Реальный мир и хорошие модели данных

Семантика и онтологии в IT

Обзорная лекция

Учебная программа ТехИнвестЛаб.ру

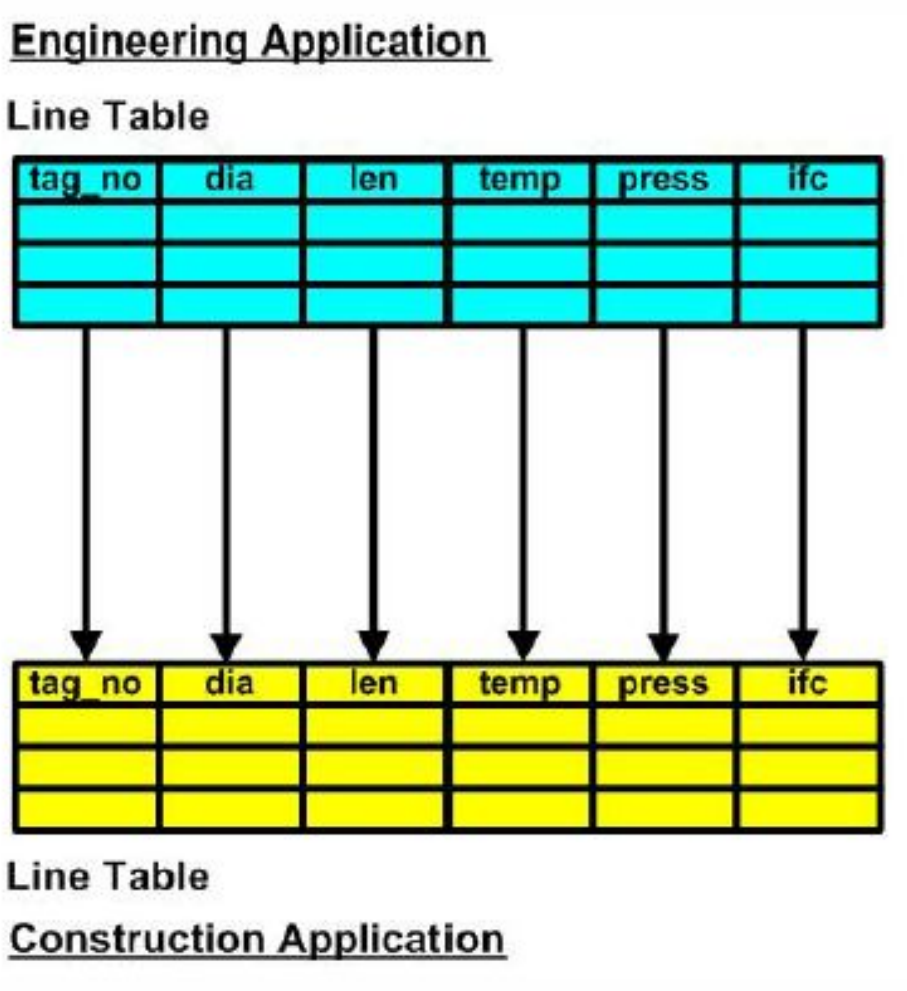
Реальность и данные

- Что есть в мире – как об этом записать в компьютере
- Модель данных
  - Структура
  - Смысл
- «Хорошие» и «плохие модели»
  - Понимание человек-человек
  - Понимание человек-компьютер
  - Понимание компьютер-компьютер

Как говорить о данных?

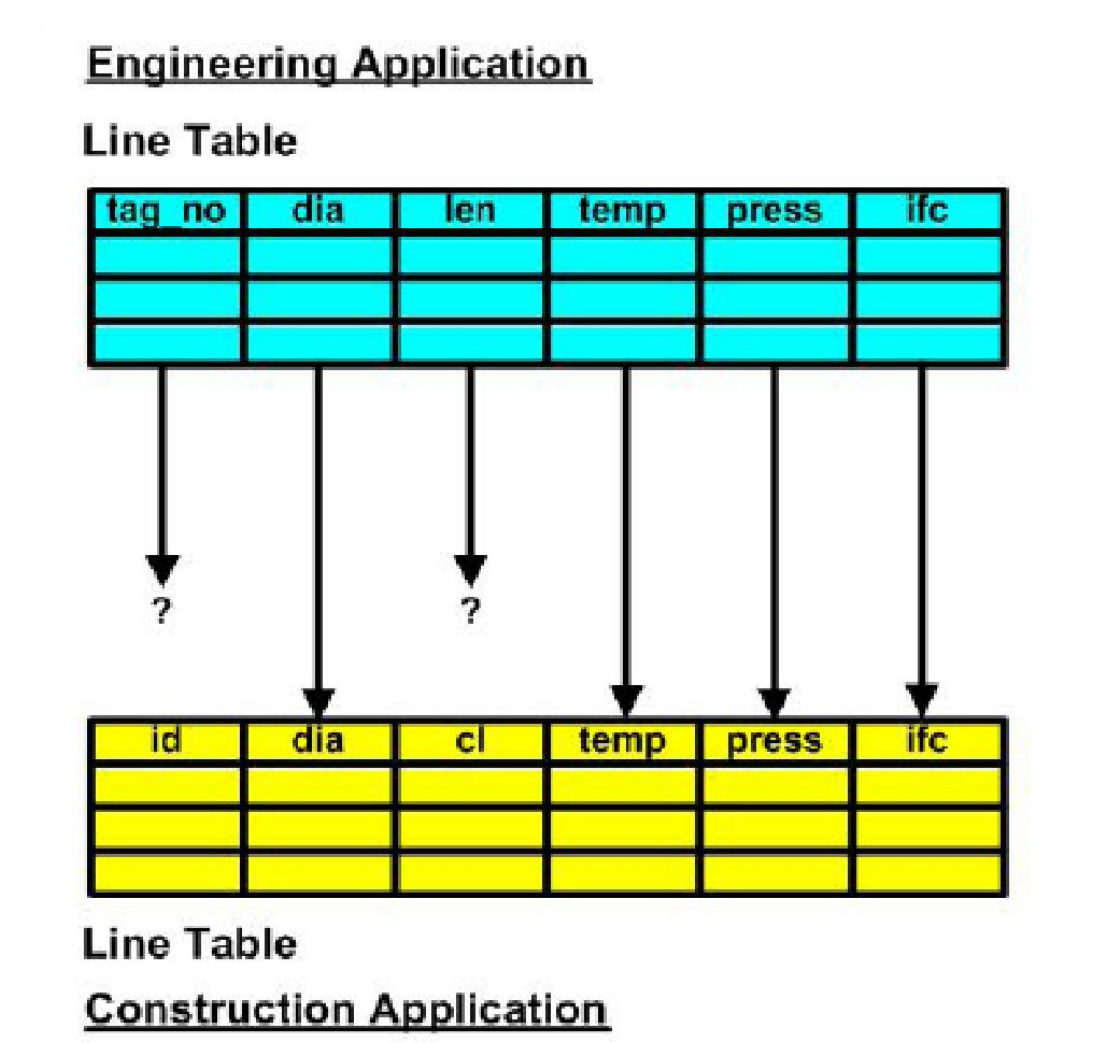
- Метамодель – модель – данные
- Языки метамоделирования
  - EXPRESS
    - EXPRESS-G
    - Текст
  - Английский
  - FOL
  - RDF/OWL (XML)
- Нужна ли граница «модель – данные»?

