

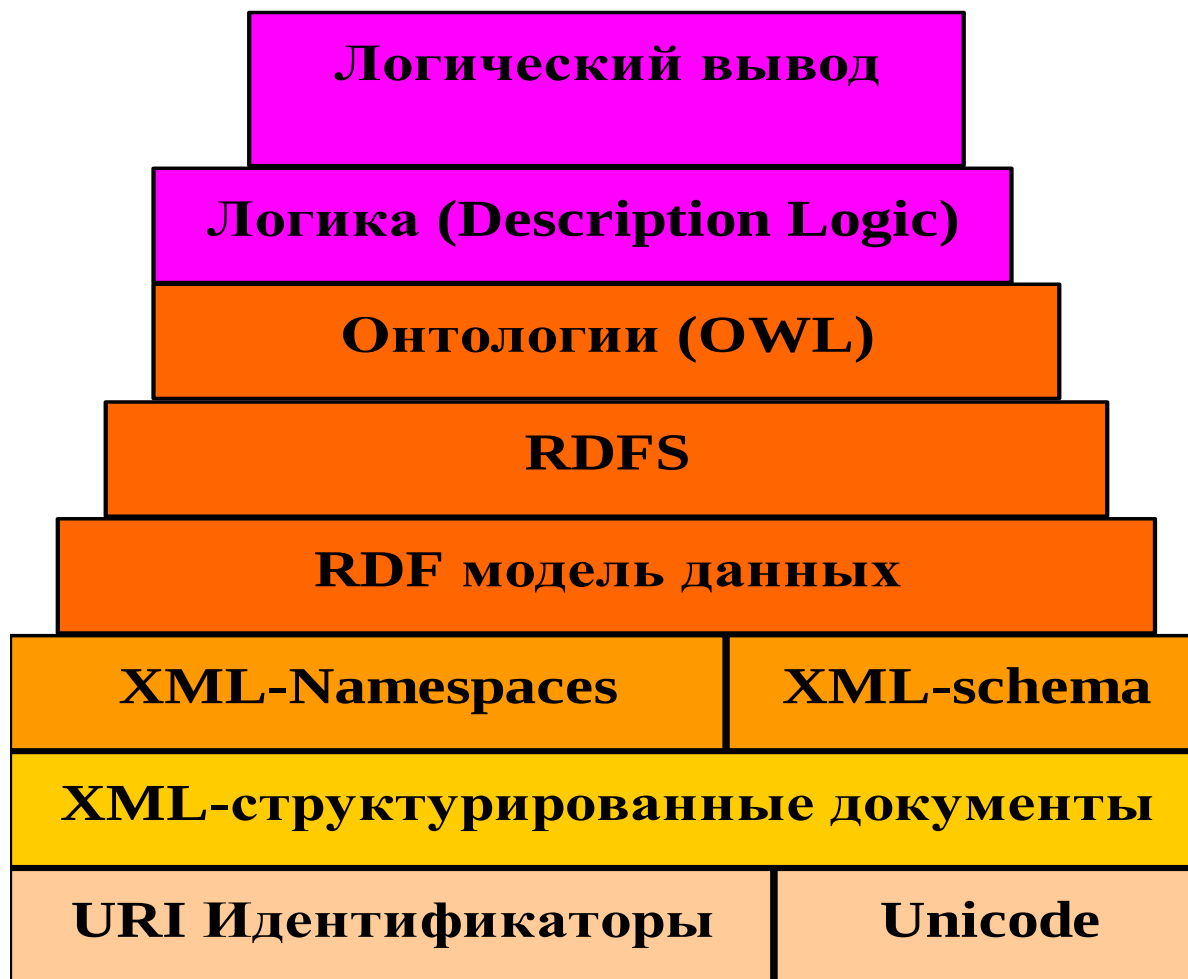
Опыт применения RDF технологий для описания общенаучной прикладной области

В.А.Серебряков

Содержание

- История вопроса
- Базовые технологии Semantic Web
 - RDF
 - RDFS
 - OWL
- Linked Open Data
- Информационная система «Научный институт РАН»
- ЕНИП РАН
- Цифровая семантическая библиотека LibMeta
- Научное Наследие
- Пространство знаний

Пирамида Semantic Web



RDF

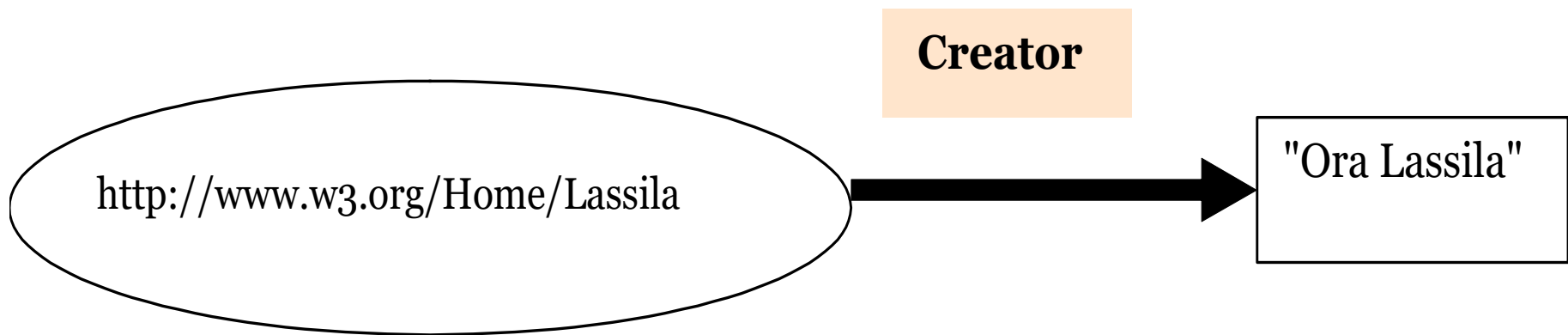
- Resource Description Framework (RDF) является базой для обработки метаданных. Она обеспечивает интероперабельность между приложениями, обменивающимися машинно-понимаемой информацией через Web. RDF придает особое значение средствам, обеспечивающим автоматическую обработку Web ресурсов.

Формальная модель RDF

1. Имеется множество, называемое *Ресурсы*.
2. Имеется множество, называемое *Литералы*.
3. Имеется подмножество *Ресурсов*, называемое *Свойства*.
4. Имеется множество, называемое *Утверждения*, каждый элемент которого – это тройка вида
 $\{\text{предикат (свойство), субъект (ресурс), объект (ресурс)}\}$,
где предикат – это свойство (элемент *Свойств*), субъект – это ресурс (элемент *Ресурсов*) и объект – либо ресурс, либо литерал (элемент *Литералов*).

