Описание ресурсов и метаданные

Курс: Концептуальное моделирование предметных областей

План лекции

- Документы и ресурсы
- Язык описания ресурсов RDF
- Язык запросов SPARQL
- Онтологии и RDF
- RDF-a
- Метаданные
- Семантическое аннотирование

Пара слов об XML и других языках разметки данных

- Дерево XML (eXtensible Mark-Up Language)
  - Элемент – пользовательский тег, набор атрибутов тега и их значений, содержимое тега
  - Схема документа XML Schema – определение структуры деревьев
- Хорош для документов, но не для ресурсов
  - Семантика фрагментов документов
  - Схема определяет разрешённую структуру документа, но не семантику ресурса
  - Не сопоставляется с онтологическим словарём
  - Нет идентификации ресурсов в разных документах
- То же касается распространённых языков обмена данными, таких как JSON

Пример XML

```
<course year="2021">
    <title>Conceptual Modeling</title>
    <teacher>
       <name>Nikolay Skvortsov</name>
       <email>nskv@mail.ru</email>
     </teacher>
     <lecture number=1>
      <theme>Semantic Web
                  Architecture</theme>
           …
      </lecture>
      …
</course>
```

Требования к описанию ресурсов

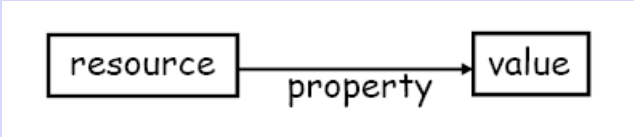
- Глобальная уникальная идентификация
- Независимость от последовательности описания
- Читаемость человеком и машиной
- Унифицированное определение семантических единиц: классов, отношений, объектов.

URI – глобальные идентификаторы

- XML как базовый язык документов
- Разные документы могут описывать один и тот же ресурс
  - могут, хотя и не должны иметь глобальный идентификатор
- URI (Uniform Resource Identifier)
  - Универсальное именование ресурсов Веба
  - тот же URI = тот же ресурс

RDF – язык описания ресурсов

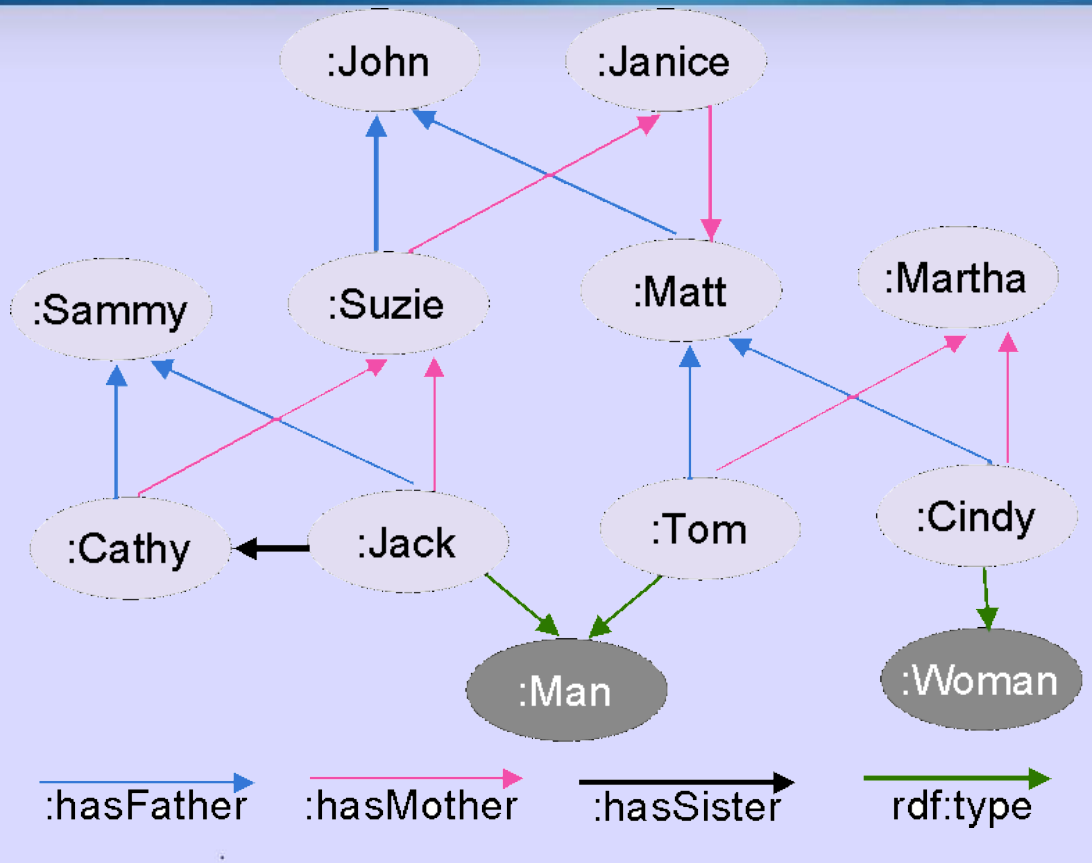
- Простейшая концептуальная модель
- Триплеты (субъект, предикат, объект)
  - Субъект – ресурс
  - Предикат – свойство ресурса
  - Объект – значение свойства (ресурсы или литералы)