Совершенный мир



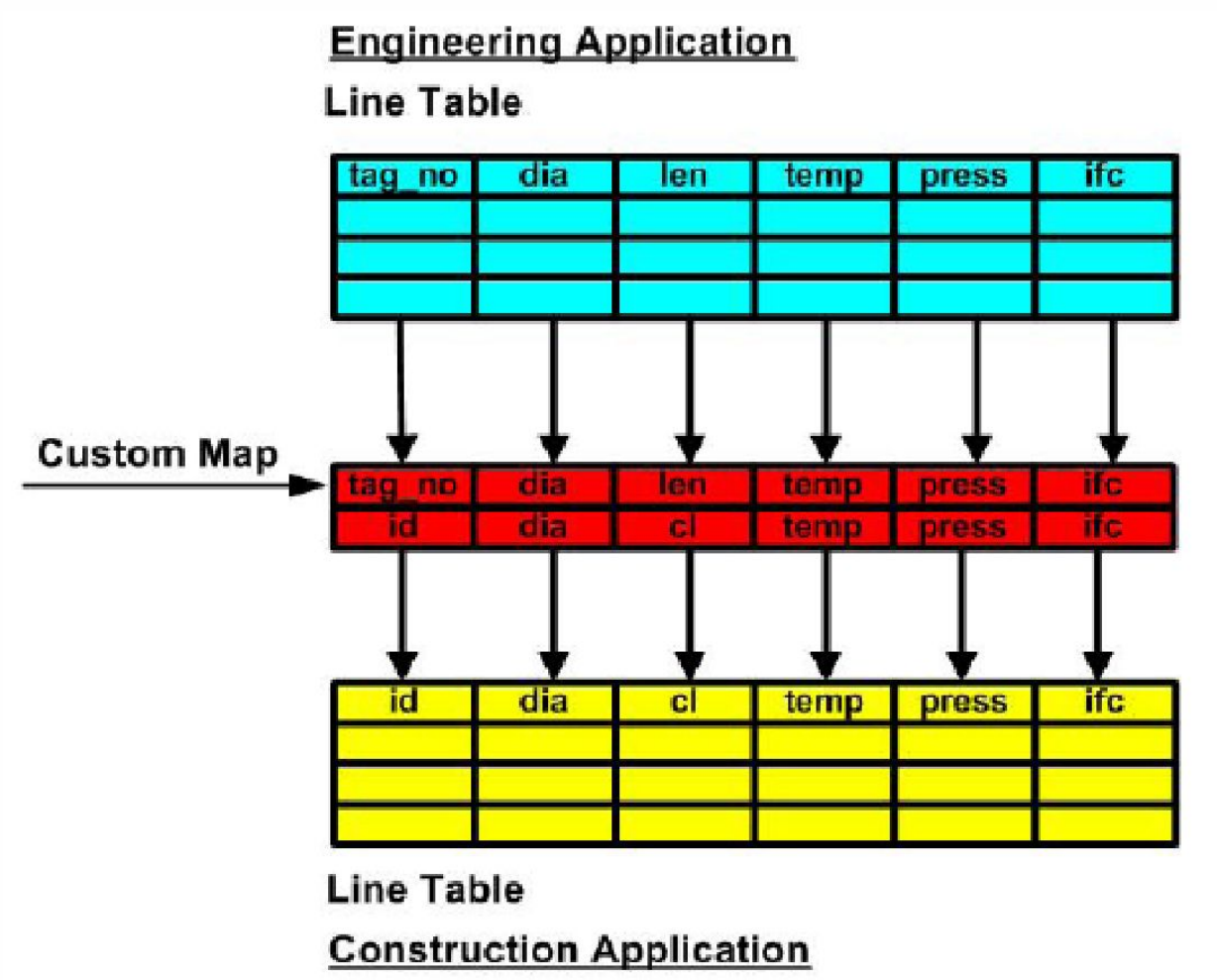
Использована диаграмма FIATECH

Реальная жизнь



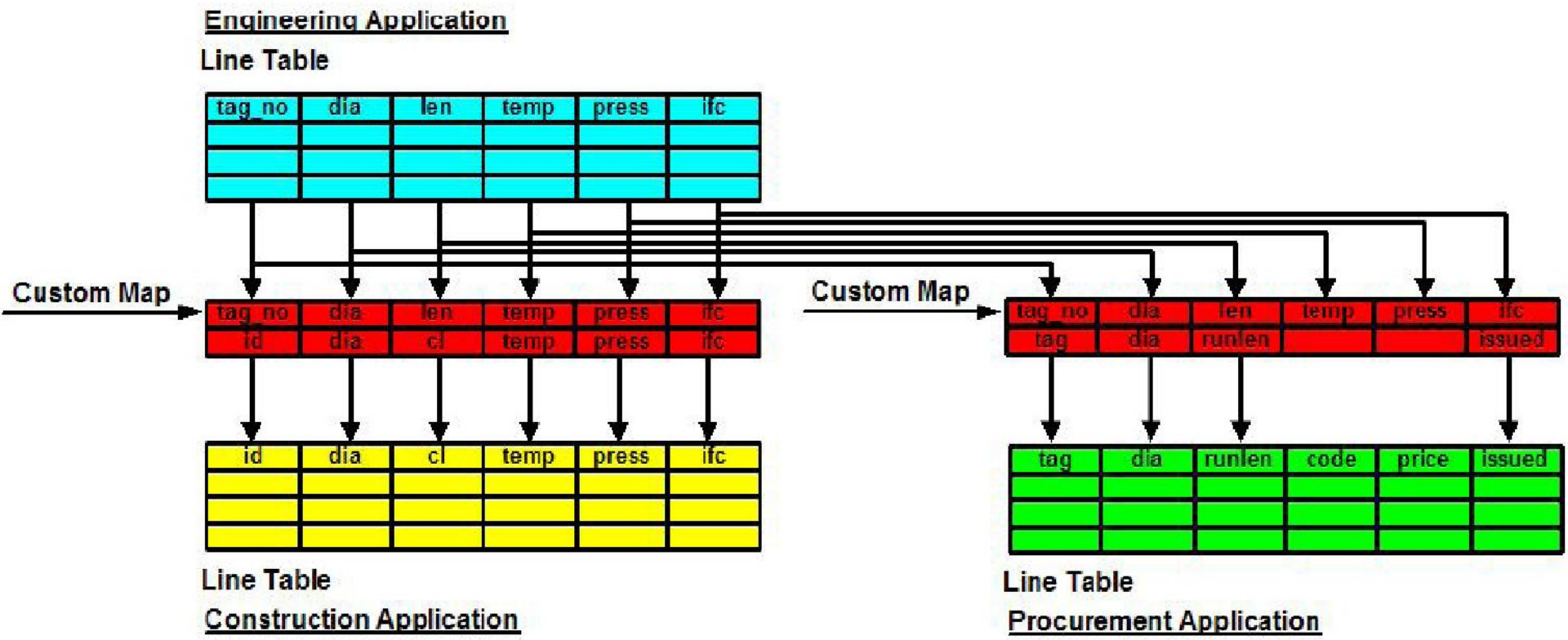
Использована диаграмма FIATECH

Частное решение



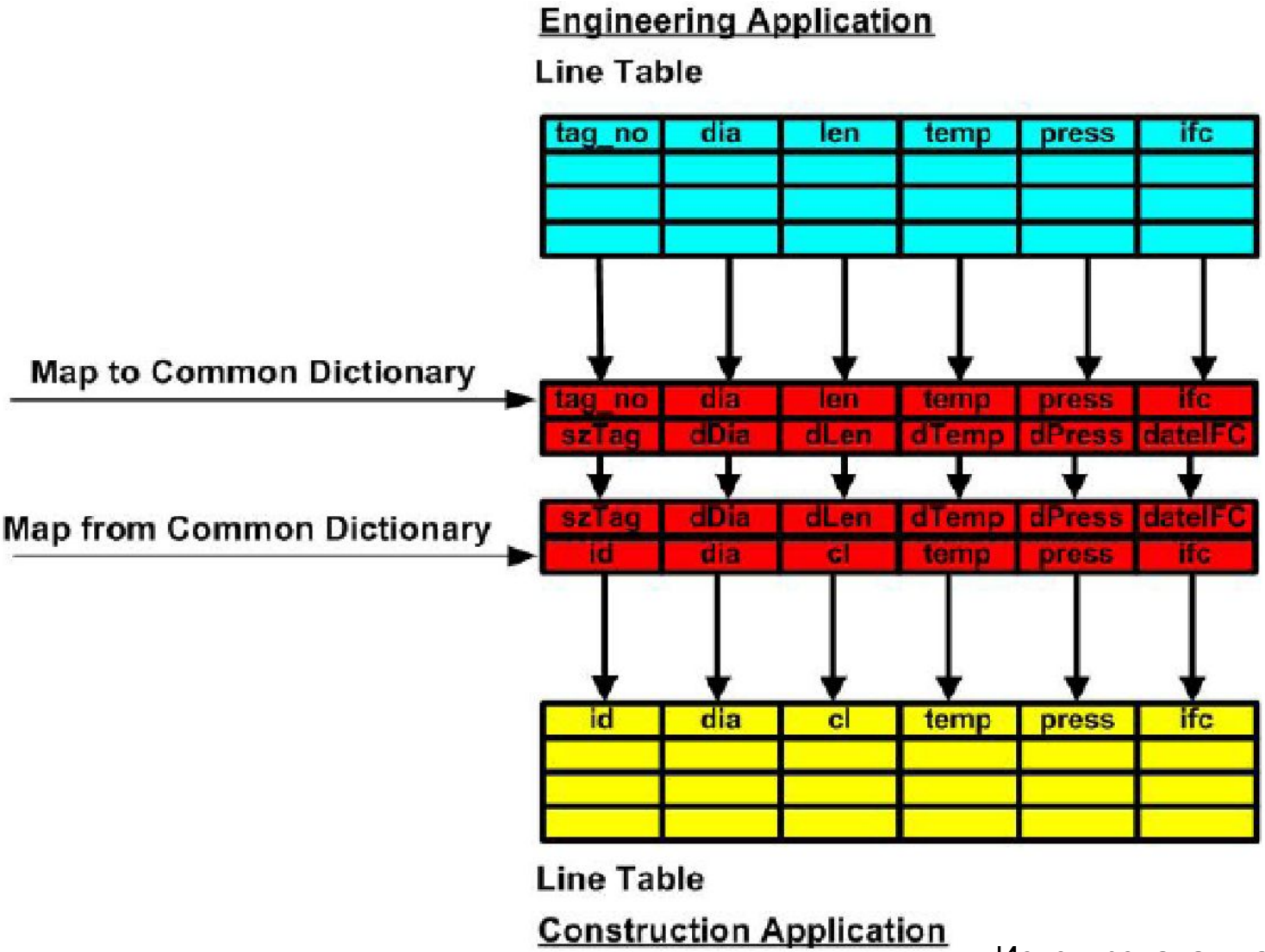
Использована диаграмма FIATECH