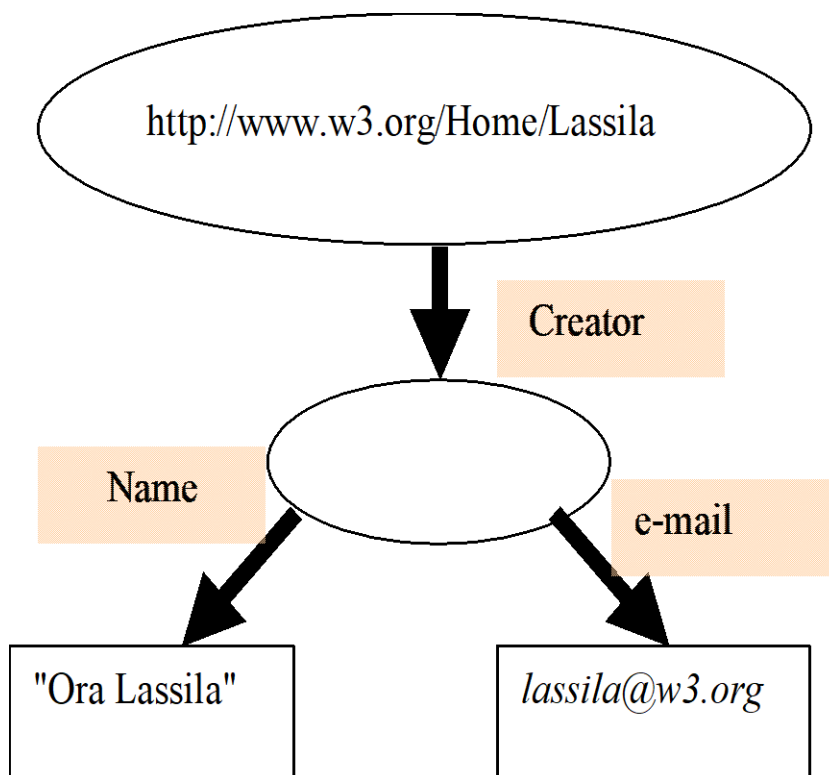
Простая вершина и дуга

Subject (Resource)	http://www.w3.org/Home/Lassila
Predicate (Property)	Creator
Object (literal)	"Ora Lassila"



Свойство со структурным значением

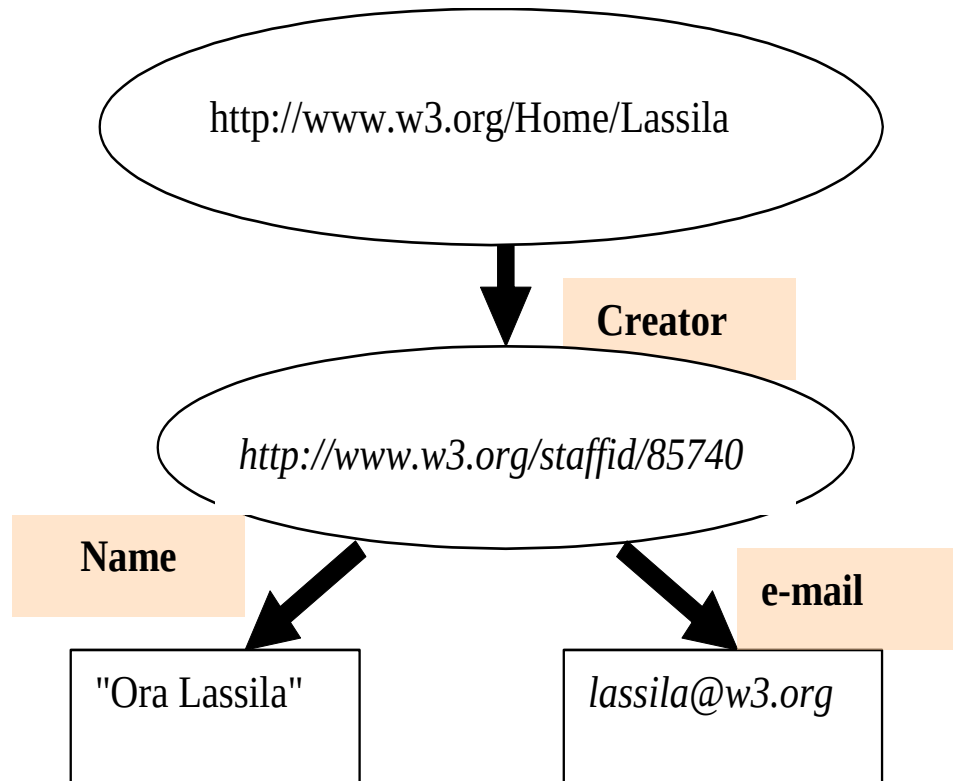
Персона с именем *Ora Lassila* имеет email lassila@w3.org и является создателем ресурса <http://www.w3.org/Home/Lassila>.



Структурное значение с идентификатором

Персона , на которую имеется ссылка

<http://www.w3.org/staffid/85740> имеет имя Ora Lassila и имеет почту lassila@w3.org. С этими свойствами создан ресурс <http://www.w3.org/Home/Lassila>.



Преимущества использования RDF модели данных

- С помощью HTTP URI, как глобального уникального идентификатора для данных и терминов словарей, RDF модель данных по своей сути предназначена для использования в глобальном масштабе и позволяет каждому ссылаться на все.
- Клиенты могут посмотреть любой URI в RDF графе в Интернете для получения дополнительной информации. Таким образом, каждая RDF тройка является частью глобальных веб-данных, и каждая RDF тройка может быть использована в качестве отправной точки для просмотра этого пространства данных.
- Модель данных позволяет установить RDF связи между данными из различных источников.

Пространство имён XML

Пространство имён в XML ([англ. XML namespace](#)) — это стандарт, описывающий именованную группу имён элементов и атрибутов, служащую для обеспечения их уникальности в [XML](#)-документе. Все имена элементов в пределах пространства имён должны быть уникальны.

XML-документ может содержать имена элементов и атрибутов из нескольких словарей XML. В каждом словаре задано своё пространство имён — так разрешается проблема неоднозначности имён элементов и атрибутов.

URI, URL, URN

URL (Uniform Resource Locator) – это строка символов, которая используется для идентификации какого-либо ресурса, но только по его адресу, по его местоположению.

URN служит для обозначения уникального имени ресурса, неважно, где этот ресурс располагается в данный момент времени или вообще. Такая природа URN (независимость от адреса) позволяет ресурсам перемещаться с одного места на другое. URN позволяет получить доступ к ресурсу по различным сетевым протоколам, обращаясь к одному и тому же имени.

URI является либо [URL](#), либо [URN](#), либо одновременно обоими.