- Модель данных
  - направленный помеченный граф
- Унифицированные идентификаторы ресурсов
  - Всякий ресурс имеет URI
- Неименованные узлы
  - не имеют URI
- Пространства имён
  - Разновидности ресурсов, свойств и типов значений принадлежат словарям предметной области, определяющим их семантику
- Контейнеры
  - В качестве значений возможны bag, sequence, alt
- Синтаксисы модели
  - Графический (граф)
  - Триплеты (N3, turtle)
  - XML-сериализация (RDF/XML)

Данные RDF

:Jack rdf:type :Man
:Jack :hasParent :Sammy
:Jack :hasParent :Suzie
:Jack :hasSister :Suzie



Пример: Jean has a friend born on the 21st of April

_:p1 – неименованный ресурс (blank node)

foaf: - пространство имён словаря ”Friend of a Friend”

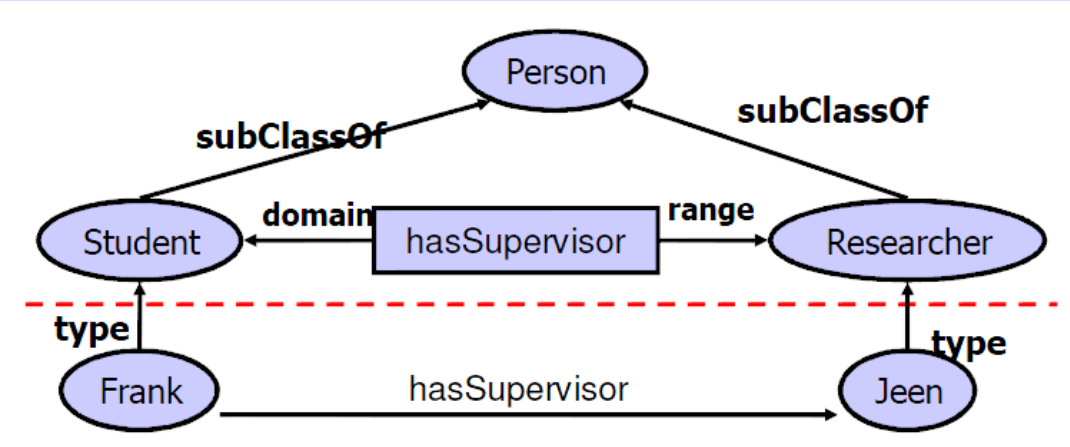
ex:Jean foaf:knows _:p1

_:p1 foaf:birthDate 04-21

RDF-схема

- RDF-Schema
- Определяет словарь для RDF
  - Уровень метаданных для множества предикатов RDF
  - Предопределённая семантика для RDF
  - Утверждения RDF-Schema определяют ограничения для описаний RDF
- Элементы языка
  - rdfs:Class – множество ресурсов с общими свойствами
    - rdfs:subClassOf – свойство, определяющее связь подмножества между множествами ресурсами
    - rdf:type – свойство, определяющее принадлежность ресурса классу
  - rdf:Property – определяет свойства ресурсов
    - rdfs:subPropertyOf – связь подмножества между свойствами
    - rdfs:domain – обозначает класс субъектов (области определения)
    - rdfs:range – обозначает класс объектов (области значений)
  - типы данных XML-Schema

