УДК 004:81’37

СЕМАНТИЧЕСКОЕ ЯДРО ЦИФРОВОЙ ПЛАТФОРМЫ

Н.В. Максимов[1,a], О.Л. Голицына[1,b], М.Г. Ганченкова[1,c], Д.В. Санатов[2,3,d], А.В. Разумов[3,e]

[1] НИЯУ МИФИ, Высшая инжиниринговая школа, Москва, Россия
[2] ООО «Медиа-Лаб», Москва, Россия
[3] Фонд «Центр стратегических разработок «Северо-Запад», Санкт-Петербург, Россия
[a]nv-maks@yandex.ru, [b]olgolitsina@yandex.ru, [c]MGGanchenkova@mephi.ru,
[d]dmsa@yandex.ru, [e]a.razumov@csr-nw.ru

Аннотация

Дано обобщённое описание интегрированного комплекса декларативных и процедурных средств, в совокупности обеспечивающих согласованное информационное (семантическое) представление сложных объектов на всех этапах их жизненного цикла. Знания (факты) представляются системой онтологий разного уровня, а структура - таксономиями и классификациями. Онтология определяется как система трёх взаимосвязанных систем (функциональной, понятийной и знаковой), а система таксономий представляет классы объектов и процессов, характерных для основных «координат» деятельности. Информационный поиск, как компонент семантического ядра, рассматривается как сложный самосогласованный процесс конструирования нового знания, где знание – это информация (тексты находимых документов), связываемая с контекстом задачи и представлениями пользователя. Такой контекст целенаправленно или косвенно задаётся пользователем посредством предопределённых семантических структур (таксономий, онтологий) либо посредством динамически формируемых компонентов (словников, выборок и т.д.). Это составляет существо семантического когнитивного поиска, когда система не только реализует отбор документов традиционными методами поиска, но и формирует образ информационной потребности, что, в свою очередь, позволит системе синтезировать комплексные, аспектно-ориентированные ответы. Предлагаются автоматизированные технологии поддержки лингвистического обеспечения, основанные на дистрибутивно-статистическом анализе как потоков объектного знания и неявного знания (извлекаемого системой в процессе взаимодействия), так и компонентов понятийно-терминологических систем. Представленные в статье средства апробированы в рамках разработанного программного комплекса xIRBIS-ML, предназначенного для организации семантического поиска в массивах данных сложных инженерных объектов.

Ключевые слова: *семантическое ядро, цифровые платформы, семантический поиск, онтологии.*

Цитирование: *Максимов, Н.В.* Семантическое ядро цифровой платформы / Н.В. Максимов, О.Л. Голицына, М.Г. Ганченкова, Д.В. Санатов, А.В. Разумов // Онтология проектирования. – 2018. – Т. 8, №3(29). – С.412-426. – DOI: 10.18287/2223-9537-2018-8-3-412-426.

Введение

Согласно государственной программе «*Цифровая экономика Российской Федерации*» [1] эффективное развитие рынков и отраслей (сфер деятельности) в цифровой экономике возможно только при наличии развитых платформ, технологий, институциональной и инфраструктурной сред.

Для сложных технических и социальных систем, к которым относятся и указанные в упомянутой государственной программе цифровые платформы (ЦП), характерно то, что они состоят из большого числа неоднородных элементов и подсистем, которые обладают значительным разнообразием внутренних и внешних связей и могут находиться во множестве состояний, в том числе в процессах становления и эволюции. Поэтому важнейшей подзадачей задачи управления жизненным циклом (ЖЦ) сложных систем является, с одной стороны,

развитие Интернета вещей, а с другой - управление знаниями на всех этапах ЖЦ этих систем. Именно последнее обеспечит возможность саморазвития сложных систем, синтетически включающих компоненты уровня вещей и уровня знаний. То есть управление знаниями на базе общей информационной ЦП в течение всего ЖЦ – это то, что обеспечивает единый, интегрированный подход к созданию, сбору, организации, распределению знаний, а в итоге приведёт к эффективному взаимодействию и использованию информационных и технологических ресурсов.

Рост актуальности вопросов развития ЦП, трансформация рынков проектирования и эксплуатации сложных инженерных объектов в таких отраслях, как атомная промышленность (где в настоящее время идёт активный поиск новых технологий управления знаниями), авиа- и судостроение, энергетика, существенно повысил актуальность тематики семантического поиска и сформировал запрос на соответствующие промышленные технологии.

Настоящая статья представляет собой обобщённое описание семантического ядра - комплекса декларативных, процедурных и информационных средств, которые в совокупности могут обеспечить согласованное семантическое представление сложных объектов, к множеству которых относятся и системы представления и управления знаниями. Разработанные[1] и использованные подходы и решения представлены кратко, в объёме, который необходим для указания места и роли компонентов.

1 Цифровая платформа и семантическое ядро

ЦП – система алгоритмизированных взаимоотношений значимого количества участников рынка, объединённых единой информационной средой, приводящая к снижению транзакционных издержек за счёт применения пакета цифровых технологий и изменения системы разделения труда [1]. Другими словами, ЦП – это площадка, поддерживающая комплекс автоматизированных процессов и согласованное использование широкого спектра цифровых продуктов (услуг) значительным количеством потребителей.

Одним из вопросов развития ЦП является взаимодействие технологических платформ между собой и взаимодействие ЦП с экосистемой. Основой такого взаимодействия являются декларативные и процедурные средства, обеспечивающие интероперабельность информационных систем (ИС), а также стандартизация информационных технологий (ИТ). При этом, согласно [2], одним из ключевых условий интероперабельности современных платформ является наличие семантического слоя организации данных.

Функциональность семантического слоя определяется характеристиками семантического ядра, которое представляет собой систему декларативных и процедурных средств, обеспечивающих согласованную автоматическую и автоматизированную (в том числе и интерактивную) идентификацию, поиск и анализ информации и знаний.

Семантический слой организации данных, по сути являющийся субплатформой, – это синтетическое соединение трёх компонентов:

- документальных информационно-поисковых систем и баз данных, обеспечивающих углублённый семантический поиск и анализ разнородной информации;
- человеко-машинных информационных интерфейсов, обеспечивающих персонифицируемое представление познавательной траектории пользователя;

[1]Работа основана на результатах научных исследований и практических разработок в области информационного поиска и управления знаниями, проводимых коллективом под руководством проф. Н.В. Максимова в течение более двух десятилетий. В ней также учён опыт разработки и внедрения Информационно-аналитической системы IRBIS (©1992-2018), используемой рядом ведущих информационных центров и организаций для создания промышленных документальных баз данных.

