

Концептуальное
моделирование
предметных областей
для решения задач над
данными





План лекции

- Мотивация курса
- Организация концептуальной информации
- История и перспективы концептуального моделирования
- Стек семантических технологий
- Содержание курса
- Организационные вопросы



Мотивация курса

- Определение семантики данных для решения задач над ними
- Абстрактные спецификации в основе представления данных и манипулирования ими
- Решение проблем неоднородности данных при решении задач
- Взаимодействие исследователей над научными данными
- Автоматизация работы с данными



Недостатки используемых технологий

- **Web-технологии**
 - Разметка данных, рассчитанная на стиль визуального представления
 - Текстовая информация
 - Социально-ориентированное содержимое
 - Интерактивные сервисы, социальные сети
- **Неструктурированная или слабоструктурированная информация**
 - Рассчитана на человека, а не на обработку машиной
 - Эвристические подходы не гарантируют точности и полноты
- **Непостоянство**
 - Неконтролируемость и хаотичное появление и исчезновение информации
- **Неоднородность**
 - Большие объёмы неоднородной информации
 - Неидентифицируемые объекты
 - Большое количество семантически неоднородных источников данных
- **Нетривиальность задач**
 - Автоматизация обработки информации
 - Интеллектуальные интерактивные сервисы
 - Распределённые и рассредоточенные информационные системы
 - Горизонтальное масштабирование



Пути решения

- **Выработка общих словарей и семантики терминов в предметных областях**
 - Взаимодействие с исследовательскими группами
 - Взаимодействие специалистов предметных областей и специалистов в информационных технологиях и науке о данных
- **Концептуальное моделирование предметных областей**
 - Согласование понятий и семантики сущностей в предметной области
 - Связывание данных, объектов и реализаций методов с определённой семантикой предметной области
- **Семантическая интероперабельность информационных ресурсов**
 - Интеграция неоднородных информационных ресурсов средствами однородных описаний предметной области
 - Однородное представление информации в предметной области
- **Понятные машине спецификации**
 - Возможность автоматического вывода
- **Курирование научных данных, обеспечение повторного использования результатов, воспроизводимости экспериментов**
 - Возможность использования наработок и представления результатов в пригодном для использования виде
- **Создание исследовательских сообществ в предметных областях**
 - Развитие стандартов предметных областей, накопление релевантных ресурсов



История семантического моделирования (1)

- Исследование концептуального моделирования (60-70 гг.)
 - Модели данных
 - Концептуальные схемы
 - Реляционная модель
 - ER-модель данных: сущности и связи (Чен, 1976 г.)
- Семантические модели данных (70-80 гг.)
 - Семантические сети
 - Логическое программирование
 - Prolog
 - Дедуктивные базы данных
 - Базы данных + логический вывод + декларативные запросы
 - Datalog, F-logic
- Развитие языков спецификаций (90-е гг.)
 - KIF – формат обмена знаниями, логика первого порядка
 - СИНТЕЗ – отечественная разработка
 - RDF – формат описания ресурсов
- Развитие логик описаний (90-е гг.)
 - KI-One, Loom, Classic (логика описаний ALCNR)
 - Автоматизация задачи включения между двумя классами объектов
- Онтологии (90-е гг.)
 - Ontolingua (KIF в основе)
 - Использование логик описаний в онтологиях
 - OIL, OWL (логика описаний SHOIN)



Развитие семантического моделирования (2)

- **Интеллектуальные агенты**
 - Мобильная программа или модель, обеспечивающая обработку на стороне данных
 - Использование автоматического вывода на основе различных логик
- **Концепция семантического веба (2000-е годы)**
 - Базис на RDF, OWL и языках правил (2001)
 - Исследования сложности моделей OWL, создание OWL 2 (2009)
 - Развитие языков правил SWRL (2004), RIF (2009)
- **Когнитивные технологии (2010-ые)**
 - Машинное и глубокое обучение
 - IBM Watson
- **Инфраструктуры исследовательских данных**
 - Обеспечение доступа к коллекциям данных в различных дисциплинах
 - Сервисы поиска, обработки данных, вычислительных ресурсов
- **Принципы достижения интероперабельности и повторного использования данных**
 - Руководящие принципы, ведущие к решению проблем неоднородности данных и их автоматизированной обработки
- **Интернет вещей**
 - До сих пор слабо использует семантические технологии
- **Языковые модели**
 - Вероятностные методы генерации текста
- **Нейронет?**
 - Взаимодействие живых организмов и интеллектуальных агентов на основе нейрокоммуникации