Определения ресурсов на RDF и RDF-Schema



Семантика RDFS

C rdfs:subClassOf D	$\forall X(C(X) \Rightarrow D(X))$	$C \sqsubseteq D$
P rdfs:subPropertyOf R	$\forall X \forall Y(P(X,Y) \Rightarrow R(X,Y))$	$P \sqsubseteq R$
P rdfs:domain C	$\forall X \forall Y(P(X,Y) \Rightarrow C(X))$	$\exists P \sqsubseteq C$
P rdfs:range D	$\forall X \forall Y(P(X,Y) \Rightarrow D(Y))$	$\exists P^{-} \sqsubseteq D$

Правила вывода в RDF и RDFS (примеры)

- x p y.
  => p rdf:type rdf:Property.
- p rdfs:domain x. u p v.
  => u rdf:type x.
- p rdfs:range x. u p v.
  => v rdf:type x.
- u r v.
  => u rdf:type rdfs:Resource. v rdf:type rdfs:Resource.
- u rdfs:subClassOf v. v rdfs:subClassOf w.
  => u rdfs:subClassOf w.
- u rdfs:subClassOf x. v rdf:type u.
  => v rdf:type x.

Язык запросов для RDF

- SPARQL (SPARQL Protocol and RDF Query Language)
- Запросы к графу RDF
  - Переменные ?x
  - PREFIX – определение пространств имен, фигурирующих в запросе
  - SELECT <список возвращаемых переменных>
  - WHERE { <список шаблонов RDF-триплетов с переменными на месте субъектов или объектов> }
  - Возвращает множество наборов ресурсов в соответствии со списком возвращаемых переменных
- OPTIONAL - {?x :name ?name } – имя может быть пустым
- FILTER regex(?name, "^J") – условия фильтрации
- INSERT, CONSTRUCT – генерация триплетов

Пример: дядей x1 является x3

```
SELECT ?x1?x3 ?name
WHERE {
      ?x1 rdf:type :Person .
      ?x1 :hasParent ?x2 .
      ?x2 :hasBrother ?x3 .
      ?x1 :hasName ?name .
}
```

Базы и словари RDF

- RDF-базы данных:
  - Dbpedia - извлечение структурированной информации из данных Wikipedia
  - Geonames – база географических объектов
  - DBLP – база данных публикаций по информатике
  - PubMed – база данных публикаций по медицине
  - UniProt – база данных белков
  - OpenCyc – объёмная онтологическая база данных
  - LOD – технология связанных открытых данных
- Словари
  - Dublin Core (DC) – атрибуты библиотечных метаданных
  - Friend-of-a-Friend (FOAF) – словарь описания людей, их отношений и деятельности
  - Semantically-Interlinked Online Communities (SIOC) – словарь онлайн-сообществ
  - Description of a Project (DOAP) – словарь для описания проектов
  - Simple Knowledge Organization System (SKOS) –словарь для представления стандартизованных таксономий
  - Creative Commons (CC) – словарь для описания лицензий

Ограничения RDFS

- RDFS
  - классы и подклассы
  - свойства (отношения) и подсвойства
  - область определения и значений свойств
- OWL
  - Операции на классах
  - Ограничения области определения и значения свойств определённых типов
  - Ограничения множественности и кванторов всеобщности и существования
  - Свойства отношений: транзитивность, инверсность, симметричность