- лингвистического обеспечения, построенного на базе гибридных методов лингвистического, лингво-статистического и структурного анализа текстов, обеспечивающих построение адекватных смысловому содержанию документов их формализованных семантических образов, над которыми определены операции, в том числе, корреспондирующие с процессами познания.

Это триединство рассматривается в контексте общего процесса синтеза, представления и поиска знаний.

Для современного состояния рынка ИТ-продуктов, ориентированных на работу с семантикой (цифровой инжиниринг, управление знаниями, семантическая обработка текстов) характерно то, что эти системы, как правило, создаются либо как закрытые проприетарные промышленные продукты, либо как узкоспециализированные решения, либо как академические проекты, что накладывает существенные ограничения на интероперабельность и, как следствие, на их широкое или совместное использование. Например, для полноценной цифровизации проектов атомной энергетики необходимо в части семантики совместить решения систем проектирования от Intergraf, Siemens, Dassault Systems, информационные системы, ориентированные на сохранение и управление знаниями, как например, Temelin knowledge management system [3] или система сохранения знаний по быстрым реакторам [4], а также функциональные технологии, ориентированные на обработку текстов и извлечение знаний, как ABBYY Compreno [5] или RCO Fact Extractor [6], технологии OSTIS [7].

Особенно острой эта проблема становится в условиях смещения границы отраслевых рынков и связанных с этим смещений границ семантического описания предметных областей (ПрО). Необходимо создание семантической инфраструктуры, способной интегрировать семантические данные из разных отраслей промышленности и областей знаний, а значит быть открытой для использования различными компаниями и способной интегрировать сторонних поставщиков семантического программного обеспечения.

Другая проблема, преодоление которой обеспечит выход ЦП на новый уровень, - высокая динамика изменения данных и связанных с ними метаданных. То есть семантическое ядро ЦП должно обладать свойствами, обеспечивающими саморазвитие. В частности, в контексте задач семантического поиска и управления знаниями система должна идентифицировать (индексировать) информацию «на лету» в зависимости от специфики задачи пользователя, отражаемой в когнитивных инфраструктурах.

2 Основные принципы разработки и функционирования ЦП

Семантическая платформа xIRBIS-ML – ориентированная на управление знаниями информационно-аналитическая система, разрабатываемой коллективом специалистов НИЯУ МИФИ, Фонда «Центр стратегических разработок «Северо-Запад» и ООО «Медиа-Лаб». Семантическое ядро ЦП xIRBIS-ML проектируется как система, обладающая свойствами функциональной полноты по отношению к среде, включая пользователей, а с точки зрения наполнения – как способная гармонично представлять знания разных отраслей и на разных уровнях общности/детализации. Основными принципами разработки и функционирования являются:

- взаимодополнительность процессов анализа/синтеза восходящих/нисходящих информационных потоков (знаний/информационных потребностей);
- интеграция функций и основных элементов информационного сопровождения этапов цикла генерации-использования знаний/данных на основе модели информационных представлений, основанной на принципах общей теории систем;

- полнота и избирательность информационного поиска, обеспечиваемые системой механизмов поиска и поиском в ассоциированных внешних информационных ресурсах (ИР);
- динамическая синхронизация взаимосвязи информационных и метаинформационных компонентов, основанная на общесистемной модели информационных представлений когнитивных процессов и управления;
- технологии оперативного построения и анализа лингвистического обеспечения (словников, рубрикаторов, тезаурусов, онтологий ПрО).

3 Интерактивный семантический поиск

Одной из основных и сложных задач семантики является смысловое отождествление содержания документов поисковым запросам. Такие подходы основываются на тезисе «смысл слова определяется его окружением» и, по существу, реализуют выявление контекста и, таким образом - определение (выбор) смысла слова.

При этом, говоря о семантическом поиске, необходимо понимать две особенности, которые важны для такого рода систем.

Первая – это то, что слово (словосочетание, выражение и даже весь текст) из найденного документа будет восприниматься в контексте *пользователя*. То есть смысл будет формироваться преимущественно на основе ПрО решаемой пользователем задачи и зависеть от полноты и специфики представления ПрО. При этом контекст ПрО, а именно общепринятые понятия и отношения между ними, может быть взят из базы знаний в общем случае интегрированной онтологии ПрО. Контекст пользовательского представления может быть сформирован динамически в процессе поиска путём построения виртуальной онтологии задачи из базовых онтологий и онтологических описаний отдельных решений [8]. Это помогает учитывать знания, которые в явном виде не присутствуют в конкретном высказывании/запросе, но существенно влияют на его смысл.

Вторая - это то, что пользователь «выстраивает» образ искомого решения, извлекая из текстов смысловые (и текстовые) фрагменты и связывая их таким образом, чтобы получаемая понятийная конструкция была непротиворечивой и обеспечивала бы решение его проблемы.

Технологической основой реализованной в xIRBIS-ML среды информационного взаимодействия «пользователь-система» является поисковый интерфейс, представляющий обобщённое рабочее пространство пользователя, ориентированное на процессы генерации знаний. Принципиально рабочее пространство включает две составляющие: собственно интерфейс формирования/развития запроса и обработки выдачи, а также структуру систематизации знаний пользователя – так называемый когнитивный рубрикатор (КР).

Поисковый интерфейс, помимо фактографического, тематического, семантического поиска с использованием чётких и нечётких механизмов отбора, вербальных или гипертекстовых технологий, позволяет пользователю осуществлять комплексный поиск (мультиобъектный, многоэтапный), например, для задач мониторинга проекта. Система, формируя поисковые образы и выдачи, готовит альтернативы, а структуры систематизации, протоколирующие поиск и идентифицирующие результаты, - задают направления «предпочтительного» развития. Эта задача решается как традиционными механизмами поиска, так и за счёт интерактивных средств визуального пространства, процедурно связывающих операционные объекты разного типа. В последнем случае поисковая траектория реализуется путём перемещения по ассоциированным визуальным элементам в находимых документах. Элементами могут быть слова или фрагменты текста, выделяемые по критерию соответствия понятиям, присутствующим в запросе или в профиле пользователя, а также динамически генерируемые гипертек-

стовые ссылки. Движущей силой, инициирующей очередные шаги поиска, является проблемная ситуация – информационная неопределённость, осознаваемая пользователем, а также дисбаланс в наполнении тематических подразделов проблемы. Автоматическое «отслеживание» таких дисбалансов обеспечивается использованием в качестве интегральной информационной структуры КР, включающего как информационные (документы, запросы, ссылки на ассоциированные ресурсы и т.п.) и метаинформационные (словари ПрО, классификации, рубрикаторы, тезаурусы, онтологии) компоненты, так и результаты аналитической обработки [9]. То есть поисковая система, помимо собственно отбора документов по заданному пользователем запросу, может способствовать выявлению предмета поиска (технологии извлечения и идентификации неявных знаний), а также обеспечить согласование на концептуальном и лингвистическом уровнях представления предмета поиска на стороне системы и пользователя.