Много частных решений



Использована диаграмма FIATECH

Общий словарь



Использована диаграмма FIATECH

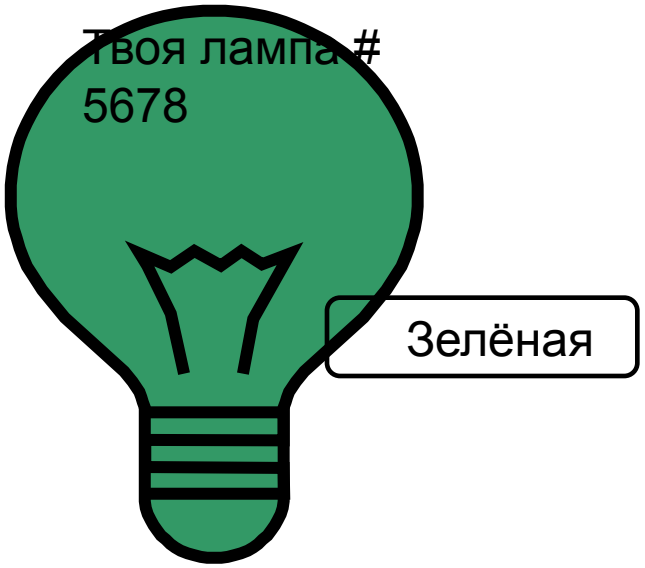
«Хорошая» модель

- Разделяемая
- Понятная
- Описывающая реальный мир
  - А какой мир «реален»?
  - Страшное слово – «онтология»