Уровень онтологий OWL в модели RDF

- OWL + RDF
  - Для языка OWL определено пространство имён RDF
    - owl:Class, owl:ObjectProperty, owl:AllValuesFrom, …
  - В описаниях RDF при использовании средств из словаря пространства имён OWL применяется семантика OWL
  - Семантика OWL создаёт дополнительные ограничения для ресурсов RDF, проверяемые специализированными средствами для OWL
- RDF расширяемый язык
  - Подобным образом в описания на RDF может быть вменена семантика любой другой модели, помимо OWL

OWL в RDF

Студент должен быть зарегистрирован на курсы в количестве от 3 до 6

```
:Student  rdfs:subClassOf  _:a
:Student  rdfs:subClassOf  _:b
_:a  rdfs:subClassOf   owl:Restriction
_:a  owl:onProperty  RegisteredTo
_:a  owl:minCardinality  3
_:b  rdfs:subClassOf  owl:Restriction
_:b  owl:onProperty  RegisteredTo
_:b  owl:maxCardinality  6
```

RDFa и метаданные RDF

- HTML+RDFa

```
<div typeof="foaf:Person" xmlns:foaf="http://xmlns.com/foaf/0.1/">
<p property="foaf:name">Alice Birpemswick</p>
<p>Email: <a rel="foaf:mbox" href="mailto:alice@example.com">
alice@example.com</a></p>
<p>Phone: <a rel="foaf:phone" href="tel:+70123456789">
(012)345-67-89</a></p>
</div>
```

- RDF

```
_:x rdf:type foaf:person
_:x foaf:name “Alice Birpemswick”
_:x foaf:mbox mailto:alice@example.com
_:x foaf:phone tel:+1-617-555-7332
```

- С ресурсами в произвольных представлениях можно связать описывающие их метаданные в RDF

Метаданные

- Снабжение документов описанием Dublin Core

```
:dcProvTutorial a foaf:Document;
        dct:title “PROV Tutorial”;
        dct:creator :daniel;
        dct:created "2013-08-25”;
        dct:replaces :tutorialDraft;
```

- Описание семантики данных с помощью аннотирования понятиями онтологии предметной области
  - Схема данных

    Person
    - pictures: Picture;
    - birthYear: integer;
  - Аннотация

    :Person rdf:type artontology:Painter
    :Picture rdf:type artontology:Picture
  - Возможность семантического поиска данных в предметной области
  - Возможность автоматического обогащения данных

Семантическое аннотирование

- Аннотирование ресурсов в терминах онтологических понятий определяет их семантику в предметной области
- Способы аннотирования
  - Экземпляр понятия с помощью rdf:type (наиболее распространённое, но наименее выразительное)
    - :Algol rdf:type astront:BynaryStar
  - Определением множества в терминах понятий { x | … }
  - Определение нового подпонятия как выражения в терминах онтологии
- Возможно одновременно использовать аннотирование в терминах нескольких онтологий
  - Онтология предметной области
  - Онтология происхождения данных
  - Онтология качества данных
  - Онтология, описывающая метамодель и т. д.
- Поиск по аннотациям позволяет найти семантически релевантные ресурсы

Метаданные происхождения

- **Происхождение** данных – метаданные об источниках, принадлежности, средствах генерации, истории изменения данных и о других реквизитов о данных
- Задачи происхождения данных
  - описание источников и принадлежности данных
  - выражение качества данных
  - метаинформация об операциях обработки данных (lineage)
- Применение происхождения данных
  - Базы данных
  - Потоки работ
  - Семантический веб
  - Представление знаний
  - Информационный поиск

Происхождение данных в СУБД

- Линии происхождения (lineage) данных – метаданные, сопровождающие кортежи базы данных, предоставляющие вместе с результатами запросов информацию о пути обработки данных
- Модели происхождения данных в реляционных базах данных
  - why-provenance – какие исходные кортежи участвовали в генерации данного кортежа
  - how-provenance – какие операции над какими кортежами участвовали в генерации данного кортежа
  - where-provenance – место хранения исходных данных для генерации данного кортежа (отношение-кортеж-атрибут)
- Модели выбираются линейной сложности
- Средства: СУБД Chimera

Пример происхождения данных в СУБД (how)