Такая структура, соединяющая интенсиональное и экстенсиональное начала процесса познания отдельного субъекта и при этом представленная в распределённой сетевой среде, позволит осуществлять управляемый мониторинг как документальной, так и понятийно-терминологической составляющей. Это по сути является реализацией системного подхода и даёт возможность видеть в явной форме новые характеристические признаки, определять способы выделения подсистем и на основе свойств соответствия и симметрии обнаруживать связи (в т.ч. и противоречия) с другими системами классификации. Пользователь, осуществляющий поиск интересующей его информации, получает «многослойную» картину ПрО, что в итоге позволит ему систематизировать найденное и сделать более обоснованный выбор «траектории» поиска, учитывающий не только его представление о предмете поиска, но и сопоставительные оценки состояния и тенденций ПрО.

4 Информационные и лингвистические компоненты

Блок ЦП, практически представляющий собственно семантику в реализации xIRBIS-ML, содержит два типа логически взаимосвязанных компонентов.

Первый тип – это собственно «рабочая» информация, использование которой в сфере основной деятельности обеспечивает воспроизводство целевого продукта по его описанию (документации, схеме и т.п.). Это т.н. задокументированные знания (зафиксированные на носителе определённым способом), включая фактографические (таблицы данных, свойств и т.д.) и документальные виды информации (статьи, монографии, учебные пособия и т.д.), представляющие теории, гипотезы, эксперименты, критический опыт.

Второй тип – это информация справочного характера, обеспечивающая, в основном, представление и нахождение целевой (рабочей) информации. Сюда относятся:

- понятийно-терминологические системы (словари, тезаурусы, онтологии, глоссарии), являющиеся инструментами познания и средствами фиксирования знаний на носителях;
- классификационные схемы (таксономии, рубрикаторы и т.д.), обеспечивающие единообразие «членения» ПрО, исходя из целевых, организационных или методологических представлений.

Перечисленные и другие виды и формы вторичной информации представляют объективно-исторически и технологически сложившийся ряд [10] функционально-ориентированных инструментальных форм представления существа (семантики) информационных единиц с той или иной полнотой и точностью [11]. При этом с точки зрения характера источника смысла – определений, гипотез, теорий, базовых методов, частных решений и т.п. - различаются онтологии задач, решений, онтологии физических свойств и единиц измерения и т.д.

Кроме того, существуют и развиваются онтологии терминосистем, как, например, проекты WordNet [12], РуТез [13], FrameNet [14].

Практически построение онтологий осуществляется на базе следующих трёх блоков - источников информации:

- первый блок – фундаментальные знания, которые составляют статьи, монографии, отчёты, диссертации и т.п., представляющие объекты модельного уровня;
- второй блок – базовые компетентностные знания, в том числе учебная и методическая информация, которая представлена учебниками, пособиями, методическими указаниями, хрестоматиями и т.п., содержащими описание существа (содержания) предмета. Кроме того, в этот блок входят справочно-методические материалы – учебные программы, оценочные средства, которые определяют структуру знания и технологию познания, в том числе опорные и контрольные точки;
- третий блок – знания практического уровня, которые представлены проектной, конструкторской, технологической, нормативной документацией, экономическими, экологическими, аналитическими отчётами и т.п.

При этом в практике ИС понятийная основа устойчиво фиксируется в информационно-поисковых тезаурусах – наиболее известной и технологичной упорядоченной форме понятийно-знаковых систем.

Таксономические системы через упорядоченность концептуальных (классов) и реальных (экземпляров) составляющих, а также признаков классификации (существенных свойств) представляют целостность семантического пространства. В свою очередь онтологии различных уровней дискретно представляют «фрагменты» ПрО через ситуативные связи экземпляров объектов (знаний) различного уровня.

Роли лингво-семантических компонентов определяются исходя из того, что основными пространствами жизни объекта являются:

- пространство основной деятельности субъекта, где по форме существования выделяются два подпространства – абстрактных объектов (концептуальные модели, теории, составляющие предмет деятельности) и конкретных объектов (физические объекты ЖЦ);
- пространство информационной деятельности субъекта, предопределяющее способы и формы представления объекта в виде информационных сообщений;
- время как фактор, обусловливающий изменение знания и условий его применения, и как свойство, позволяющее фиксировать знание в виде дискретных макрообъектов.

Таким образом, для структурно-систематизированного представления состоявшегося (сгенерированного, систематизированного, проверенного, задокументированного) знания используется следующая «сетка координат»:

- координата «объект», задаваемая структурной таксономией, представляющей составные части объекта (узлы, детали, технологии и т.п.) с точки зрения совокупного процесса ЖЦ;
- координата «предмет», задаваемая функциональной таксономией (как «абстрактная модель ПрО») – структурой, представляющей теоретические и иные знания, относящиеся к этапам ЖЦ изделия;
- координата этапов работ (как фактор, отражающий разделение работ, предопределяемое специализацией субъекта, в процессах ЖЦ), задаваемая таксономией стадий и этапов ЖЦ и так или иначе связанной с ней таксономией форм представления знания – типов и видов документов как специфических форм и способов описания объекта.

В качестве «координат» можно использовать и любые другие системы классификации. Например, ГРНТИ, УДК, патентные и др. классификации. Кроме того, объект может идентифицироваться также и ключевыми словами, а глубинные семантические связи могут отражаться онтологиями.

При этом, подобно соотношению практической и модельной составляющих в целенаправленной деятельности в информационной деятельности выделяются ситуационная (на вторичном уровне наиболее полно представлена онтологиями) и структурно-концептуальная (на вторичном уровне представлена таксономиями) составляющие. Первая отражает комбинаторную природу деятельности и языка, позволяющую выражать (обозначать) новые смыслы «старыми» знаками. Вторая представляет базис, который обеспечивает преемственность и «вектор» развития ПрО, а в части процессов информационных коммуникаций является основой для «узнавания» нового за счёт апелляций к общим смыслам.