Определения для работы с семантикой данных

- Семантика
 - Изучение значения единиц языка (естественного, искусственного)
- Формальная семантика
 - Интерпретации языков путём их формального описания в математических терминах
- Понятие
 - мысленное представление, используемое для классификации сущностей мира по некоторым признакам, которое рассматривается как эквивалент сущности для некоторых целей
- Концептуализация
 - процесс осмысления предметной области, формирования понятий для идентификации и классификации сущностей
 - результат концептуализации как процесса
- Модель
 - Абстрактное, отвлечённое от реальности представление, отражающее наиболее общие свойства объектов
- Концептуальная модель предметной области
 - абстрактное описание предметной области, независимое от аспектов реализации систем, в рамках которых оно используется, определяющее структуру и поведение сущностей в предметной области
- Модель данных
 - совокупность языков определения данных и манипулирования данными
- Концептуальная схема
 - формальное описание концептуальной модели предметной области средствами языка концептуального моделирования
- Декларативная спецификация
 - описание предмета или требуемого результата с помощью его свойств и ограничений
 - В отличие от императивной спецификации, описывающей способ получения результата
- Представление знаний
 - структурирование знаний с целью формализации процессов решения задач в определенной предметной области
 - В базах знаний используются декларативные подходы к спецификации задач



Основные для этого курса средства спецификации

- Модель требований
 - Декомпозиция постановки задач на подзадачи
 - Операционализация требований
- Онтология
 - Явным образом определённая спецификация концептуализации
 - Формальное представление множества понятий предметной области и связей между этими понятиями
 - Словарь предметной области, содержащий точные определения или аксиомы, ограничивающие смысл терминов
 - Онтология = словарь предметной области + логическая теория
- Концептуальная схема
 - Определение структур, типов и поведения объектов предметной области
 - Абстрактное представление данных в предметной области
- Поток работ
 - Спецификация процесса решения задачи с использованием структур данных с помощью определения порядка, условий и ограничений вызова методов



Различие онтологий и концептуальных схем

Концептуальные схемы

- Представление структур данных и их поведения
- Абстрактное представление данных
- Соглашение разработчиков о представлении данных
- Гипотеза замкнутого мира
- Экземпляр – данные, относящиеся к определённому объекту
- Для работы с системами

Онтологии

- Теория предметной области
- Наиболее общие свойства понятий
- Соглашение сообщества о трактовке и использовании понятий
- Гипотеза открытого мира
- Экземпляр – отнесение объекта к понятию предметной области
- Общедоступны



Виды организации терминов и понятий

- **Словарь**
 - список слов или терминов естественного или искусственного языка
 - список всех слов, используемых в документе или коллекции документов
 - набор слов, известных или используемых агентом (человеком или компьютером)
- **Глоссарий**
 - список терминов предметной области с вербальными определениями
- **Тезаурус**
 - словарь, дополненный информацией о синонимах, омонимах, родовидовых связях терминов, отношениях части/целого, ассоциативно связанных терминах
- **Иерархия**
 - множество сущностей, частично упорядоченное в соответствии с некоторым отношением
- **Система классификации**
 - отнесение объектов к группам (классам) по некоторому критерию
 - Класс – множество однотипных объектов или объектов с общим свойством
- **Таксономия**
 - упорядоченная (обычно иерархическая) система классификации
- **Рубрикатор**
 - иерархическая система классификации, предназначенная для систематизации информационных фондов, массивов или изданий