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

ExternalTours

	name	destination	type	price
t_3:	BayTours	San Francisco	cable car	$50
t_4:	BayTours	Santa Cruz	bus	$100
t_5:	BayTours	Santa Cruz	boat	$250
t_6:	BayTours	Monterey	boat	$400
t_7:	HarborCruz	Monterey	boat	$200
t_8:	HarborCruz	Carmel	train	$90

```
SELECT   e.destination, a.phone
FROM     Agencies a,
         (SELECT name,
             based_in AS destination
         FROM Agencies a
         UNION
         SELECT name, destination
         FROM ExternalTours ) e
WHERE    a.name = e.name
```

Result of Q_2:

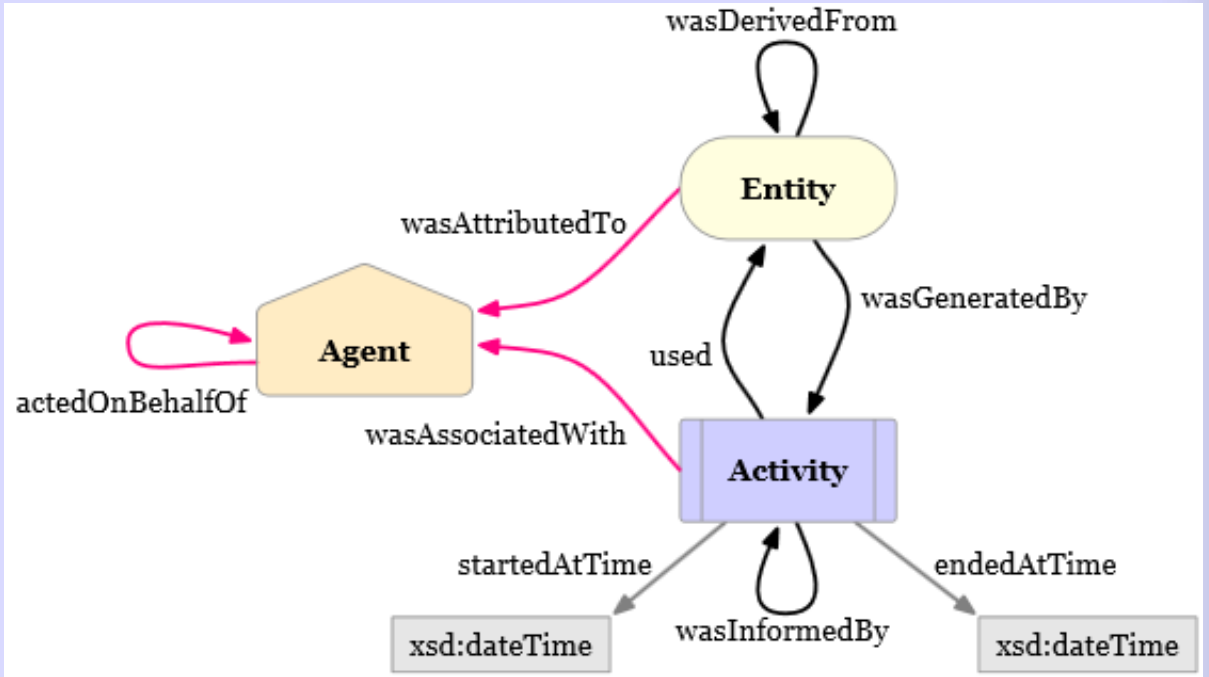
destination	phone	
San Francisco	415-1200	$t_1 \cdot (t_1 + t_3)$
Santa Cruz	831-3000	t_2^2
Santa Cruz	415-1200	$t_1 \cdot (t_4 + t_5)$
Monterey	415-1200	$t_1 \cdot t_6$
Monterey	831-3000	$t_1 \cdot t_7$
Carmel	831-3000	$t_1 \cdot t_8$