С точки зрения идентификации информации и знаний важным и перспективным свойством онтологического представления семантики является то, что, в отличие от линейного поискового образа в нём представлены ситуации (факт имманентной или ситуативной взаимосвязи двух объектов).

Онтология с точки зрения общей теории систем определена в [15] как совокупность трёх взаимосвязанных систем:

$$O =< S_f, S_c, S_t \equiv>,$$

где S_f - функциональная система (объекты и связи действительности) определяется как

$$S_f =< M_f, A_f, R_f, Z_f >,$$

где M_f множество объектов, A_f множество характеристических свойств, R_f – множество функциональных отношений, Z_f закон композиции, т.е., правил и схем упорядочения объектов (таксономия ПрО);

S_c - понятийная система, определённая как

$$S_c =< M_c, A_c, R_c, Z_c >,$$

где M_c – множество понятий ПрО, A_c – множество признаков систематизации понятий (мерономия), R_c – классы/подклассы парадигматических отношений, Z_c – закон композиции (схема упорядочения);

S_t - терминологическая система, определённая как

$$S_t =< M_t, A_t, R_t, Z_t >,$$

где M_t - множество терминов, A_t – множество свойств, R_t – множество отношений эквивалентности и включения, а также лингвистических отношений, Z_t – закон композиции (грамматика);

$\equiv$ - операция сопоставления элементов различных систем на уровне знаков, обеспечивающая их тождество в функциональной, понятийной и терминологической системах.

Автоматическое построение онтологий на основе представленных на естественном языке текстов, т.е. формирование связей между выделенными в тексте понятиями, основывается на преобразовании лингвистических отношений в функциональные. При этом набор типовых, так называемых функциональных, связей для ПрО обычно ограничен (подробно см. [16, 17]).

Автоматически построенные понятийно-терминологические графы могут быть отредактированы средствами интерактивных человеко-машинных процедур, которые в этом случае реализуют принцип дополнительности: представление знания в виде извлекаемых из текста ключевых слов и отношений система осуществляет с «устоявшейся» точки зрения (статистической значимости), а человек, внося изменения и дополнения в построенный системой образ (список терминов, граф), фиксирует отличия, характеризующие новизну и специфику по отношению к устоявшемуся и усредненному представлению знаний.

Однако при практическом построении таких объектов возникают существенные сложности. Богатство лексики естественного языка приводит к тому, что применение к построенным по текстам онтологиям теоретико-графовых операций (например, объединение или пересечение) для определения их семантической близости не даёт результата: понятия (точнее, обозначающие их знаки), выделенные в сопоставляемых текстах, не пересекаются. Определённый выход из этой ситуации заключается в приведении лексики построенных онтологий к лексике согласованной понятийно-терминологической структуры (например, тезауруса).

Представление систем, входящих в определение онтологии, в виде графов позволит использовать при формализации выполнения операций над онтологиями аксиомы теоретико-графовых операций. В качестве основных операций используются бинарные операции объединения и пересечения и унарные - построения аспектного представления и масштабирования онтологий, с помощью которых можно, в том числе, синтезировать новые онтологии (так называемые прикладные онтологии), отражающие ПрО в заданном аспекте. Отличительной особенностью реализации операций над онтологиями является возможность использования для повышения семантической связности структур понятийного и терминологического уровней [18].

На рисунке 1 представлен фрагмент результата операции объединения онтологий, построенных по текстам конструкторских документов ПрО «Атомная энергетика»: «Компоновка основного оборудования», «Прочность и сейсмостойкость», «Внутрикорпусные устройства», «Технические решения при модернизации реакторной установки», «Источники излучения». Общие узлы, входящие в нескольких онтологий функционального уровня, выделены на рисунке темно-серым фоном, а узлы, соответствующие понятийной системе – серым. Тезаурусные связи обозначены: *RT* (Related Term) – связь с ассоциативным дескриптором, *BT* (Broader Term) – связь с вышестоящим дескриптором, *NT* (Narrower Term) – связь с нижестоящим дескриптором. В результате использования в качестве общего понятийного базиса тезауруса INIS МАГАТЭ [19] удалось повысить семантическую связность объединения путём добавления тезаурусных маршрутов.

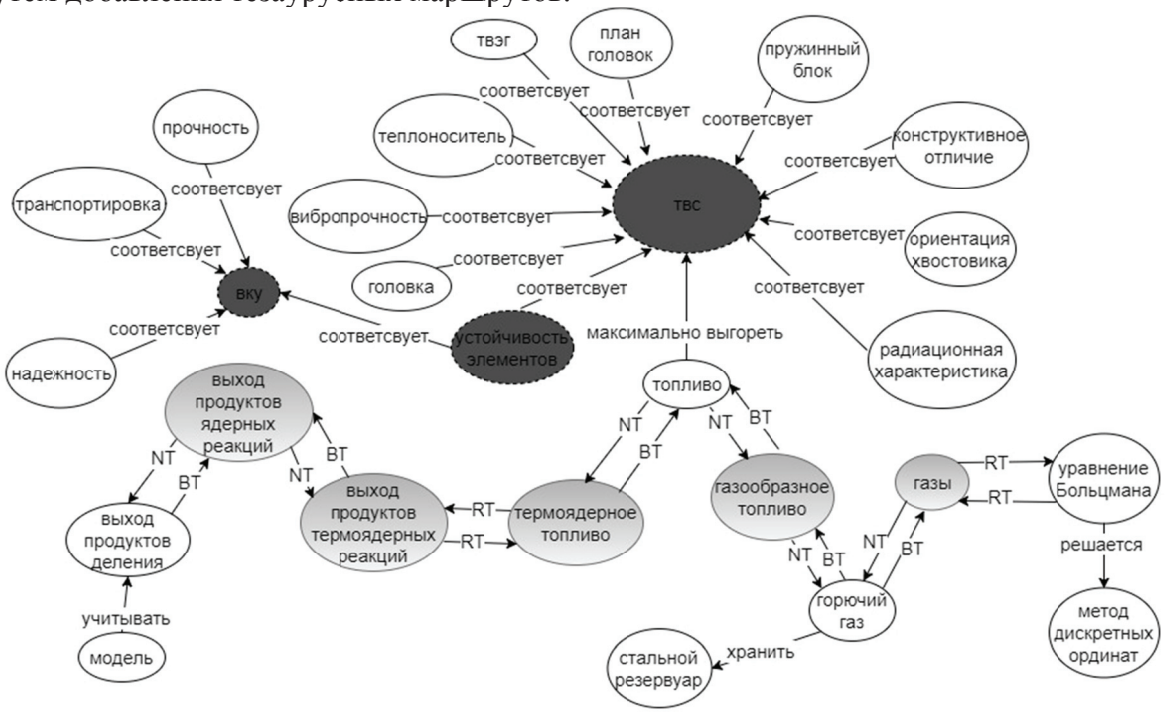


Рисунок 1 – Фрагмент графа - результата операции объединения онтологий

Следует отметить, что при построении онтологии не всегда удаётся передать контекст конкретного термина, если на структурном уровне этот термин (как отдельная вершина) не будет связан доопределяющим термином. Для преодоления подобных проблем предлагается использовать взвешенный ориентированный метаграф