URL — это URI, который, помимо идентификации ресурса, предоставляет ещё и информацию о местонахождении этого ресурса. А URN — это URI, который только идентифицирует ресурс в определённом [пространстве имён](#) (и, соответственно, в определённом [контексте](#)), но не указывает его местонахождение. Например, URN `urn:ISBN:0-395-36341-1` — это URI, который указывает на ресурс (книгу) 0-395-36341-1 в пространстве имён [ISBN](#), но, в отличие от URL, URN не указывает на местонахождение этого ресурса.

URN

пространство
имен

urn:isbn:4990223511

схема

уникальный идентификатор
ресурса

:issn:1082-9873

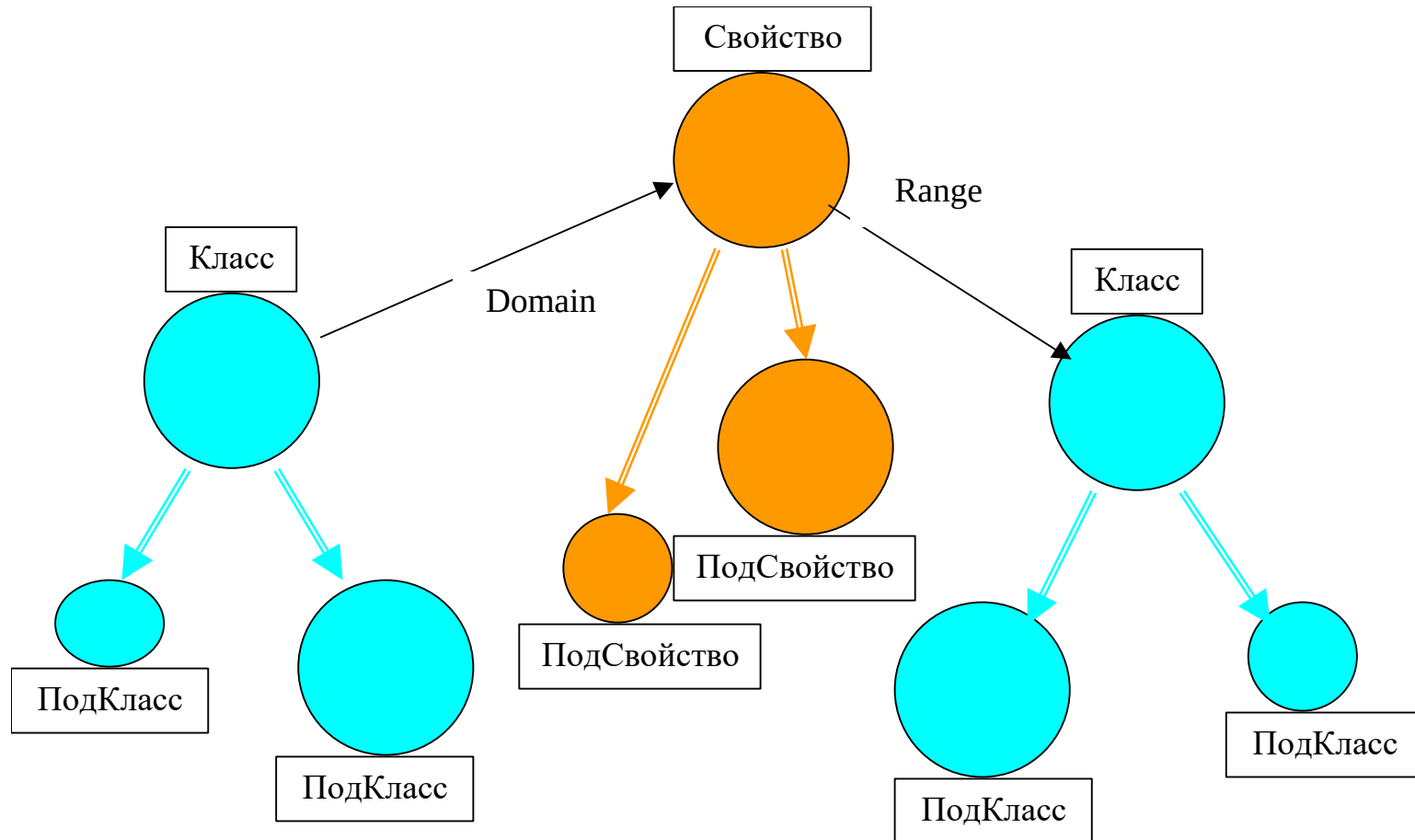
:doi:10.1000/1

:oasis:names:specification:docbook:dtd:xml:4.1

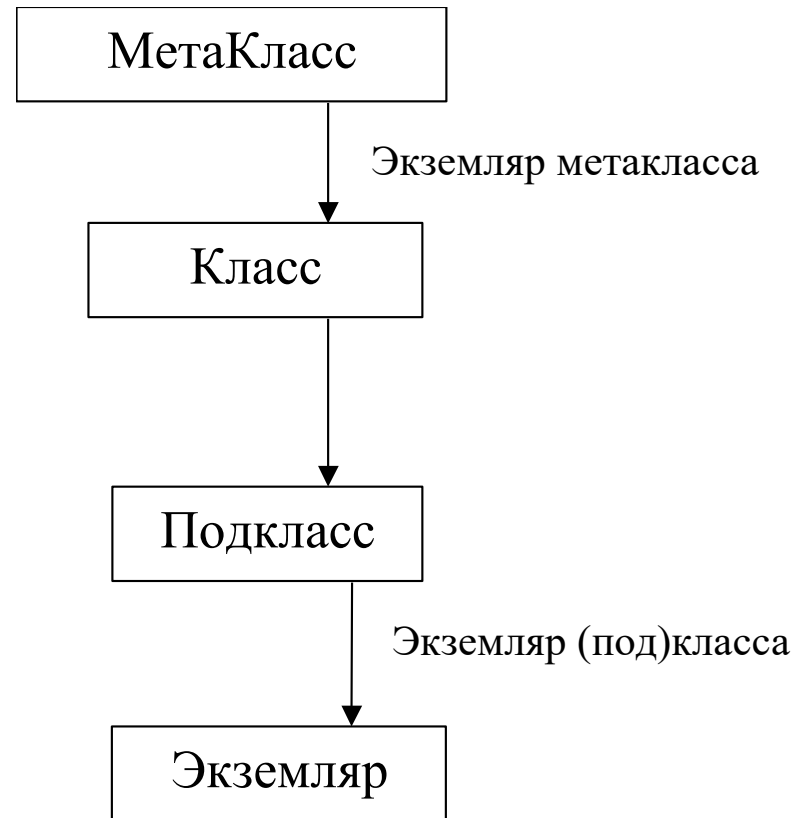
RDF Schema

Система типов позволяет ресурсам быть определенными как экземпляры одного или более классов. Кроме того, классы могут быть организованы иерархически (`rdfs:Class` и `rdfs:subClassOf`), которые используются при спецификации прикладных схем. Также и свойства могут быть организованы иерархически.