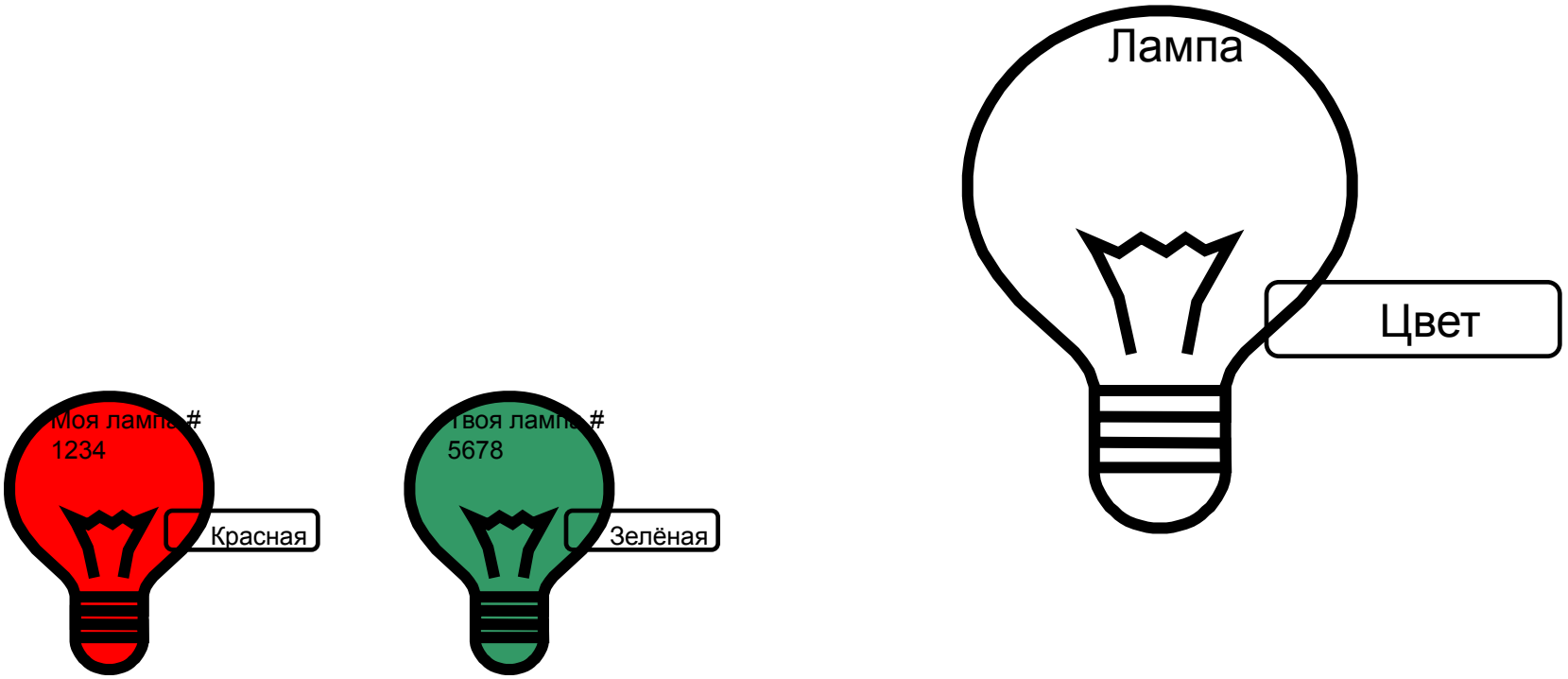
Традиционные парадигмы моделирования данных

Табличная	Компьютерных записей	Сущностная	Пример
Строка	Запись	Отдельная сущность	Моя машина ТТТ-123 99
Клетка	Поле	Отдельный атрибут	Красная
Таблица	Файл	Тип сущностей	Машина
Колонка	Тип поля	Тип атрибута	Цвет машины