$$MG_f = \langle V_f, V_f^M, X_f \rangle,$$

где V_f - множество вершин (терминов); V_f^M - множество метавершин, каждая из которых в свою очередь может быть метаграфом; X_f - множество дуг (отношений), определённых на множестве $V_f \cup V_f^M$, и $\forall x_i \in X_f : x_i = \{vb_i, ve_i, \langle tr_i, A_i \rangle\}$, где $tr_i \in TR$ (TR - множество типов отношений), A_i - множество характеристических атрибутов отношений, соответствующих дугам.

На рисунке 2 приведён пример метаграфа, построенного для следующего фрагмента текста: «*Клапаны на вертикальных сосудах следует устанавливать на верхнем днище. Клапаны не допускается использовать для регулирования давления в сосуде или группе сосудов. Изготовитель обязан поставлять клапаны с паспортом и руководством по эксплуатации*».

Выше отмечалось, что онтология обладает свойствами системы. Определяя системный базис - подмножества свойств и типов отношений, соответствующих аспекту, или функцию отображения на другую онтологию, можно построить подсистемы и проекции, которые сами являются системами. Это означает, что приведённое выше графовое представление онтологии обладает характерным свойством мультиграфов. Отметим также, что поскольку пара объектов ПрО может быть связана несколькими типами отношений, то граф можно классифицировать как гиперграф.

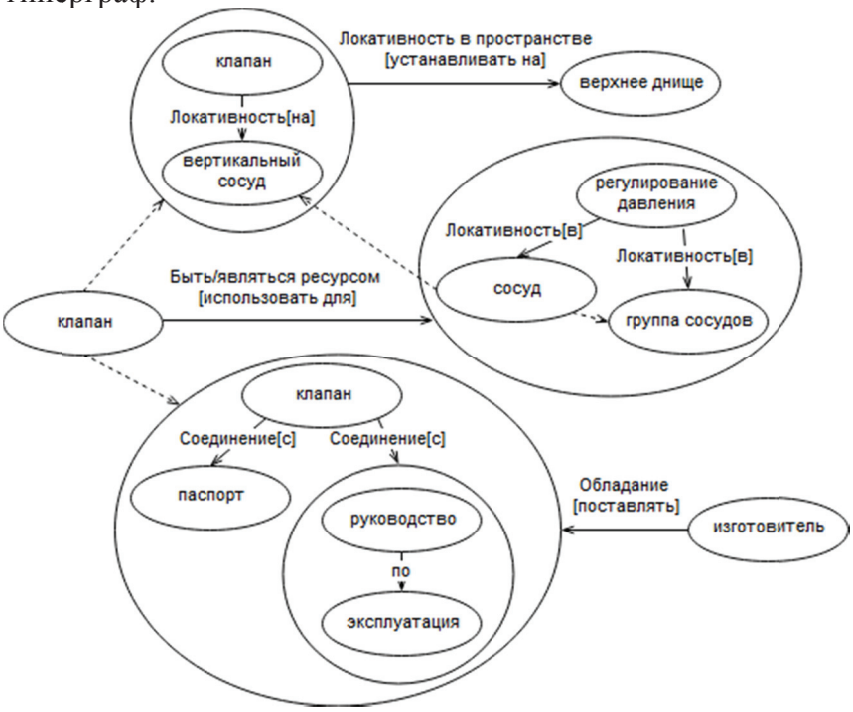


Рисунок 2 - Пример метаграфа с типизированными отношениями (в квадратных скобках приведена лингвистическая конструкция из текста, пунктирными линиями обозначены иерархические связи)

5 Архитектура и функциональные подсистемы

ИТ-среда всегда отличалась своей гетерогенностью, несмотря на многочисленные попытки её глобальной стандартизации. Современная тенденция, диктуя свои требования, за-

ставляет отвечать на подобный вызов созданием кроссплатформенных систем, способных работать с одинаковой эффективностью независимо от ИТ-инфраструктуры конечного пользователя. Одним из ответов на подобное требование является агентный принцип и наличие развитого программного интерфейса, позволяющего гибко и эффективно подстраиваться под текущие пользовательские тренды и ставшего неотъемлемой частью сложных систем. В частности, платформа xIRBIS-ML позволяет интегрировать лингвистические компоненты нескольких производителей.

Основными отличиями xIRBIS-ML являются:

- семантическая интеграции информации, относящейся к разным видам деятельности и всем этапам ЖЦ, при которой информационные объекты рассматриваются как имеющие двойное назначение: порождаются и используются в целевой деятельности пользователя и являются метаинформацией в процессах информационного обеспечения этой деятельности;
- ориентированное на задачи когнитивного поиска многоаспектное индексирование разнородной информации, в том числе с использованием взаимодополняющих технологий лингвистических процессоров и эвристических методов, обеспечивающих глубокое семантическое индексирование и последующее выделение аспектов (динамическое индексирование);
- комплексное использование разнотипных методов и моделей отбора документов;
- гибкий поисковый интерфейс, ориентированный на задачи анализа и синтеза информации;
- объединённая объектно-ориентированная модель, совмещающая как данные информационных процессов, так и метаданные, определяющие параметры обработки.

5.1 Основные подсистемы xIRBIS-ML

Подсистема сбора *и обработки документов* обеспечивает импорт документов с преобразованием в унифицированный хорошо структурированный формат путём выделения структурных элементов, а также многоаспектное индексирование содержания, в том числе с возможностью интерактивного редактирования автоматически приписываемых классификационных кодов, списков ключевых слов и онтологий. Построение многоаспектного понятийного образа документа производится путём представления содержания в виде гиперграфа извлекаемых из текста понятий и типизируемых отношений, что позволяет не только динамически формировать онтологию знаний и выявлять новые понятия и связи, но также перейти к графически управляемым навигационным технологиям концептуального поиска. В итоге это обеспечивает более точный поиск, а интерактивные технологии индексирования – извлечение экспертных знаний для актуализации семантических компонентов.