Классы и свойства



Метакласс, класс, подкласс, экземпляр



Онтология

"Онтология – это явная формальная спецификация разделяемой концептуализации предметной области. " [Tom Gruber, 1992г.]

Онтология дает подробную концептуализацию (т.е. мета информацию), которая описывает семантику (смысл) данных. Ее функция аналогична схеме базы данных.

Но:

- Язык определения онтологий синтаксически и семантически богаче, чем распространенные подходы в базах данных.
- Онтология должна быть разделяемой и содержать согласованную терминологию, т.к. она используется для совместного использования и обмена информацией.
- Онтология предоставляет теоретическое описание предметной области, а не структуру хранилища данных.

Онтологии и Web

Так как XML фактически является стандартом для обмена данными в Web, то желательно также обмениваться онтологиями, используя синтаксис XML, что упростит написание синтаксических анализаторов.

Это требование приводит в свою очередь к языкам, основанным на XML, т.е. определению языка с использованием XML.

Инструменты онтологий

- Работа с классами как множествами ресурсов
 - перечисление экземпляров
 - пересечение «и»
 - объединение «или»
 - дополнение «не»
 - не пересекающиеся «отделенные» классы
 - эквивалентность...
- «Анонимные» классы ресурсов (удовлетворяющие заданным ограничениям).
- Обратные свойства, транзитивность, симметричность, рефлексивность...
- Мощность, перечисление значений...
- Локальные ограничения на свойства в контексте классов...

Преимущества использования онтологий

Онтологии дают общий словарь. Они обеспечивают:

- **Интероперабельность** различных наборов данных путем использования разделяемых понимаемых машинами структур
- **Предоставляют концептуальную схему** набора данных вне зависимости от формата, структуры, сложности или размера
- Могут быть использованы **как базис для поиска и извлечения** данных на основе структуры контента
- Онтологии дают **семантическое описание** для программ моделирования, используемых в прикладных областях.
- Онтологии **поддерживают логический вывод**, что позволяет извлекать знания из имеющихся

Разработка онтологии

Разработка онтологии включает определение :

1. понятий предметной области как классов
2. связей между этими понятиями/классами
3. свойств этих понятий/классов
4. областей определения и областей значений свойств
5. ограничений на классы и свойства
6. создание базы знаний, содержащей экземпляры определенных понятий

Язык запросов SPARQL

SPARQL — [язык запросов к данным](#), представленным по модели [RDF](#), а также [протокол](#) для передачи этих запросов и ответов на них. SPARQL является рекомендацией [консорциума W3C](#) и одной из технологий semantic web. Предоставление SPARQL-точек доступа ([англ. SPARQL-endpoint](#)) является рекомендованной практикой при публикации данных в semantic web.

Точка доступа SPARQL — это служба, поддерживающая протокол запросов SPARQL. Точка доступа позволяет пользователю делать запросы к RDF базе. Сервер обрабатывает запрос SPARQL, переданный по HTTP, и возвращает ответ. Таким образом, точки доступа SPARQL являются [API](#) к RDF базам.

Linked Open Data

Создатель WWW и инициатор Semantic Web и Linked Open Data Tim Berners-Lee в 2006 сформулировал четыре принципа разработки Linked Open Data.

Основные принципы

•Первый принцип связанных данных

использование URI ссылок не только для идентификации веб-документов и цифрового контента, а также объектов реального мира и абстрактных понятий. Это могут быть материальные вещи, такие как люди, места и автомобили, и более абстрактные, такие, как отношения.

•Второй принцип связанных данных

использование HTTP URI, для идентификации объектов и абстрактных понятий, что позволяет этим URI быть разыменованными в соответствии с протоколом HTTP в описание выявленных объектов или понятий.

Основные принципы

- ***Третий принцип связанных данных***

использование в качестве единой модели данных для публикации структурированных данных Web - Resource Description Framework (RDF), простой основанной на графах модели, которая была разработана для использования в контексте веб.

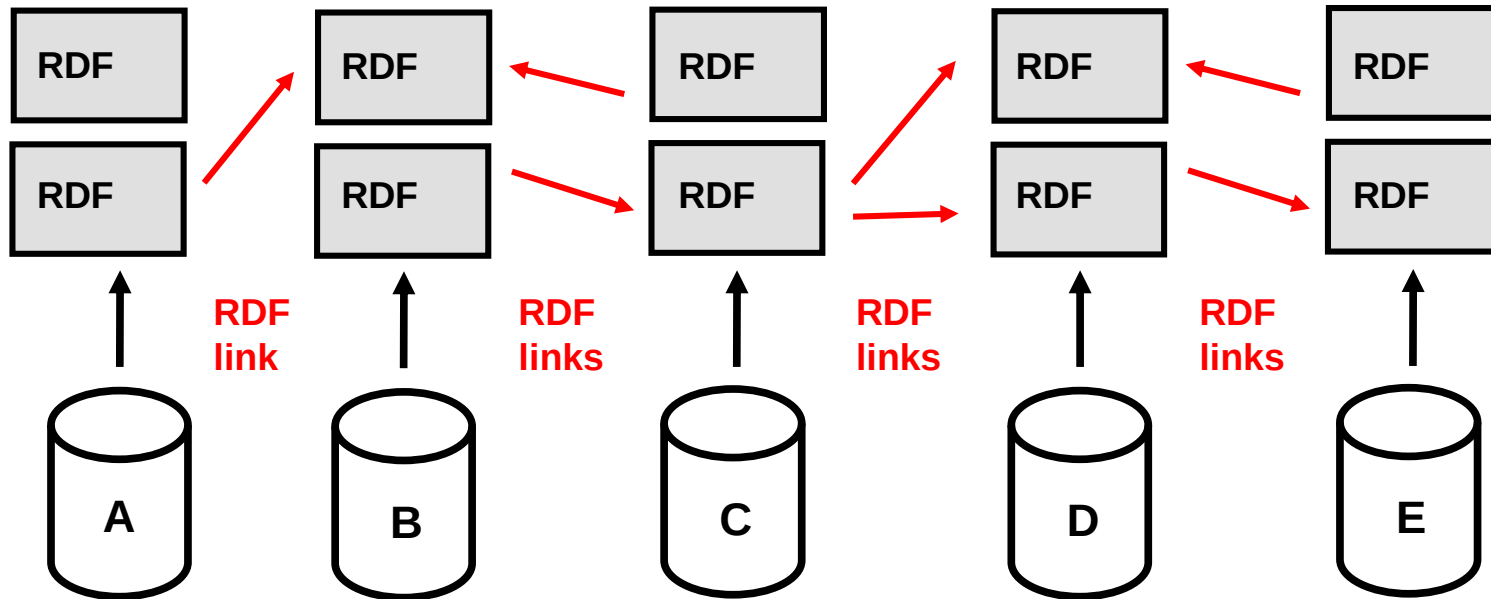
- ***Четвертый принцип связанных данных***

Включать RDF утверждения на другие URIs таким образом, чтобы можно было бы обнаруживать связанные вещи. Например, ссылка может быть установлена между человеком и местом, или между местом и компанией. В отличие от классического Web, где гиперссылки не типизированы, гиперссылки, которые соединяют вещи в контексте связанных данных, имеют типы.