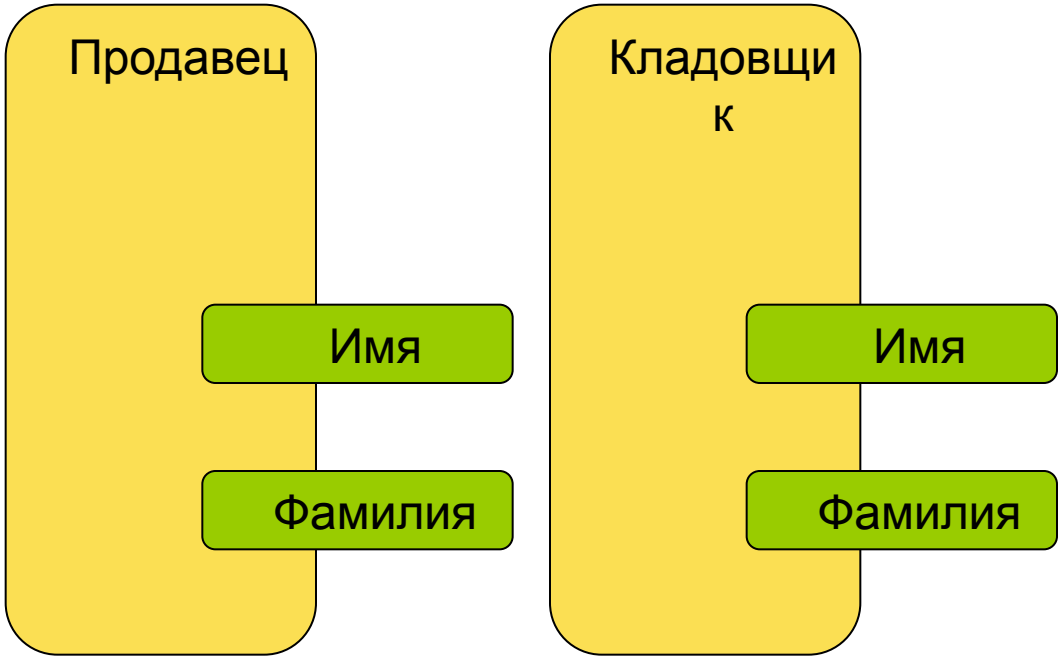
Предметы и атрибуты



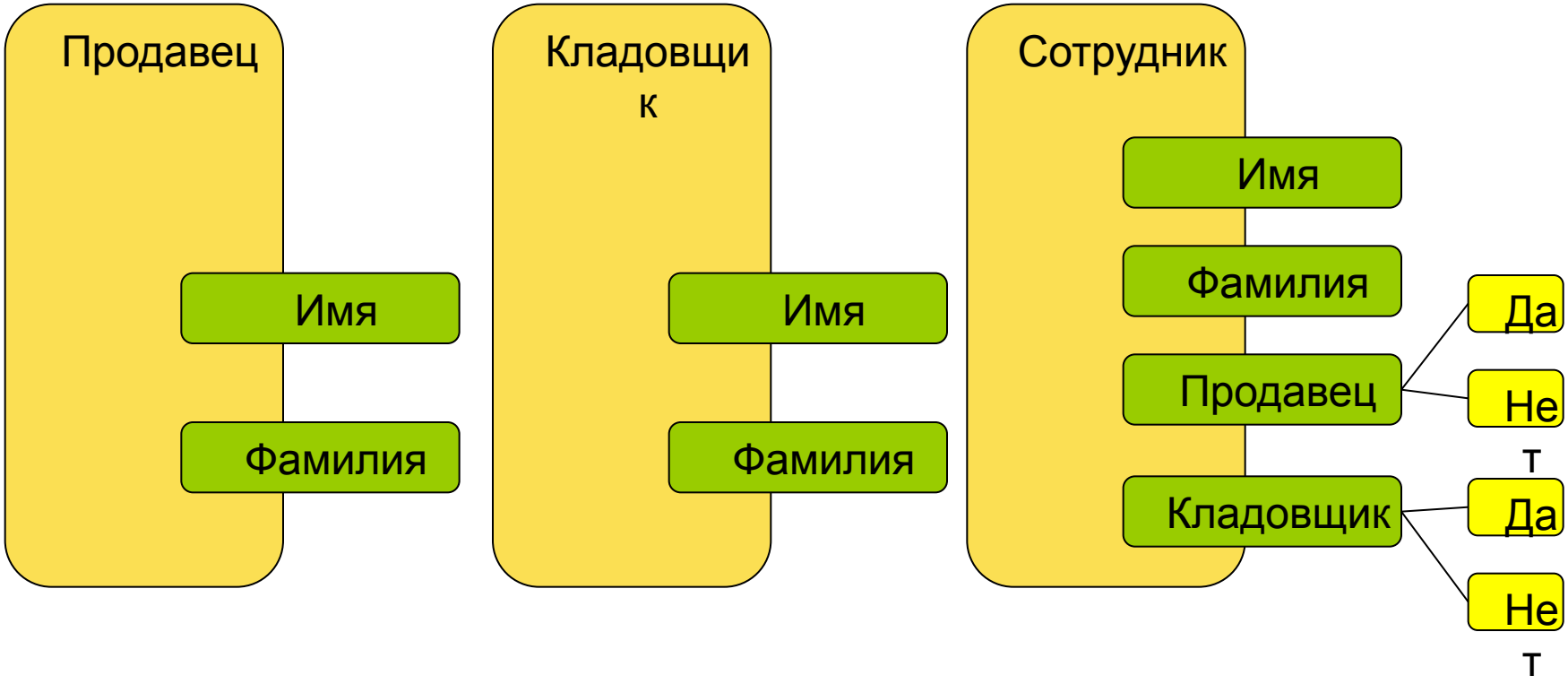
Типы сущностей и атрибутов