Стандарт W3C происхождения данных

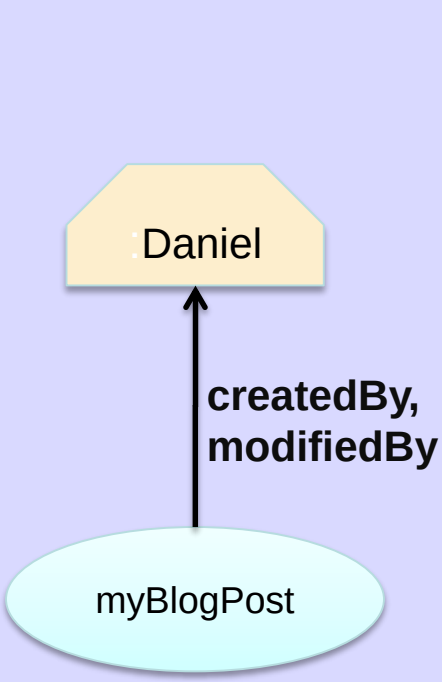
- Набор документов
  - PROV-DM: модель происхождения данных
  - PROV-O: реализация модели в языке онтологий OWL
- Метаданные PROV-O отвечают на вопросы
  - Кто имел отношение к генерации данных
  - Кто владеет данными
  - Как данные менялись от версии к версии
  - Как другие данные повлияли на них
  - Какими инструментами генерировались данные
  - и другие

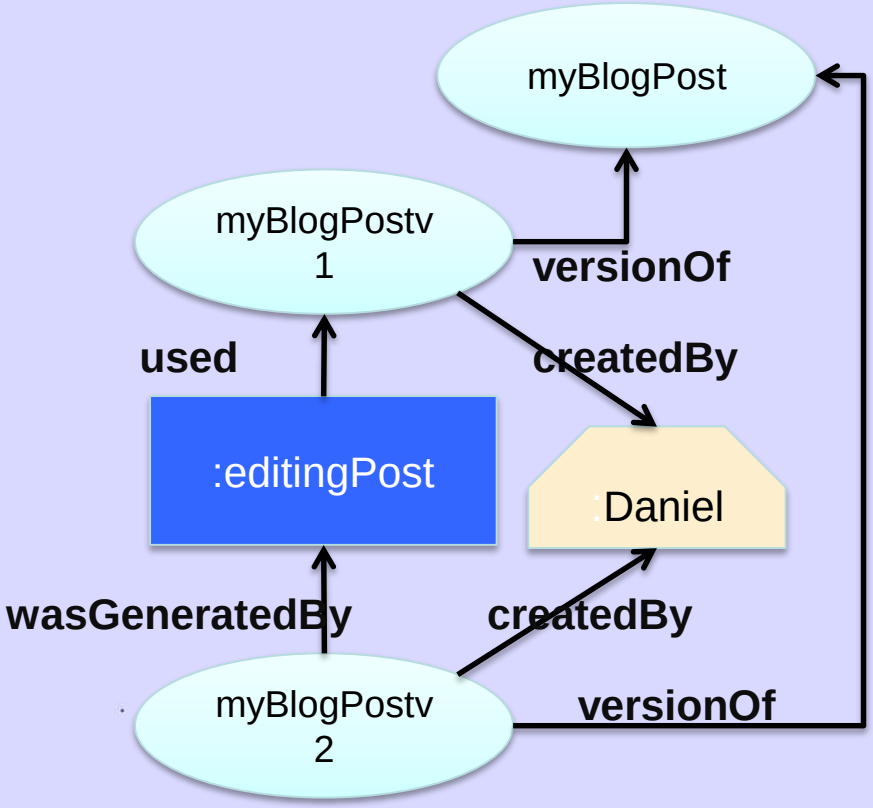
PROV-O

- Agent – агент, активный по отношению к данным
  - человек, организация, программа
- Entity – описываемая сущность
  - единичная сущность, план, множество, комплект
- Activity – деятельность, совершаемая агентом по отношению к данным
- Связи
  - Взаимодействие сущностей, агентов и деятельностей