Linked Open Data

Использование семантических технологий для

1. Публикации структурированных данных в Web,
2. Установления ссылок между данными из одного источника на данные из других источников.



LOD предоставляет

- ***Объединяющая модель данных.***

LOD опирается на RDF как единую, объединяющую модель данных. Обеспечивая глобальную уникальную идентификацию сущностей, а также позволяя использовать различные схемы для представления данных, RDF модель данных специально разработана для использования для глобального обмена данными. В отличие от других методов публикации данных в Интернете, она опирается на широкий спектр различных моделей данных.

- ***Стандартизированный механизм доступа к данным.***

LOD берет на себя способы использования протокола HTTP. Это соглашение позволяет источникам данных быть доступными с использованием обычных браузеров и позволяет использовать поисковые механизмы поисковиков. С другой стороны, Web APIs используются с помощью других подходящих интерфейсов.

LOD предоставляет

- ***Обнаружение данных на основе гиперссылок.***

С помощью URI как глобального идентификатора для сущностей, LOD позволяет устанавливать гиперссылки между сущностями в различных источниках данных. Эти ссылки связывают данные в едином глобальном пространстве данных и позволяют LOD приложениям обнаруживать новые источники данных во время выполнения.

- ***Само-описательные данные.***

LOD облегчает интеграцию данных из различных источников, опираясь на общие словари, делая определение этих словарей достижимым, а также путем предоставления терминов из различных словарей для связи с другими словарями по ссылкам.

ЕНИП РАН\НИ РАН

- Портал РАН
- НИ РАН
- ЕНИП РАН

Основной профиль метаданных

к основному профилю метаданных ЕНИП была отнесена общая поддержка следующих четырех основных групп информационных сущностей:

«Участники научной деятельности» – от «Персон», образующих разнообразные организационные объединения от формальных («Организации» и «Подразделения») до неформальных («Коллективы», «Сообщества», «Рабочие группы»).

- «Научная деятельность», в частности, «Проекты», а также «Научные мероприятия» – как разовые, так и повторяющиеся, такие как «Конференции», «Семинары», «Симпозиумы».
- «Результаты научной деятельности», в которые могут входить «Веб-сайты», «Базы данных», «Экспериментальные данные» и их «Математические модели», «Программные системы», в частности, «Научные вычислительные приложения», «Экспериментальные установки», «Изобретения», «Технологии», и т.п.
- «Документы и публикации» – научные труды, статьи, отчеты, диссертации сотрудников, возможно, административные постановления. Примерами специализации публикации могут служить, например, «Тезисы конференций», «Диссертации» и т.п.

Прикладные профили метаданных

- Библиографическая информация
- Информация о разработках и продуктах
- Математическая информация
- Конференции и поддержка проведения конференций
- Коллекции и архивы
- Поддержка библиотечной деятельности

Классы Основного профиля метаданных

- Класс *Лицо* - субъект деятельности, то есть действующее лицо проектов и мероприятий, автор или участник разработок и публикаций. К этому классу относятся персоны, группы и коллективы, организационные единицы.
 - Подкласс *Персона* – описывает разного рода метаинформацию о людях.
 - Подкласс *Организационная единица* - преКласс представляет организационные единицы.
 - Подкласс *Организация* - организация в юридическом смысле как частный случай организационной единицы.
 - Подкласс *Подразделение* - организационная единица, являющаяся частью некой организации в юридическом смысле.
- Класс *Должность* – эта структура представляет должность некоторого лица в организации, проекте, рабочей группе и т.д. Для должностей указывается название должности, должностное лицо.
 - Подкласс *Штатная должность* - должность (штатная единица), занимаемая некоторым лицом в организации или подразделении, с указанием работодателя, сотрудника, названия должности и рабочей контактной информации.
- Класс *Деятельность* - Общий суперкласс для различных видов деятельности (проекты, мероприятия и т.д.). Деятельность имеет даты начала и окончания, исполнителей и описательные характеристики.
 - Подкласс *Проект* - описывает официально зарегистрированные проекты: планируемые, проводимые и завершённые.
 - Подкласс *Мероприятие* – описывает разного рода мероприятия, частным случаем которых являются конференции, семинары и т.д.
- Класс *Результат деятельности* - описывает разработки («продукты»), документы и прочие результаты деятельности. Различные расширения данного класса определяются специализированным профилем метаданных о разработках.
 - Подкласс *Документ* - к этому классу относятся разного рода документы и публикации, как печатные, так и цифровые. Документ рассматривается как частный случай разработки.
 - Подкласс *Web-документ* - документ, доступный из Сети и не являющийся официальной публикацией.

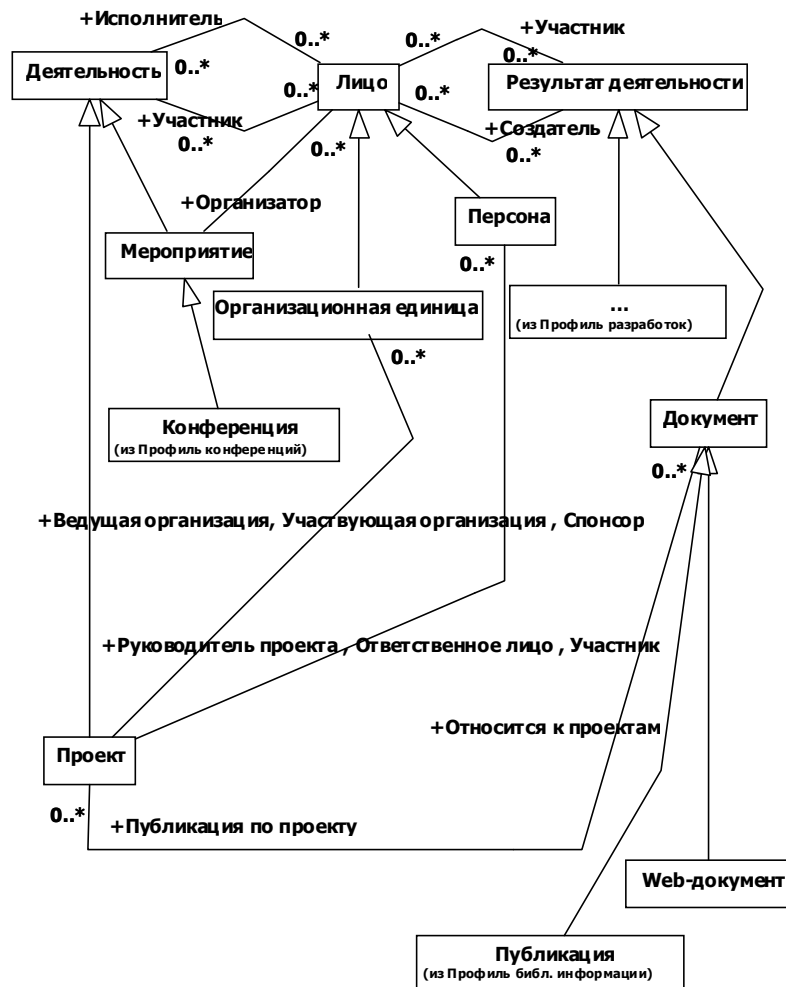
Основной профиль метаданных

- Лицо (организация или персона)
- Деятельность
- Результат деятельности
- Мероприятие
- Организационная единица
- Организация
- Подразделение
- Должность
- Персона
- Конференция
- Проект
- Документ
- Должность