Подсистема информационного поиска в распределённых ресурсах, помимо классических механизмов поиска по чётким/нечётким критериям и с реформулированием запроса по обратной связи, также обеспечивает переадресацию и адаптацию запроса для проведения поиска во внешних Интернет-ресурсах с учётом их особенностей, в том числе синтаксиса поисковых языков, что обеспечит более точный индивидуализируемый отбор информации.

Подсистема статистического анализа документальных потоков и лексики обеспечивает формирование распределений различных информационных срезов, а также построение и комплексный анализ временных рядов профилированных потоков документов и лексики. Это позволяет средствами *комплексного* дистрибутивно-статистического анализа потоков документов и запросов выявлять тенденции, источники и взаимозависимости, свойственные ПрО. Для отображения и анализа используются компоненты деловой графики.

Подсистема анализа и ведения лингвистического обеспечения ориентирована на поддержку пользовательского информационного пространства и обеспечивает построение и ведение иерархических терминологических (словарных и классификационных) структур ПрО, которые могут быть использованы в качестве средств систематизации ПрО, а также для автоматической классификации документов. Динамическое обновление состава и связей терминов основано на лингвистическом и статистическом анализе текстов документов и баз данных.

Подсистема мониторинга хода выполнения и анализа содержания проектов реализует: мониторинг показателей по всем этапам ЖЦ проекта; статистический анализ и систематизацию информацию о проекте и ассоциированных ПрО в локальной БД, а также в других ИР; автоматическое формирование сводных статистических портретов основных действующих лиц и т.п.

5.2 Объектная модель xIRBIS-ML

Основу интеграции информационной среды и управления поисковой навигацией xIRBIS-ML составляет объектная модель информационной среды, включающая расширенную объектную модель документа. Структура среды определяется трёхзвенностью совокупной ИС «*пользователь – автоматизированная ИС (АИС) – ИР*», где АИС, как промежуточное звено, должна обеспечивать согласование сторон – пользователя и ИР – каждая из которых имеет свою специфику организации и поведения.

Объектная модель трёхзвенной информационной среды специфицирует взаимодействие рабочего пространства пользователя (объект КР) и пользовательского интерфейса системы (задаваемого матрицей взаимосвязи функций поиска и обработки с интерфейсными операционными объектами).

Информационный ресурс, как объект рабочего пространства, характеризуется тремя группами параметров, определяющими:

- собственно ресурс (идентификация, адрес, метод доступа, тип);
- содержание ИР (тематика и характер ресурса, а также метаданные об элементах данных, документах и справочных ресурсах);
- взаимодействие с ИР (поисковый вход и метод; синтаксис поискового языка; метод доступа к содержанию).

Базовым объектом является документ, характеризующийся формой представления (так называемой схемой документа), которых может быть несколько. Документ представляет собой иерархию элементов данных, для каждого из которых специфицируются: идентификатор(ы), тип, метод(ы) преобразования на входе/выходе, метод(ы) индексирования.

Заключение

Семантическое ядро, назначением которого является обеспечение семантически согласованной идентификации, поиска и анализа информации и знаний, в итоге должно обеспечивать «содержательное» отождествление текстов, сходных по смыслу, но вариантно представленных (описаний объекта в разных аспектах, разными словами, в разных формах и т.п.). В рамках сформулированных выше предложений это достигается использованием системного подхода: состоявшееся знание (факты) представлено онтологиями разного уровня, а его структура - таксономиями и классификациями. Причём онтология – это, в соответствии с семиотической моделью, система трёх взаимосвязанных систем (функциональной, понятийной и знаковой), а система таксономий представляет классы объектов и процессов, характерных для основных «координат» деятельности. Исходя из того, что новое знание всегда связано с

известным, с помощью введённых операций над онтологиями становится возможным приведение описаний (их онтологических образов) к сопоставимым формам. Процесс конструирования нового знания, где знание – это информация, связанная с контекстом, реализуется средствами информационного поиска как компонента семантического ядра.

В рамках предложенных решений контекст явно или неявно определяется пользователем с помощью либо семантических структур (таксономий, онтологий ПрО), либо динамических компонентов (словников, КР), что позволяет говорить о полноценном семантическом поиске, когда система не только реализует более или менее интеллектуальный отбор документов, но и формирует выражения самой информационной потребности (в терминах, понятных системе), что позволит синтезировать *комплексные*, аспектно-ориентированные ответы.

Изложенные положения и решения семантического ядра в той или иной степени были использованы при разработке ряда опытных образцов и промышленных систем.

Список источников

[1] Государственная программа «Цифровая экономика Российской Федерации». Утверждена распоряжением Правительства Российской Федерации от 28 июля 2017 г. № 1632-р.

[2] ГОСТ Р 55062-2012 Информационные технологии (ИТ). Системы промышленной автоматизации и их интеграция. Интероперабельность. Основные положения. Утвержден и введен в действие Приказом Федерального агентства по техническому регулированию и метрологии от 13 ноября 2012 г. № 751-ст**.** Дата введения 2013-09-01.

[3] ***Kostiha, F***. Knowledge management at Czech nuclear power plants / F. Kostiha // *Bezpecnost Jaderne Energie*, 2010, 18(7-8), P.203-209.

[4] ***Pryakhin, A***. Fast reactor knowledge inventory / A. Pryakhin, A. Stanculescu, Y. Yanev // *International Journal of Nuclear Knowledge Management*, 2009, 3.2, P.199-209.

[5] Технология анализа и понимания текстов на естественном языке ABBYY Compreno. - https://www.abbyy.com/ru-ru/infoextractor/compreno/.

[6] Инструментарий компьютерного анализа текстовой информации RCO Fact Extractor. - http://www.rco.ru/

[7] ***Голенков, В.В.*** Проект открытой семантической технологии компонентного проектирования интеллектуальных систем. Часть 1: Принципы создания / В.В. Голенков, Н.А. Гулякина // *Онтология проектирования*. – 2014. - №1(11). – С. 42-64.

[8] ***Левенчук, А.*** Онтология, схема/онтология инженерного проекта и схемное/онтологическое мышление. - https://ailev.livejournal.com/1159110.html.

[9] ***Максимов, Н.В.*** Структура и компоненты операционного визуального пространства интерактивного поиска научной информации / Н.В. Максимов, О.Л. Голицына, А.Л. Усенко // *Научная визуализация.* – 2014. – Т. 6, № 4. – С. 96-106.