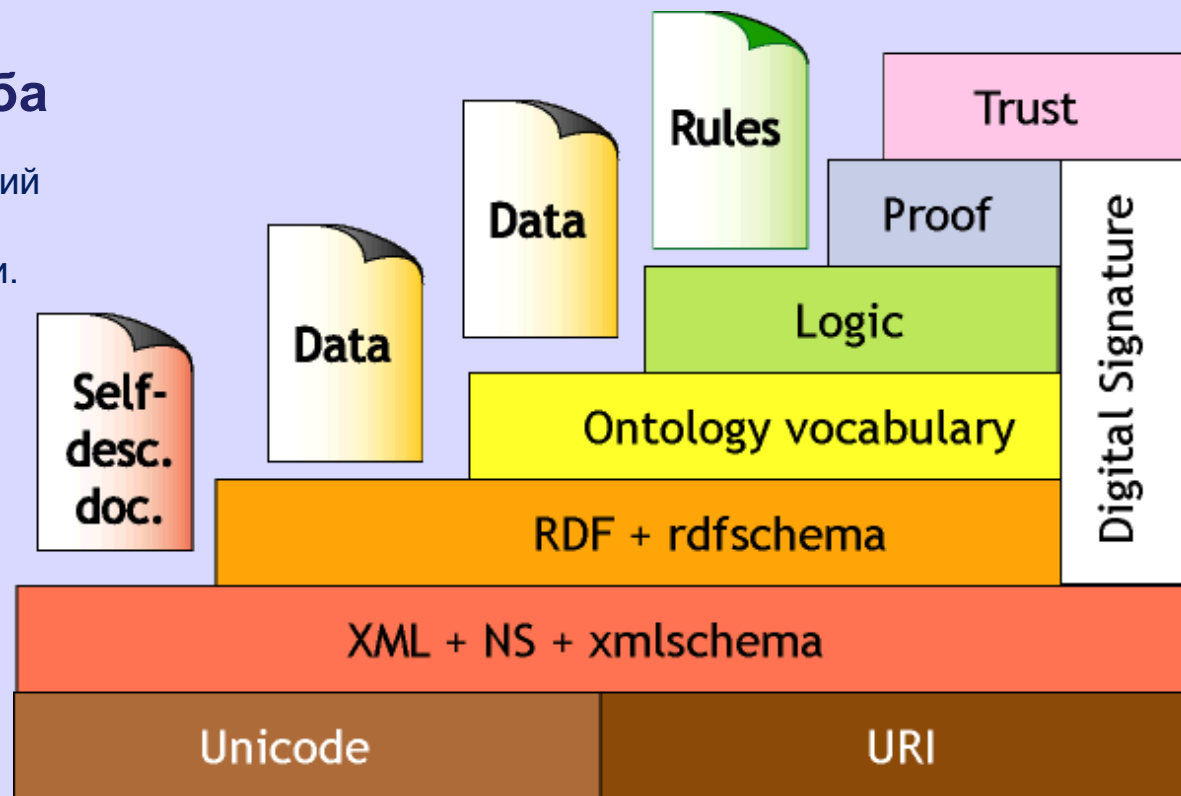
Требования к семантическому вебу

- Идентификация ресурсов, мест, агентов
 - Использование глобальных идентификаторов
- Семантическая интероперабельность приложений
 - Веб данных, существующих вне рамок приложений
 - Общие словари понятий, общедоступная (shared) семантика
 - Развитие методов идентификации понятий
- Данные, снабжённые семантикой
 - Семантическая разметка данных вместо разметки стиля
- Возможность машинной обработки данных и автоматического решения задач
 - Формальность и декларативность моделей данных
 - Развитие методов автоматического вывода
 - Учёт сложности и разрешимости определённых задач в моделях
- Семантический веб
 - «Семантический веб является вебом информации, требующей действий над ней – информации, полученной из данных на основе семантической теории для интерпретации символов»
 - «Семантическая теория обеспечивает учет смысла, при котором логическая связь терминов устанавливает интероперабельность систем»



Архитектура семантического веба

Стек семантических технологий семантического веба
Слоёный пирог Т. Бернерс-Ли.





