можно резюмировать следующим образом:

- на основе онтологии строится база данных библиотеки (Рис. 1-4);
- онтология может использоваться для создания точки доступа LOD к данным библиотеки;
- на основе онтологии строится граф знаний библиотеки (Рис. 5).

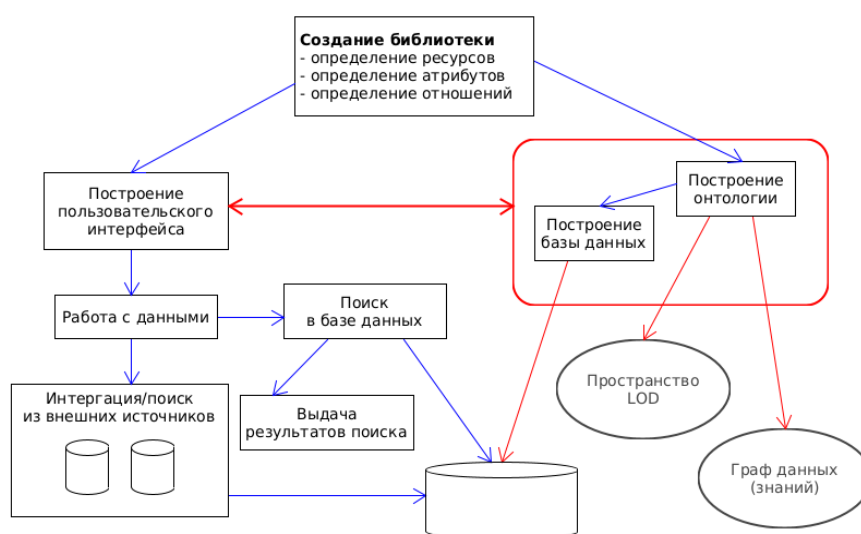


Рис.5. Схема построения графа знаний LibMeta.

3. Примеры графа знаний научного института

Представление предметной области в виде тезауруса и онтологии составляет основу для формирования графа знаний для организации навигации по данным библиотеки.

Граф знаний представляет собой базу знаний, которая использует модель данных или топологию, структурированную графом, для интеграции данных. Граф знаний – это граф, свойства вершин которого, дуги (отношения между вершинами) и методы работы с ними определяются приложениями, характерными для обработки «знаний», что в свою очередь определяется семантикой этого понятия и может меняться от приложения к приложению. Граф знаний - это граф с определенной семантической интерпретацией вершин и дуг (бинарных отношений).

Все данные (статьи, персоны, ключевые слова и т.д.) семантической библиотеки, благодаря онтологическому представлению могут быть представлены в виде графа знаний.

На Рис. (Рис. 6а, 7а) представлен граф с вершинами «публикация» и «понятие тезауруса» и их связи через ребра с вершинами другого уровня, где цветом выделены уровни вершин соответствующего уровня. На Рис. 6б, 7б представлены соответствующие странички цифровой библиотеки LibMeta.



Рис. 6а. Связи публикаций с вершинами разного уровня, представленными различными цветами.

Доступ к данным, через граф знаний обеспечивается онтологией библиотеки и реализуется через интерфейс библиотеки.

Просмотр объекта

Методы интегрирования обыкновенных дифференциальных уравнений.

Тип объекта [Литература ОДУ](#)
 URI <http://libmeta.ru/resource/publications#odu#25023>
 Атрибуты
 Год издания - 1955
 Название - Методы интегрирования обыкновенных дифференциальных уравнений.
 Авторы ОДУ - [Матвеев Н.М.](#)
 Код - 28
 Издательство - Л.: Изд-во Ленинградского университета

Связанные понятия из тезауруса, в которых объект используется в качестве значения атрибута

- [Параметра введения обший метод \(\$F\(x, y, y'\) = 0\$ \)](#)
- [Вариация произвольной постоянной \(Лагранжа метод\)](#)
- [Н. М. Гюнтера метод решения Якоби ОДУ первого порядка](#)
- [Метод интегрирующего множителя ОДУ первого порядка](#)
- [Бернулли ОДУ](#)
- [Лежандра преобразование \(\$F\(x, xy' - y, y'\) = 0\$ \)](#)
- [ОДУ первого порядка в полных дифференциалах](#)
- [ОДУ первого порядка линейное](#)
- [ОДУ первого порядка, разрешенное относительно производной](#)
- [Особая точка ОДУ первого порядка](#)
- [Особое решение \(особый интеграл\) ОДУ первого порядка](#)
- [Риккати ОДУ общее](#)
- [Точка покоя автономной системы второго порядка](#)
- [Центр \(точка покоя\)](#)
- [Решение ОДУ первого порядка](#)
- [Эйлера метод множителя](#)
- [Пирсона ОДУ](#)
- [Звляча о траекториях](#)
- [Дарбу ОДУ первого порядка](#)
- [ОДУ первого порядка \$n\$ -ой степени](#)
- [ОДУ первого порядка, не разрешенное относительно производной](#)
- [Лагранжа ОДУ первого порядка](#)
- [Клеро ОДУ...](#)

Рис. 6б. Связи публикаций.