[10] ***McGuinness, D.L.*** Ontologies Come of Age. In In D. Fensel, J. Hendler, H. Lieberman, & W. Wahlster (Eds.), Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, MIT Press, 2002. - P.171-191.

[11] ***G. Van Heijst, A. Th. Schreiber, and B. J. Wielinga.*** Using Explicit Ontologies in KBS development. *International Journal of Human-Computer Studies*. 1997; Vol. 46, Iss. 2-3, P.183-292.

[12] WordNet. A Lexical Database for English. - https://wordnet.princeton.edu/.

[13] ***Лукашевич, Н.В.*** Проектирование лингвистических онтологий для информационных систем в широких предметных областях / Н.В. Лукашевич, Б.В. Добров // *Онтология проектирования.* – 2015. - Т.5 - №1(15). – С. 47-69.

[14] FrameNet Project. - https://www.framenet.icsi.berkeley.edu/fndrupal/.

[15] ***Голицына, О.Л.*** Онтологический подход к идентификации информации в задачах документального поиска. / О.Л. Голицына, Н.В. Максимов, О.В. Окропишина, В.И. Строгонов // *Научно-техническая информация.* Сер. 2. – 2012. – № 5. – С.1-9.

[16] ***Golitsina O.L., Maksimov N.V., Okropishina O.V., Okropishin A.E***. Semantic Identification of Text in the Class of Tasks of Information Retrieval // Proceedings of ICAI'14, WORLDCOMP'14, July 21-24, 2014, Las Vegas, Nevada, USA. CRSEA Press, USA. – 2014. – Vol. I. – P.47-52. – DOI: 10.3103/s0005105512030028.

[17] ***Голицына, О.Л.*** Онтологический подход к идентификации информации в задачах документального поиска: практическое применение / О.Л. Голицына, Н.В. Максимов, О.В. Окропишина, В.И. Строгонов // *Научно-техническая информация.* Сер. 2. – 2013. – № 3. – С.1-8. – DOI: 10.3103/S0005105513020027.

[18] ***Максимов, Н.В.*** Методологические основы онтологического моделирования документальной информации // *Научно-техническая информация*. Серия 2: – 2018. – №3. – С.6-22. – DOI: 10.3103/S0005105518020036.
[19] INIS/ETDE Thesaurus. - https://inis.iaea.org/inis/m/products-services/thesaurus/index.html.

SEMANTIC CORE OF DIGITAL PLATFORM

N.V. Maksimov[1,a], O.L. Golitsina[1,b], M.G.Ganchenkova[1,c], D.V. Sanatov[2,3,d], A.V. Razumov[3,e]

[1] *High Engineering School of National Research Nuclear University MEPhI, Moscow, Russia*
[2] *Media-Lab LLC, Moscow, Russia*
[3] *Center for Strategic Research "North-West", St. Petersburg, Russia*
[a]*nv-maks@yandex.ru,* [b]*olgolitsina@yandex.ru,* [c]*MGGanchenkova@mephi.ru,* [d]*dmsa@yandex.ru,*
[e]*a.razumov@csr-nw.ru*

Abstract

This article gives a generalized description of the integrated complex of declarative and procedural means, which together provide a consistent information (semantic) representation of complex objects at all stages of their life cycle. Knowledge (facts) is represented by a system of ontologies of different levels, and the structure - by taxonomies and classifications. Ontology is defined as a system of three interrelated systems (functional, conceptual and sign), and the system of taxonomies represents classes of objects and processes' characteristic of the main "coordinates" of activity. Information retrieval, as a component of the semantic core, is considered as a complex self-consistent process of constructing new knowledge, where knowledge is information (texts of found documents) associated with the context of the task and the user's views. Such context is purposefully or indirectly determined by the user by means of predefined semantic structures (taxonomies, ontologies) or by means of dynamically formed components (dictionaries, samples, etc.). This is the essence of the semantic cognitive search, where the system not only implements the selection of documents by traditional retrieval methods, but also forms an image of information needs, which, in turn, will allow the system to synthesize complex, aspect-oriented answers. Automated technologies of linguistic support based on distributive-statistical analysis of both object knowledge and implicit knowledge flows (extracted by the system in the interaction process) and components of conceptual-terminological systems are proposed. Scientific and educational potential can be widely used to create a basic component of information and linguistic support in the early stages. The tools presented in the article are tested in the framework of the developed software xIrbis-ML, designed for semantic search of data in knowledge bases for complex engineering objects.

Key words: *semantic core, digital platforms, semantic search, ontology.*

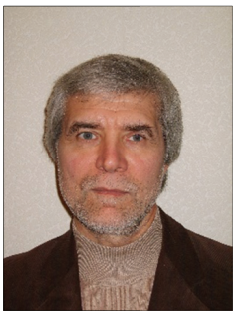
Citation: *Maksimov NV, Golitsina OL, Ganchenkova MG, Sanatov DV, Razumov AV.* Semantic core of digital platform [In Russian]. *Ontology of designing.* 2018; 8(3): 412-426. - DOI: 10.18287/2223-9537-2018-8-3-412-426.

References

[1] State Program «Digital Economy of the Russian Federation» [In Russian]. Approved by Russian Federation Government Executive Order dated July 28, 2017 No. 1632-r.
[2] GOST R 55062-2012 Industrial automation systems and integration. Interoperability. General position. [In Russian] Approved and put into action by Order of the Federal Agency for technical regulation and metrology dated November 13, 2012. N 751-st. Date of implementation 2013-09-01.
[3] ***Kostiha, F***. Knowledge management at Czech nuclear power plants. *Bezpecnost Jaderne Energie*, 2010; 18(7-8): 203-209.
[4] ***Pryakhin A, Stanculescu A, Yanev Y.*** Fast reactor knowledge inventory // *International Journal of Nuclear Knowledge Management*, 2009, 3.2: 199-209.
[5] Technology of analysis and understanding of natural language texts ABBYY Compreno. - https://www.abbyy.com/ru-ru/infoextractor/compreno/.
[6] Tools for computer analysis of text information RCO Fact Extractor. - http://www.rco.ru/.
[7] ***Golenkov VV, Guliakina NA***. The project is an open semantic technology component of designing intelligent systems. Part 1: Principles of creation [In Russian] *Ontology of designing*. – 2014; 1(11): 42-64.