Пример описания происхождения данных





Отношения в PROV-O

- Agent – wasAssociatedWith – Activity
  - Кто ответственен за изменение документа?
- Activity – used – Entity
  - При создании документа использована литература?
  - Для получения результата использован запрос?
  - Для процесса вычисления использованы входные данные?
- Entity – generatedBy – Activity
  - Как получен результат?
- Entity – wasDerivedFrom – Entity
  - На основе каких документов создан данный документ?
  - Какие данные участвовали в получения результата?
- Activity – wasInformedBy – Activity
  - Какая деятельность предшествовала данной деятельности?
  - Какие шаги нужды для решения задачи?
- Entity – wasAttributedTo – Agent
  - Кто автор документа?
- Agent – actedOnBehalfOf – Agent
  - Кто запустил инструмент обработки данных?
- Entity – hadPrimarySource – Entity
  - Откуда взяты данные?

Применение метаданных происхождения

- Оценка достоверности, подлинности, актуальности, точности данных
- Контроль источников данных
- Журналы обработки данных и применения методов
- Журналы прохождения тестов
- Отладка реализаций методов
- Обоснование критериев для слияния данных
- Сравнение работы разных реализаций
- Систематизация версий
- и другие

Качество данных

- измерения, метрики и модели
- проверка данных на соответствие требованиям
  - требования предметной области
  - требования методов анализа
  - требования задачи
- мониторинг
  - агрегация и анализ статистики
- улучшение качества данных
  - Граничит с очисткой данных
  - Обогащение данных

Состав описания модели качества данных

- **Измерения** (dimensions) качества данных – характеристики, измерительные индикаторы, оценивающие различные аспекты
- **Метрики** (или техники) – алгоритмы, эвристики, вывод знаний или обучение, обеспечивающие решение задач, относящихся к качеству данных
- **Методологии** – методические указания о комплексном использовании метрик для достижения эффективной оценки качества данных

Измерения качества данных

- объём данных (volume)
- полнота (completeness)
- точность (accuracy)
- целостность (consistency)
  - ограничения (constraints)
  - Коррекция данных
- временные (time-related) измерения качества

Полнота данных

- *Полнота кортежа – заполнение всех атрибутов кортежа*
- *Полнота атрибута – отсутствие null-значений для атрибута в отношении*
- *Полнота отношения – отсутствие null-значений в отношении в целом*

ID	Name	Surname	BirthDate	Email
1	John	Smith	03/17/1974	smith@abc.it
2	Edward	Monroe	02/03/1967	NULL
3	Anthony	White	01/01/1936	NULL
4	Marianne	Collins	11/20/1955	NULL

not existing

existing but unknown

not known if existing

Точность данных

- погрешность косвенных воспроизводимых измерений

$$\Delta F = \sqrt{\sum_{i=1}^{n} \left(\Delta x_i \frac{\partial F}{\partial x_i} \right)^2}$$

- Синтаксическая точность
  - Rman Holiday нет в домене названий фильмов, но есть Roman Holiday (проверка по другим источникам данных)
- Семантическая точность
  - Curtiz не является режиссёром фильма Casablanca
- Оценка влияния
  - От точности данных зависит корректное решение задач идентификации объектов, поиска дубликатов
  - Точность данных критична для ключевых атрибутов

Метрики качества данных

- Пример используемых метрик в определённой задаче
  - Полнота данных – относительное количество непустых значений атрибута для всех объектов
  - Точность данных – взятая из исходных данных точность или рассчитанная
  - Объем данных – ожидаемое количество возвращаемых кортежей
  - Возраст данных – разность текущей даты и даты создания данных
  - Целостность – соответствие набору определённых правил
  - Надёжность – результирующее значение, рассчитанное по определённой формуле на основании значений других метрик
- Методы вычисления характеристик качества в виде метрик могут быть выбраны совершенно другие
- Методология оценки качества в данном случае включает перечисленные метрики, обобщаемые метрикой Надёжность

Целостность данных

- Ограничения целостности
- Ключевая зависимость
  - Для набора атрибутов нет двух кортежей с одинаковым набором значений
- Зависимость включения
  - Внешними ключи в базах данных
- Функциональные зависимости
  - Формулы зависимости
- Коррекции данных (data edits)
  - marital status = married $\wedge$ *age* < *14*