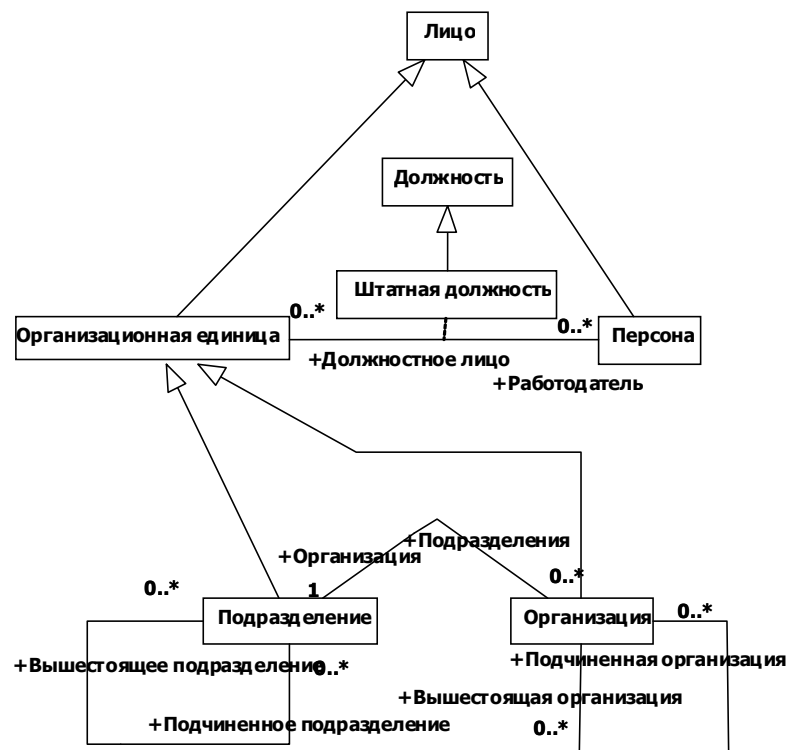
Примеры отношений (свойств)

- Исполнитель
- Участник
- Организатор
- Ведущая организация
- Участвующая организация
- Спонсор
- Участник
- Создатель
- Ответственное лицо
- Руководитель проекта
- Относится к проектам
- Публикация по проекту
- Должностное лицо
- Работодатель
- Организация
- Подразделение
- Вышестоящее подразделение
- Подчиненная организация
- Подчиненное подразделение

Основной профиль метаданных (1)



Основной профиль метаданных (2)



Специализированные/прикладные профили метаданных

Помимо базового набора элементов метаданных, предложения ЕНИП включают ряд специализированных/прикладных профилей метаданных, среди которых можно выделить следующие:

- *Библиографическая информация* - представление метаданных об официально зарегистрированных печатных изданиях (публикациях).
- *Информация о разработках и продуктах* – (в частности, программное обеспечение, Web-сайты и системы, базы и наборы данных, техническое обеспечение и оборудование).
- *Математическая информация* – поддержка специфики описания ресурсов, имеющих отношение к математике,
- *Конференции и поддержка проведения конференций* .

Специализированные/прикладные профили метаданных

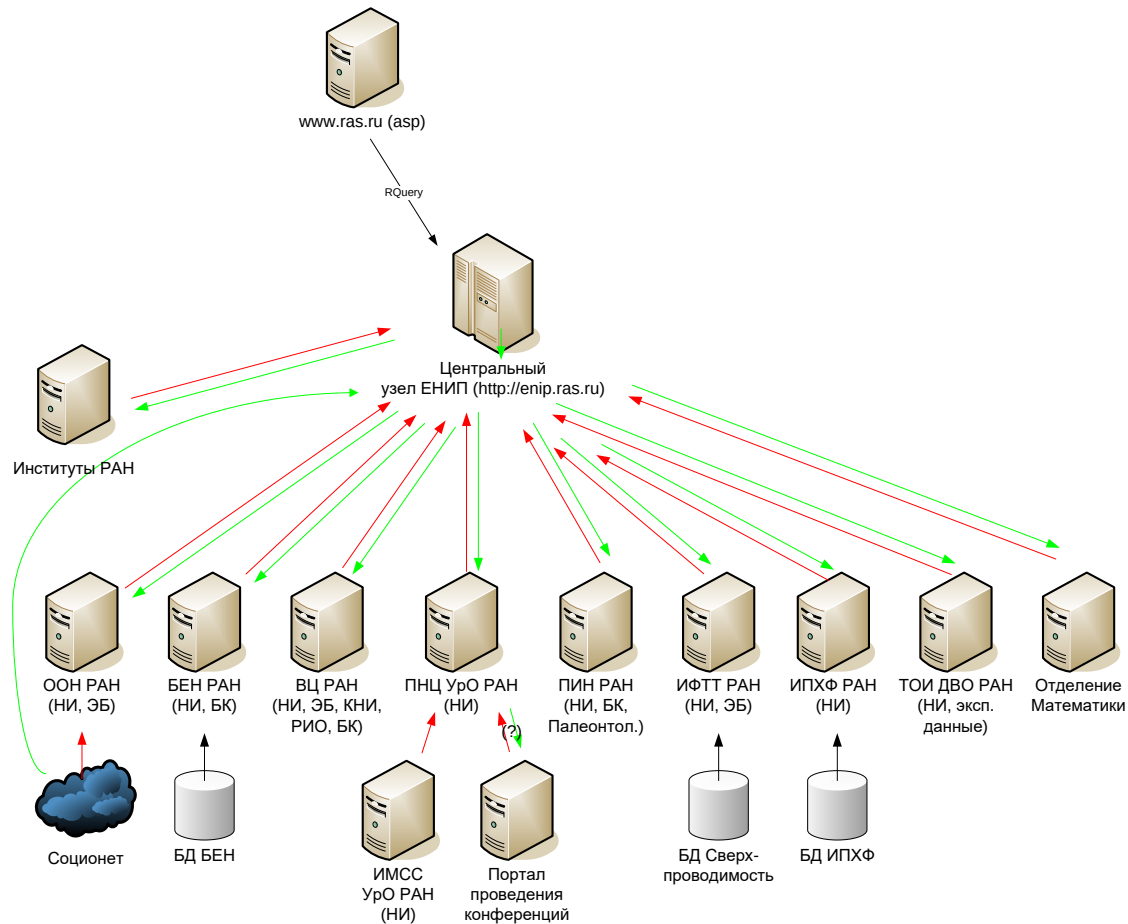
- *Коллекции и архивы* – поддержка описания коллекций, объединений физических и/или электронных элементов. К этому профилю относятся архивы и цифровые архивы, библиотечные и музейные каталоги и коллекции, эе коллекции текста, изображений, звуков, данных, программных продуктов и пр. (в частности, д специализированных/прикладных профилей метаданных специализированных/прикладных профилей метаданных иски, веб-подборки), разнообразные физические коллекции (картин и пр.), коллекции мероприятий (например, серии лекций).
- *Поддержка библиотечной деятельности* – поддержка информации для межбиблиотечного обмена и библиотечной деятельности; расширяет библиографическую специализацию понятием единиц хранения изданий (экземпляров изданий), а также библиотечных сервисов.
- *Поддержка издательской деятельности* – метаданные для описания плана издательства редакционно-издательского отдела организации; этот профиль может быть использован в качестве единого формата предоставления списка публикаций авторами и отделами организации в план издательства.