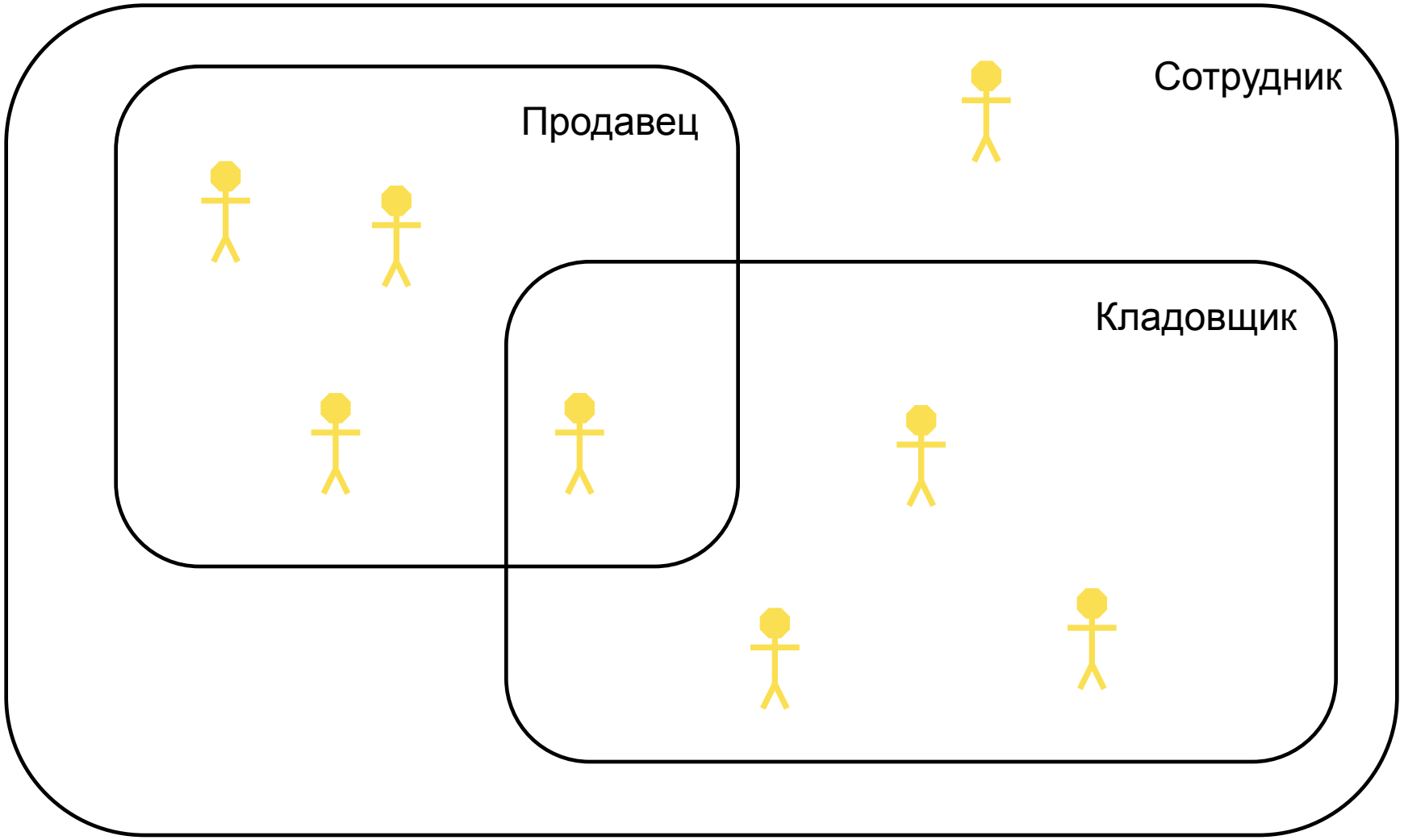
Проблема сущностей и атрибутов (1)



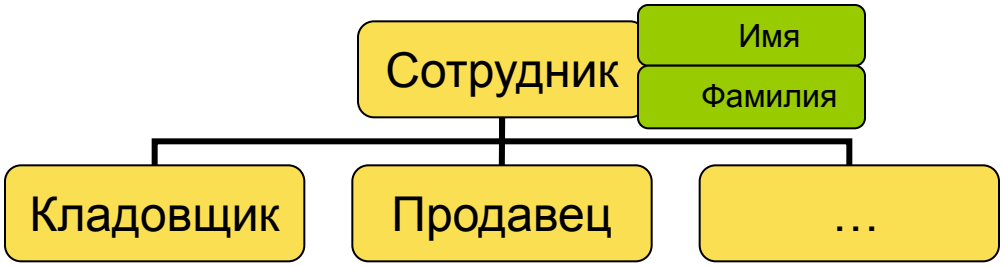
Проблема сущностей и атрибутов (2)