[8] ***Levenchuk A.*** Ontology, scheme / ontology of engineering project and schematic / ontological thinking [In Russian]. - https://ailev.livejournal.com/1159110.html.
[9] ***Maksimov NV, Golitsina OL, Usenko AL***. The structure and components of the operational visual space for scientific interactive information retrieval. *Scientific visualization*. – 2014; 6(4): 72-95.
[10] ***McGuinness D.*** Ontologies Come of Age. In In D. Fensel, J. Hendler, H. Lieberman, & W. Wahlster (Eds.), Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, MIT Press, 2002. - P.171-191.
[11] ***G. Van Heijst, Schreiber ATh, Wielinga BJ.*** Using Explicit Ontologies in KBS development. *International Journal of Human-Computer Studies*. 1997; 46(2-3): 183-292.
[12] WordNet. A Lexical Database for English. - https://wordnet.princeton.edu/.
[13] ***Loukachevitch NV, Dobrov BV.*** Developing linguistic ontologies in broad domains [In Russian]. *Ontology of designing*. – 2015; 5(1): 47-69.
[14] FrameNet Project. - https://www.framenet.icsi.berkeley.edu/fndrupal/.
[15] ***Golitsina OL, Maksimov NV, Okropishina OV, Strogonov VI.*** The ontological approach to the identification of information in tasks of document retrieval. *Automatic Documentation and Mathematical Linguistics*. – 2012; 46(3): 125-132. – DOI: 10.3103/s0005105512030028.
[16] ***Golitsina OL, Maksimov NV, Okropishina OV, Okropishin AE.*** Semantic Identification of Text in the Class of Tasks of Information Retrieval // Proceedings of ICAI'14, WORLDCOMP'14, July 21-24, 2014, Las Vegas, Nevada, USA. CRSEA Press, USA. – 2014; I: 47-52. - DOI: 10.3103/s0005105512030028.
[17] ***Golitsina OL, Maksimov NV, Okropishina OV, Strogonov VI.*** An ontological approach to information identification in tasks of document retrieval: A practical application. *Automatic Documentation and Mathematical Linguistics*. – 2013; 47(2): 45-51. – DOI: 10.3103/S0005105513020027.
[18] ***Maksimov NV.*** The Methodological Basis of Ontological Documentary Information Modeling. *Automatic Documentation and Mathematical Linguistics*. – 2018; 52(2): 57–72. – DOI: 10.3103/S0005105518020036.
[19] INIS/ETDE Thesaurus. - https://inis.iaea.org/inis/m/products-services/thesaurus/index.html.

Сведения об авторах



Максимов Николай Вениаминович в 1975 г. окончил Московский инженерно-физический институт по специальности «Прикладная математика». Д.т.н. (2002г.). Профессор кафедры системного анализа МИФИ. Область научных интересов: моделирование и разработка документальных информационно-поисковых систем и баз данных, лингвистическое обеспечение документальных информационно-поисковых систем и систем управления знаниями; человеко-машинные информационные системы, интерфейсы на основе когнитивных и поведенческих моделей. Автор более 120 научных работ, учебников и учебных пособий.

Nikolai Veniaminovich Maksimov graduated from the Moscow Engineering Physics Institute in 1975, D. Sc. Eng. (2002). Professor of System Analysis Department of National Research Nuclear University MEPhI. Research interests: modeling and development of documentary information search systems and databases, linguistic support of documentary information search systems and knowledge management systems; human-machine information systems, interfaces based on cognitive and behavioral models. Author of more than 120 scientific papers, textbooks and tutorials.

Голицына Ольга Леонидовна в 1980 г. окончила Московский государственный университет по специальности «Прикладная математика». Кандидат технических наук (2004 г.). Доцент кафедры системного анализа МИФИ. Область научных интересов: моделирование и разработка документальных информационно-поисковых систем и баз данных, лингвистическое обеспечение документальных информационно-поисковых систем и систем управления знаниями; проектирование баз данных. Автор более 50 научных работ, учебников и учебных пособий.

Olga Leonidovna Golitsyna graduated from the Moscow State University in 1980, Cand.Sc. (2004). She is associate professor of System Analysis Department of National Research Nuclear University MEPhI. Research interests: modeling and development of documentary information search systems and databases; linguistic support of documentary information search systems and knowledge management systems; database design.

Ганченкова Мария Герасимовна в 1997 г. окончила Московский инженерно-физический институт. К.т.н. (2002 г.). С 1995 по 2000 гг. – внедрение ИТ-систем в российском финансовом секторе. С 2000 по 2014 гг. исследовательская работа в МИФИ (Россия), Королевском институте технологий (Швеция), Университете Аальто (Финляндия). С 2017 г. – директор Высшей инжиниринговой школы МИФИ. Более 70 научных работ. Область научных интересов: физика материалов, компьютерный инжиниринг, цифровые платформы.

Maria Gerasimovna Ganchenkova graduated from the Moscow State University in 1980. Can.Sc. (2002). From 1995 to 2000 – implementation of IT-systems in the Russian financial sector. From 2000 to 2014 active research work at MEPhI (Russia), Royal Institute of technology (Sweden), Aalto University (Finland). From 2017 - Director of Higher engineering schools MIFI. More than 70 scientific works. Research interests: physics of materials, computer engineering, digital platforms.

Санатов Дмитрий Васильевич в 2006 г. окончил Санкт-Петербургский государственный университет. В 2018 г. получил дополнительное образование в МГТУ им. Баумана по направлению анализа данных. С 2006 г. по настоящее время работает в сфере стратегического консалтинга. С 2017 г. реализует проекты коммерциализации цифровых технологий в компании Медиа-Лаб, созданной как спин-офф проект ЦСР «Северо-Запад». Область научных интересов: управление изменениями в корпоративном секторе и региональном развитии, цифровизация экономики, промышленные технологические платформы.

Dmitriy Vasilievich Sanatov graduated from the St.Petersburg State University in 2006. Now he is Deputy Director in the Center for strategic research “North-West” and CEO in Media-Lab, LLC. Research interests: change management in corporate sector and regional development, digitalization of economy, industrial technology platforms.

Разумов Антон Витальевич в 2002 г. окончил Московский государственный горный университет по специальности «Автоматизированные системы управления». С 2002 по 2017 гг. – развитие ИТ-инфраструктуры промышленных компаний, разработка программного обеспечения в среде Delphi, Power Builder, внедрение информационных систем уровня ERP и CRM. Область научных интересов: базы данных, системы автоматизированного проектирования.

Anton Vitalievich Razumov graduated from the Moscow State Mining University in 2002. From 2002 to 2017, he worked on the development of it infrastructure in industrial companies, software development in Delphi, Power Builder, implementation of information systems ERP and CRM. Research interests: databases, computer-aided design.