Блок математической специализации

Блок математической специализации предназначен для информационных систем математической специализации (сайты математических организаций, издательств, тематические ИС и другие). Включение этого блока обеспечивает следующие возможности:

- Поддержка математических формул. При выводе конечному пользователю значений этих атрибутов формулы могут просматриваться тремя способами:
 1. Формулы на языке TeX (для просмотра пользователями, знающими TeX).
 2. Формулы в MathML (для просмотра страницы в браузере с установленным MathML-плагином).
 3. Формулы в виде картинок, встроенных в HTML-текст.
- Поиск по формулам и их фрагментам через систему поатрибутной полнотекстовой индексации.
- Тематическая классификация ресурсов классификатором MSC.
- Идентификация публикаций их идентификаторами в системах реферирования MathSciNet и Zentralblatt MATH.
- Поддержка вторичных страниц Math-Net.

ЕНИП РАН



НИ – субпортал «Научный институт»
 ЭБ – субпортал «Электронная Библиотека»
 КНИ – субпортал «Каталог научной информации»
 РИО – субпортал «Редакционно-издательский отдел»
 БК – субпортал «Библиотечный отдел»

— XML-загрузка данных в портал
 — Репликация данных по протоколу ЕНИП
 — Распределенный поиск по протоколу ЕНИП

Цифровая семантическая открытая библиотека LibMeta

- Цифровая библиотека
 - обеспечивает поддержку использования медийных объектов (таких как текст, аудио, видео файлы или любой их комбинации) при описании своих объектов.
- Семантическая библиотека
 - обеспечивает поддержку различных типов ресурсов и связей между ними.
- Открытая библиотека
 - предоставляет возможность связывания своих данных с данными из открытых источников.

Семантическая библиотека LibMeta

- Основана на общих понятиях обеспечивающих функционирование системы;
- Представляет инструменты для:
 - отражения специфики предметной области и формирования ее концептуальной модели
 - Поддерживает работу с классификаторами и словарями предметной области;
 - работы с открытыми источниками данных;
 - спецификации данных ПО.
- Поддерживает настраиваемые интерфейсы работы с данными;
- Поддерживает навигацию по пространству знаний.

Функциональность инструмента

- Основная функциональность обеспечивает решение задач:
 - формирование многоуровневой модели предметной области;
 - интеграция источников данных;
 - предоставление пользовательских и машиночитаемых интерфейсов.
- Список функций:
 - создание/просмотр/редактирование
 - информационных ресурсов и их экземпляров,
 - тезауруса,
 - коллекций информационных объектов;
 - подключение источников данных;
 - пакетная загрузка данных, составляющих контент библиотеки;
 - атрибутивный/семантический/полнотекстовый поиск;
 - предоставление данных в машиночитаемом формате;
 - выделение дополнительных связей между экземплярами информационных ресурсов;
 - поддержка семантических меток для описания тематической направленности;
 - ...

Разработка онтологии

Трехуровневая модель онтологии ЦПНЗ

1 уровень. Концептуализация ЦПНЗ

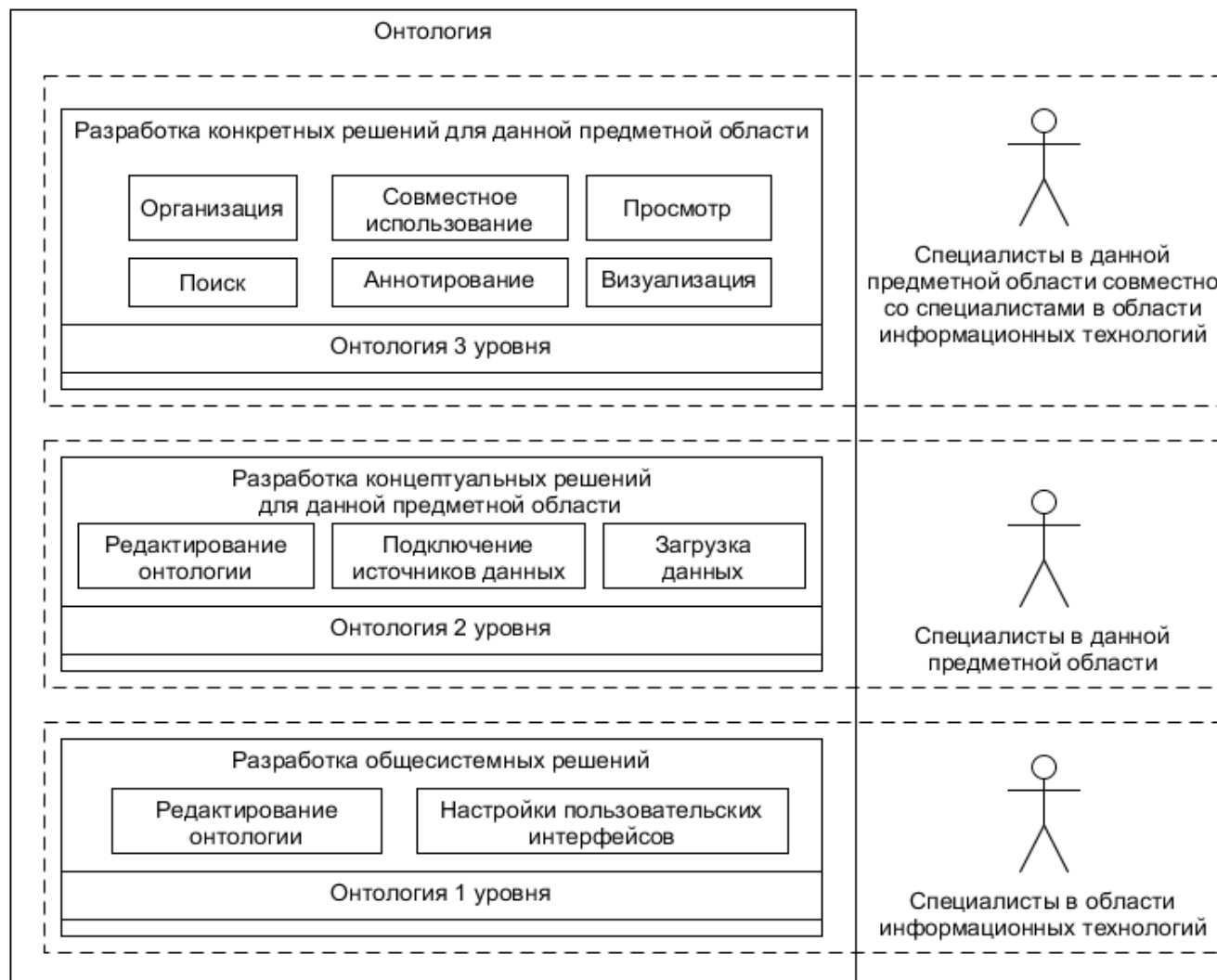
- *Информационный ресурс, Набор атрибутов, Атрибут, Тезаурус, Концепт,*