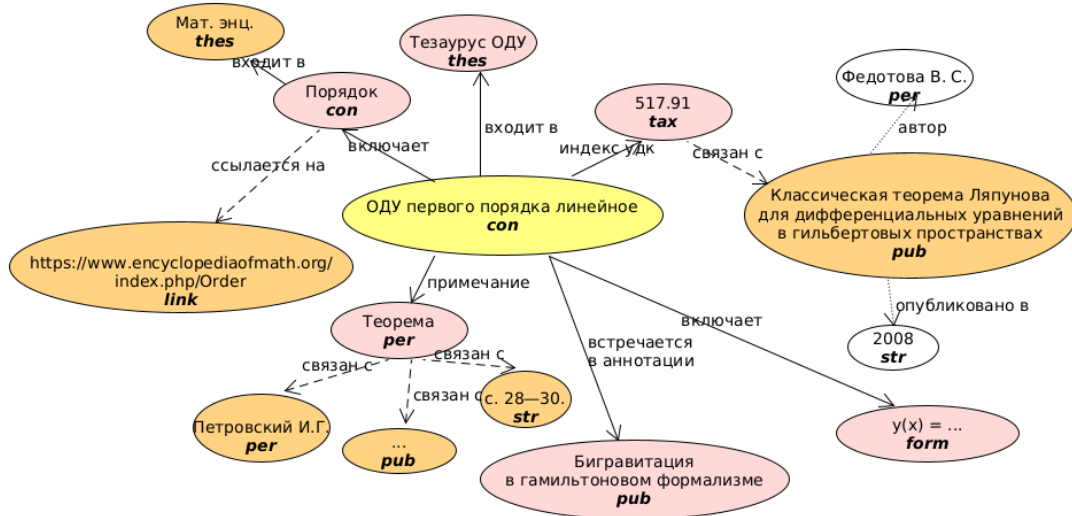


Рис. 7а. Связи понятий тезауруса с вершинами разного уровня, представленными различными цветами.

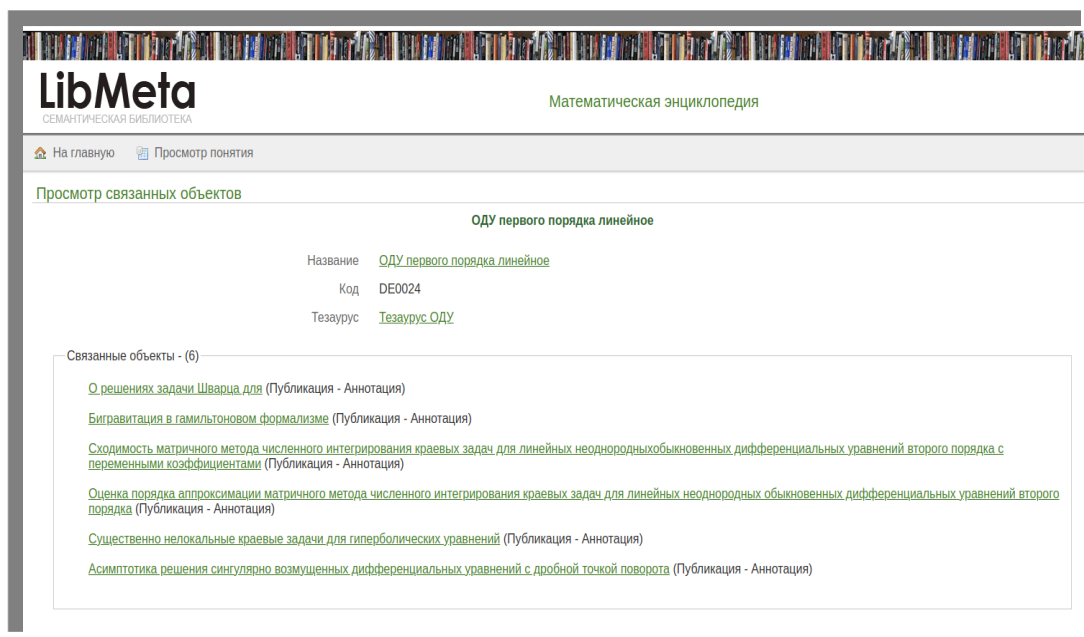


Рис. 76. Связи понятий тезауруса.