Уровни архитектуры семантического веба

- Уровень RDF-Schema
 - Идентификация ресурсов (URI)
 - Минимальная модель данных для определения ресурсов
 - Семантика ресурсов
- Уровень онтологии
 - Средства описания семантики понятий
 - Общедоступные описания концептуализации
- Уровень логики
 - Логический вывод
 - Логическое программирование
- Уровень доказательств и надёжности данных
 - Цифровые подписи
 - Криптография
 - Происхождение данных



Принятые стандарты семантического веба

- RDF (Resource Description Framework)
 - Тройки субъект-предикат-объект
- RDF-Schema (2004)
 - Классы (множества), свойства (отношения), иерархии классов и свойств, области определения и значения свойств и др.
- RDFa (RDF in Attributes)
 - Подход к семантической разметке для HTML
- SPARQL (SPARQL Protocol and RDF Query Language)
 - Язык запросов к RDF-тройкам
- OWL (Web Ontology Language)
 - Язык спецификации онтологий, основанный на дескриптивной логике SHOIN
- OWL 2
 - Три профиля, основанные на разных дескриптивных логиках с разными наборами разрешимых задач
- SWRL (Semantic Web Rule Language)
 - OWL + RuleML (datalog-подобный язык правил над классами и свойствами OWL)
- RIF (Rule Interchange Format)
 - Спецификации диалектов правил с различной семантиков
- PROV (Provenance)
 - Происхождение данных