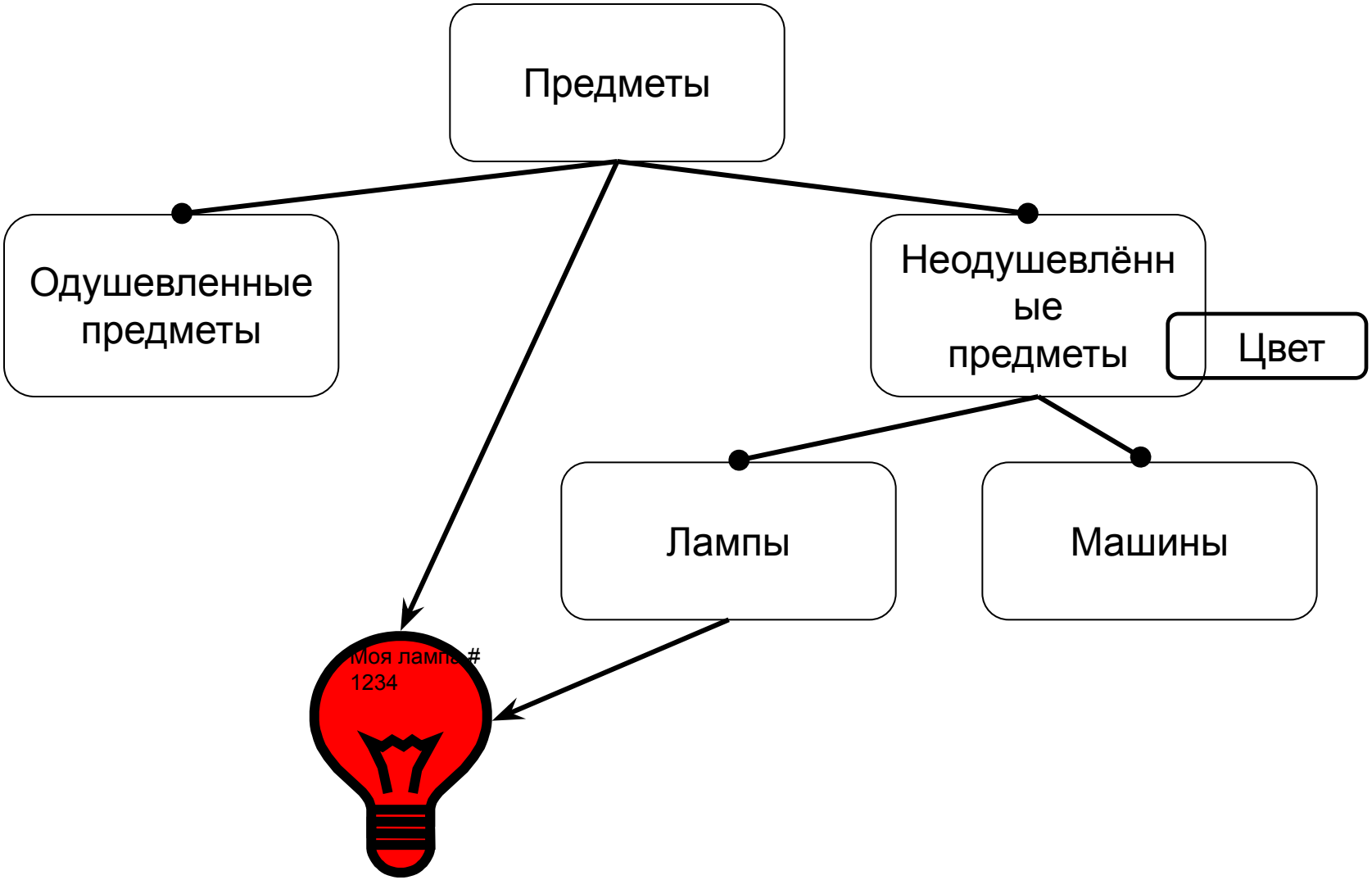
Сущности в реальном мире



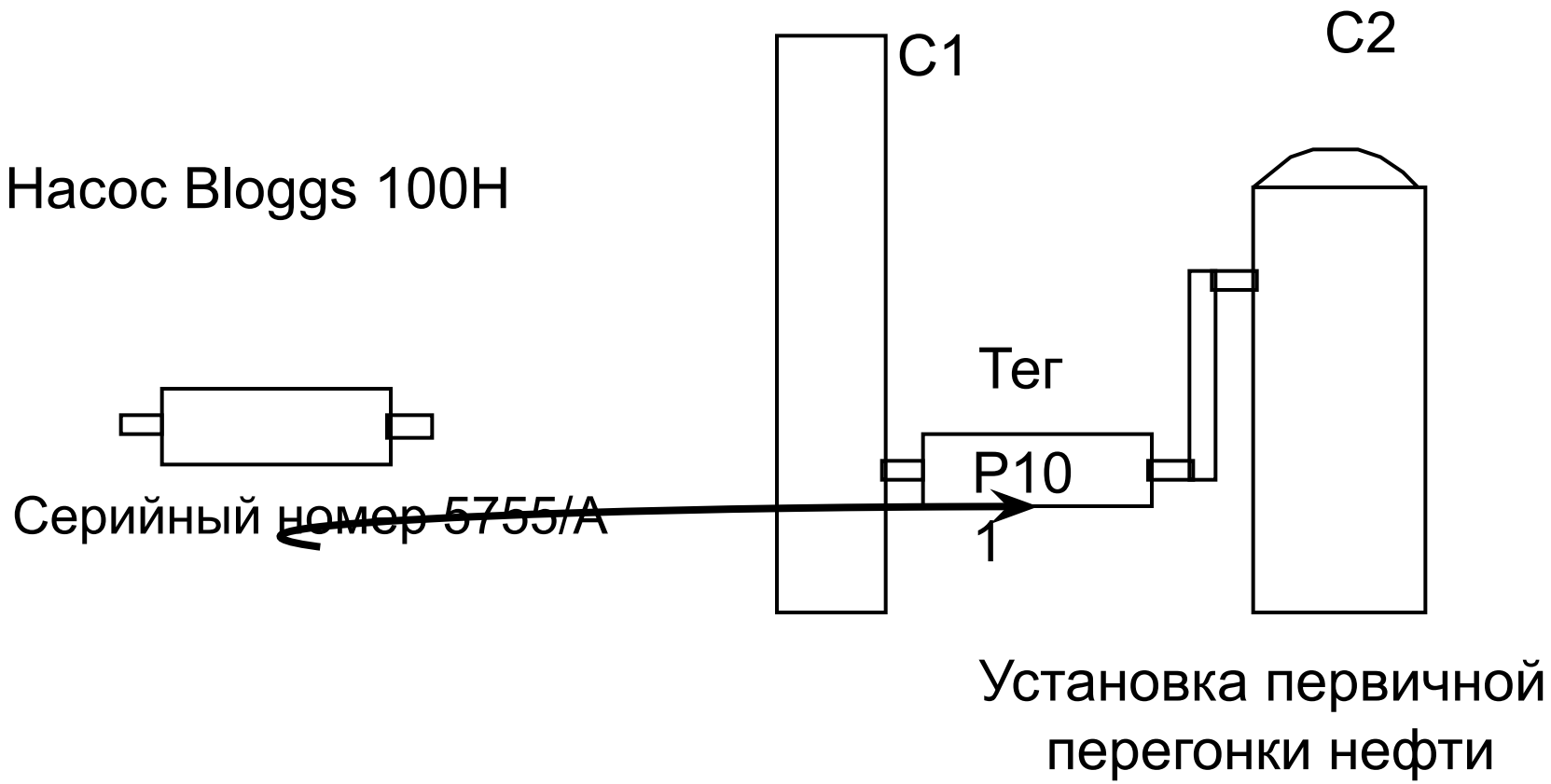
Субстанциональный подход



Аристотелева иерархия всего



Тег и серийный номер



Использована диаграмма Matthew West

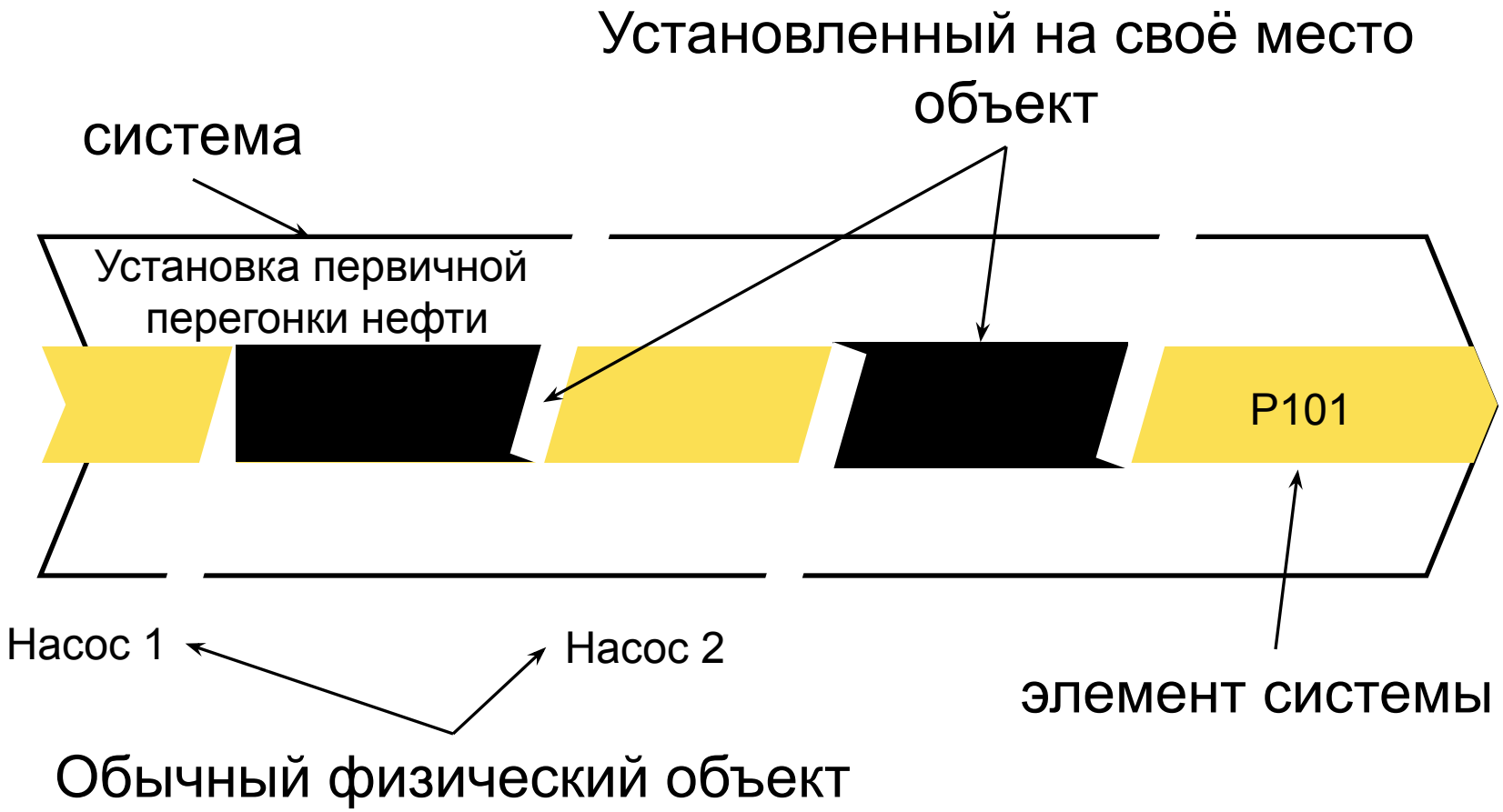
4D
+ экстенсионализм

Наряду с настоящим, существуют и прошлое, и будущее

1. Индивиды расположены как во времени, так и в пространстве , и имеют как темпоральные, так и пространственные части.
2. Если два индивида занимают одинаковую пространственно-временную область, они представляют собой одно и то же (экстенсионализм).

Использована диаграмма Matthew West

Пространственно-временная карта элемента системы