Временные измерения качества

- актуальность (currency)
- временной охват (timeliness)
- изменчивость (volatility)

Улучшение качества данных

- Подходы к **улучшению** (improvement) качества
  - очистка данных
  - правила качества данных
  - обогащение данных

W3C Data Quality Vocabulary (DQV)

- Стандартизированная модель (RDF-словарь) от консорциума W3C для описания качества данных в машиночитаемом виде
- Обеспечение возможности публикации информации о качестве данных в вебе, чтобы как люди, так и программы могли находить, интерпретировать и использовать данные, основываясь на их качестве
- Расширяет описание каталогов данных (описанные в **DCAT** - Data Catalog Vocabulary).
- Переход от описания качества в свободной документации к **формальным спецификациям**
- Возможность автоматической оценки и фильтрации данных

Ключевые компоненты модели DQV

- Показатель качества (Dimension):
  - Характеристика, которая оценивается (например, Точность, Полнота, Происхождение).
- Метод оценки (Metric)
  - Конкретный способ вычисления показателя. Описывает, как была получена оценка (например, «процент заполненных полей», «среднеквадратичное отклонение»).
- Значение качества (QualityMeasurement)
  - Результат применения метода к конкретному набору данных или его части. Содержит числовое или категориальное значение (например, "0.95", "высокое").
- Связь с данными (по ссылкам dqv:hasQualityMeasurement)
  - Привязывается к ресурсам, описанным через DCAT
    - целым наборам данных (Dataset)
    - их конкретным представлениям (Distribution)
    - сервисам доступа (DataService)
- Позволяет оценивать качество на разных уровнях агрегации:
  - Набор данных (полнота покрытия неба)
  - Отдельная запись/объект (флаги качества в астрономии)
  - Конкретное значение (погрешность измерения блеска)

Пример

```
ex:galex-catalog a dcat:Dataset ;
    dcat:title "GALEX UV Survey Catalog" ;
    dcat:description "Ultraviolet source catalog from the GALEX mission" .
ex:photometricQuality a dqv:Dimension ;
    skos:prefLabel "Photometric quality"@en ;
    skos:definition "Assesses reliability of photometric measurements based on observation flags."@e
ex:positionalAccuracy a dqv:Dimension ;
    skos:prefLabel "Positional accuracy"@en ;
    skos:definition "Accuracy of object coordinates, dependent on catalog resolution and sky region."@en .

ex:flagBasedMetric a dqv:Metric ;
    dqv:inDimension ex:photometricQuality ;
    skos:definition "Metric derived from GALEX artifact and extraction flags (Fexf, Nexf, Falf, Nalf). Value is 0 (bad), 0.5 (uncertain), or 1 (good)."@en .
ex:positionalRMSE a dqv:Metric ;
    dqv:inDimension ex:positionalAccuracy ;
    skos:definition "Root mean square error of coordinates estimated from cross-match statistics in a given sky region."@en .
```

```
ex:galex-catalog dqv:hasQualityAnnotation [
    a dqv:QualityAnnotation ;
    dqv:isMeasurementOf ex:positionalAccuracy ;
    dqv:value "0.3"^^xsd:float ;   # средняя ошибка позиционирования в угловых секундах
    prov:wasAttributedTo ex:crossMatchAnalysis
] .

ex:object1 a dcat:Record ;
    dcat:inDataset ex:galex-catalog ;
    # ... другие свойства (координаты, блеск и т.д.) ...

# Оценка фотометрического качества объекта object1
ex:object1-quality a dqv:QualityMeasurement ;
    dqv:isMeasurementOf ex:photometricQuality ;
    dqv:computedOn ex:object123 ;
    dqv:value "0.8"^^xsd:float ; # вес рассчитан по флагам
    dqv:isBasedOnMetric ex:flagBasedMetric ;
    prov:wasGeneratedBy [
        a prov:Activity ;
        prov:used ex:object123 ;
        prov:description "Applied rule: if artifact flags != 0 then value=0; else if classification probability >0.5 then value=1 else 0.5."
    ] .
```