2 уровень. Концептуализация предметной области или спецификация предметной области в терминах 1 уровня

- *Информационные Ресурсы: Событие, Формула, 3d-модель, Персона, Публикация,*
- *Атрибуты: год публикации, индекс классификатора, автор, состоялось в, упоминается с персоной,*
- *Тезаурус: Тезаурус предметной области.*

3 уровень. Спецификация предметной области в терминах 2 уровня.

Уровни онтологии ЕЦПНЗ



Первый уровень онтологии (1)

На этом уровне описываются основные понятия. Первый уровень онтологии включает следующие понятия.

Информационный объект - цифровая копия объекта реального мира или специально созданный цифровой объект, отражающий определенные свойства реального объекта;

Тематическое подпространство – совокупность информационных объектов с их свойствами и связями, относящихся к некоторому научному направлению;

класс объектов – объекты, связанные с несколькими тематическими подпространствами. К универсальным классам относятся такие объекты, как персоны, публикации, документы, события, организации и т.п.;

Атрибуты цифрового объекта – структурированные свойства (метаданные), характеризующие объект с точки зрения задач ЕЦПНЗ;

Профиль метаданных – набор атрибутов (с указанием их обязательности), описывающий объекты того или иного класса;

Первый уровень онтологии (2)

- **Предметная онтология (тезаурус)** тематического подпространства— совокупность индексов классификационных систем, ключевых терминов с их тезаурусными связями, относящихся к данному научному направлению;
- **Тезаурусные связи** – связи между двумя элементами предметной онтологии с заданным набором значений, который может устанавливаться внутри данного тематического подпространства;
- **Атрибутные связи** – связи между различными объектами как внутри одного класса, так и между объектами различных классов;

Второй уровень онтологии

На втором уровне онтологии конкретизируются понятия первого уровня применительно к общему пространству и тому или иному тематическому подпространству, в частности, в соответствии с определением, данным на первом уровне, определяются тезаурусы конкретных предметных областей. Конкретные локальные классы объектов определяются спецификой той области науки, к которой они относятся. Для микробиологии это могут быть штаммы, коллекции и базы данных микроорганизмов; для языкознания – словари, языки, корпуса текстов и т.п., для химии – реакции, вещества, химические элементы и т.п.

Третий уровень онтологии

На третьем онтологическом уровне устанавливаются профили метаданных объектов ЕЦПНЗ и связи между ними. В качестве примеров однозначных атрибутивных связей внутри одного класса можно привести связи публикаций различного уровня: статья (аналитический уровень) связана связью «опубликована в...» с определенным выпуском журнала (монографический уровень), который, в свою очередь, связан с описанием журнала (сводный уровень).

Научное Наследие

Сущности:

- Персоны
- Публикации
- Организации
- Коллекции
- Словари
- Даты

ЭБ «Научное Наследие»

Поиск

публикация / Название на языке оригинала содержит ум и

только записи пользователя:

публикации

Сортировать результаты по полю

Найдено записей: 2

[в новом окне](#)

[Вернуться в главное меню](#)

- ☐ [Натансон Исидор Павлович](#) [Простейшие задачи на максимум и минимум. Изд. 2-е.](#) – 1952. – 34 Номер МСЦ: 10081783
- ☒ [Циолковский Константин Эдуардович](#) [Ум и страсти.](#) – 1928. – 22 Номер МСЦ: 10078863

RDF-формат ВЦ РАН

```

<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:dctype="http://purl.org/dc/dcmitype/"
  xmlns:core="http://isir.ras.ru/namespace/"
  xmlns:auxiliary="http://umeta.ru/namespaces/blocks/auxiliary/"
  xmlns:kernel="http://umeta.ru/namespaces/platform/kernel/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <core:Person rdf:about="http://scirus.benran.ru/scirus/heritage/view-record.php?tbl=person&id=131">
    <core:personName rdf:parseType="Resource">
      <core:last>Циолковский</core:last>
      <core:first>Константин</core:first>
      <core:middle>Эдуардович</core:middle>
      <dc:language>ru</dc:language>
    </core:personName>
    <core:birthDate rdf:parseType="Resource">
      <core:dateValue>1857, 5 (17) сентября</core:dateValue>
    </core:birthDate>
    <core:deathDate rdf:parseType="Resource">
      <core:dateValue>1935, 19 сентября</core:dateValue>
    </core:deathDate>
    <core:birthPlace rdf:parseType="Resource">
      <rdf:value><![CDATA[с. Ижевское, ныне Рязанской области]]></rdf:value>
      <dc:language>ru</dc:language>
    </core:birthPlace>
    <core:activity rdf:parseType="Resource">
      <dc:format rdf:about="http://www.iana.org/assignments/media-types/text/html" rdf:parseType="Resource"/>
      <core:textContent>механика, астрономия, техника</core:textContent>
      <dc:language>ru</dc:language>
    </core:activity>
    <core:archiveInformation rdf:parseType="Resource">
      <dc:format rdf:about="http://www.iana.org/assignments/media-types/text/html" rdf:parseType="Resource"/>
      <core:textContent><![CDATA[Личный фонд Циолковского Константина Эдуардовича хранится в Архиве Российской академии наук: АРАН. Ф.555 - Циолковский Конст

```

Единое Цифровое Пространство Научных Знаний

Спасибо за внимание!