4. Заключение и выводы

Обсуждается новый взгляд на проблему представления научных трудов, результатов и структуры академического института. На основе семантического представления данных в цифровой библиотеке института формируется граф знаний, позволяющий осуществить навигацию по различным узлам, таким как административные единицы и персоны, научные труды и результаты и т.д.

Дальнейшие исследования посвящены разработке и совершенствованию функциональности локальной версии LibMeta – научный институт, а также реализации интерфейсов для различных категорий пользователей.

Работа представлена в рамках выполнения темы НИР «Математические методы анализа данных и прогнозирования» ФИЦ ИУ РАН.

Литература

1. Бездушный А.А., Бездушный А.Н., Серебряков В.А., Филиппов В.И. Интеграция метаданных Единого Научного Информационного Пространства РАН. М: Вычислительный центр им. А.А.Дородницына Российской академии наук, 2006. 238 с.
2. Бездушный А.А., Бездушный А.Н., Нестеренко А.К., Серебряков В.А., Сысоев Т.М., Теймуразов К.Б., Филиппов В.И. Информационная Web-система «Научный институт» на платформе ЕНИП. М: Вычислительный центр им. А.А. Дородницына Российской академии наук, 2007. 253 с.

3. Ataeva O., Serebryakov V., and Tuchkova N., Ontological approach to a knowledge graph construction in a semantic library // Lobachevskii J. of Mathematics, 44, (6) (2023) pp. 2229–2239. <https://doi.org/10.1134/S1995080223060471>
4. Ataeva O., Kornet Yu.N., Serebryakov V., and Tuchkova N., Approach to creating a thesaurus and a knowledge graph of an applied subject area // Lobachevskii J. of Mathematics, 44, (7) (2023) pp. 2577–2586. <https://doi.org/10.1134/S1995080223070077>
5. Загорулько Ю.А., Загорулько Г.Б. Искусственный интеллект. Инженерия знаний. М: Юрайт, 2018. 93 с.
6. Гаврилова Т.А., Кудрявцев Д.В., Муромцев Д.И. Инженерия знаний. Модели и методы: Учебник. СПб.: Издательство «Лань», 2016. 324 с.
7. Боргест Н.М. Научный базис онтологии проектирования // Онтология проектирования. №. 1. 2013. С. 7-25.
8. Scholarly Knowledge. Open Research Knowledge Graph. <https://orkg.org/>

References

1. Bezdushnyj A.A., Bezdushnyj A.N., Serebryakov V.A., Filippov V.I. Integraciya metadannyh Edinogo Nauchnogo Informacionnogo Prostranstva RAN. М: Vychislitel'nyj centr im. A.A.Dorodnicina Rossijskoj akademii nauk, 2006. 238 p.
2. Bezdushnyj A.A. , Bezdushnyj A.N. , Nesterenko A.K. , Serebryakov V.A. , Sysoev T.M. , Tejmurazov K.B. , Filippov V.I. Informacionnaya Web-sistema «Nauchnyj institut» na platforme ENIP. М: Vychislitel'nyj centr im. A.A.Dorodnicina Rossijskoj akademii nauk, 2007. 253 p.
3. Ataeva O., Serebryakov V., and Tuchkova N., Ontological approach to a knowledge graph construction in a semantic library // Lobachevskii J. of Mathematics, 44, (6) (2023) pp. 2229–2239. <https://doi.org/10.1134/S1995080223060471>
4. Ataeva O., Kornet Yu.N., Serebryakov V., and Tuchkova N., Approach to creating a thesaurus and a knowledge graph of an applied subject area // Lobachevskii J. of Mathematics, 44, (7) (2023) pp. 2577–2586. <https://doi.org/10.1134/S1995080223070077>
5. Zagorul'ko YU.A., Zagorul'ko G.B. Iskusstvennyj intellekt. Inzheneriya znaniy. М: YUrajt, 2018. 93 p.
6. Gavrilova T.A., Kudryavcev D.V., Muromcev D.I. Inzheneriya znaniy. Modeli i metody: Uchebnik. SPb.: Izdatel'stvo «Lan'», 2016. 324 P.

7. Borgest N.M. Nauchnyj bazis ontologii proektirovaniya // Ontologiya proektirovaniya. N 1. 2013. P. 7-25.
8. Scholarly Knowledge. Open Research Knowledge Graph.
<https://orkg.org/>