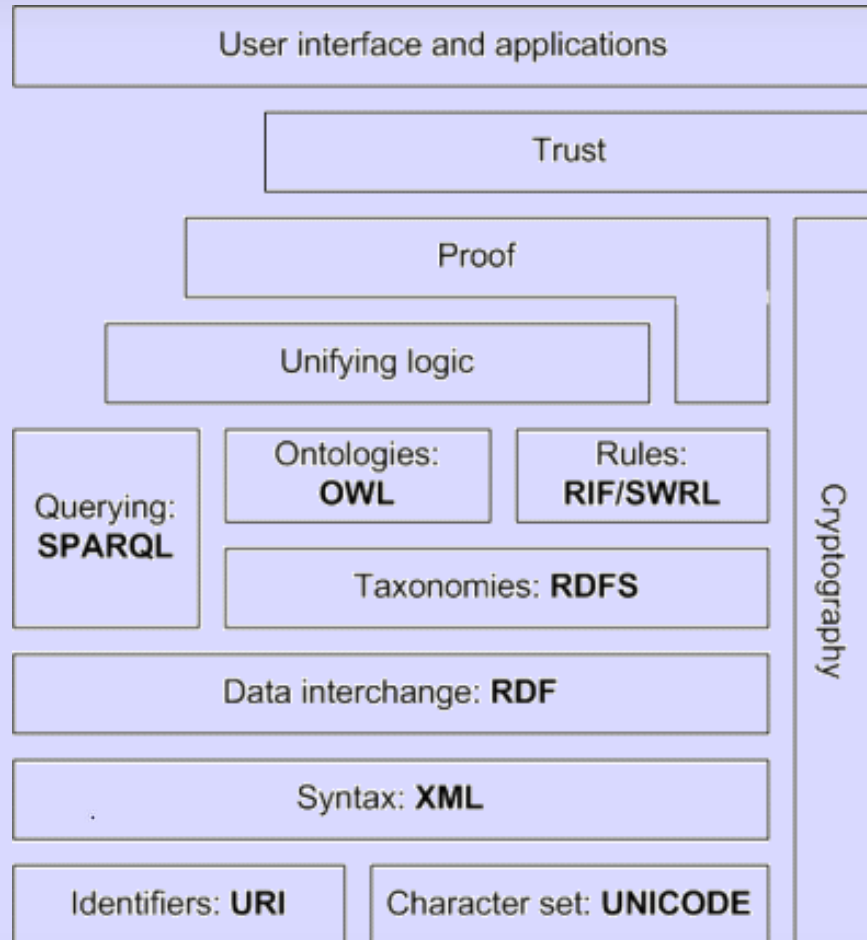


Примеры спецификаций

- **RDF**
 - `:Jack rdf:type :Man`
 - `:Jack :hasParent :Sammy`
 - `:Jack :hasParent :Suzie`
 - `:Jack :hasSister :Suzie`
- **RDFS**
 - `:hasParent rdfs:domain :Person`
 - `:hasParent rdfs:range :Person`
- **OWL**
 - `Class (:Person`
 - `SubClassOf(ObjectExactCardinality(1 :hasParent :Man))`
 - `SubClassOf(ObjectExactCardinality(1 :hasParent :Woman))`
 - `)`
 - `SubObjectPropertyOf(:hasFather :hasParent)`
 - `InverseObjectProperties(:hasParent :hasChild)`
 - `DisjointClasses(:Woman :Man)`
- **SPARQL**
 - `SELECT ?x1 ?x3`
 - `WHERE {`
 - `?x1 rdf:type :Person .`
 - `?x1 :hasParent ?x2 .`
 - `?x2 :hasBrother ?x3 .`
 - `}`
- **SWRL**
 - `hasAuncle (?x1, ?x3) ← hasParent (?x1, ?x2) · ^ hasBrother (?x2, ?x3)`



Архитектура и стандарты языков





Решение задач над концептуальными схемами

- Построение модели требований задачи
- Концептуализация предметной области
- Построение онтологии предметной области / задачи
- Построение концептуальной схемы предметной области / задачи
- Построение потока работ для решения задачи
- Публикация результатов



Онтологическое моделирование предметных областей
Описание структуры объектов в концептуальных схемах
Описание поведения объектов в концептуальных схемах
Обеспечение интероперабельности данных и их повторного
использования
Некоторые семантические технологии

Состав курса



Онтологическое моделирование

- Языки концептуальных описаний
 - Логики описания
 - Принципы логического вывода на основе понятий
- Языки семантического веба
 - RDF, OWL, SPARQL
- Принципы описания предметных областей и понятий разной природы
- Общеизвестные онтологии, согласование онтологий
- Метаданные, семантическое аннотирование объектов, происхождение данных



Концептуальные схемы

- Модели требований
- Описание структур данных и объектов
- Язык UML
- Описание декларативных методов
- Языки правил
- Поток работ, язык BPMN
- Интеграция схем и отождествления объектов
- Описания открытого и закрытого мира



Обеспечение интероперабельности

- Интероперабельность и повторное использование данных
- Коллекции ресурсов данных, методов, потоков работ
- Инфраструктуры исследовательских данных и работа в исследовательских сообществах
- Руководящие принципы обеспечения интероперабельности и повторного использования данных
- Жизненный цикл решения исследовательских задач над данными
- Спецификации происхождения данных
- Публикация результатов исследований



Некоторые семантические технологии

- Открытые связанные данные (LOD)
- Доступ к данным, основанный на онтологиях (OBDA)
- Онтологические языки спецификации сервисов (OWL-S)
- Спецификации повторно используемых потоков работ (RO-Crate)
- Нейросимволические методы



Самостоятельная работа

- Описание предметной области исследования, проводимого в магистерской работе
 - Постановка задачи
 - Определение существенных терминов
 - Моделирование предметной области
 - Построение концептуальной схемы предметной области
 - Возможно, отображение в неё исходных данных
- Для того, чтобы способствовать написанию магистерских работ и статей



Организационные вопросы

- Скворцов Николай Алексеевич
 - Институт проблем информатики ФИЦ ИУ РАН
 - nskv@mail.ru
 - +79261355404
- Лекции по четвергам в 16:20 в аудитории 582а ВМК МГУ
- Самостоятельная работа
 - Описание предметной области магистерского исследования
- Материалы будут выкладываться по адресу
 - <https://eclass.cmc.msu.ru/course/view.php?id=131>
- Отчётность в конце семестра
 - Экзамен с вопросами по лекциям