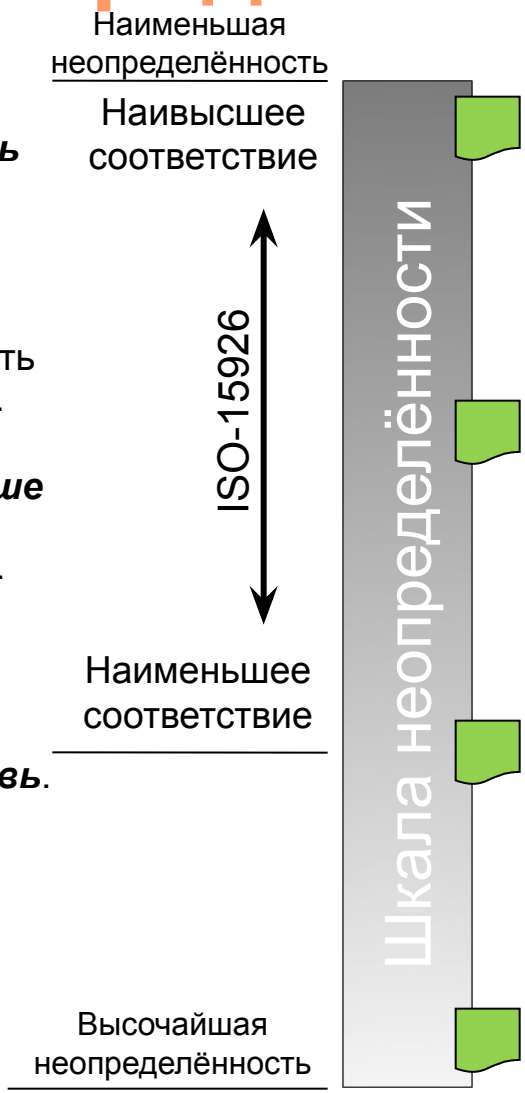


Использована диаграмма Matthew West

ISO 15926 – «английский» для данных жизненного цикла

Уменьшая неопределённость?

- Доступ к данным и обмены работают лучше всего, когда ***неопределённость*** исключена из деловых интерфейсов.
- Неопределённость между обменивающимися сторонами представляет ***риск***, и может потребовать ***значительных усилий для решения***.
- Сем выше неопределённость ***тем выше риск и затраты*** на реализацию действенного и рационального обмена.
- При появлении нового делового или технологического интерфейса могут появиться новые неопределённости, а ***затраты и риски – возникнуть вновь***.
- **Неопределённость = (Повторить) Затраты и(или) Риск**



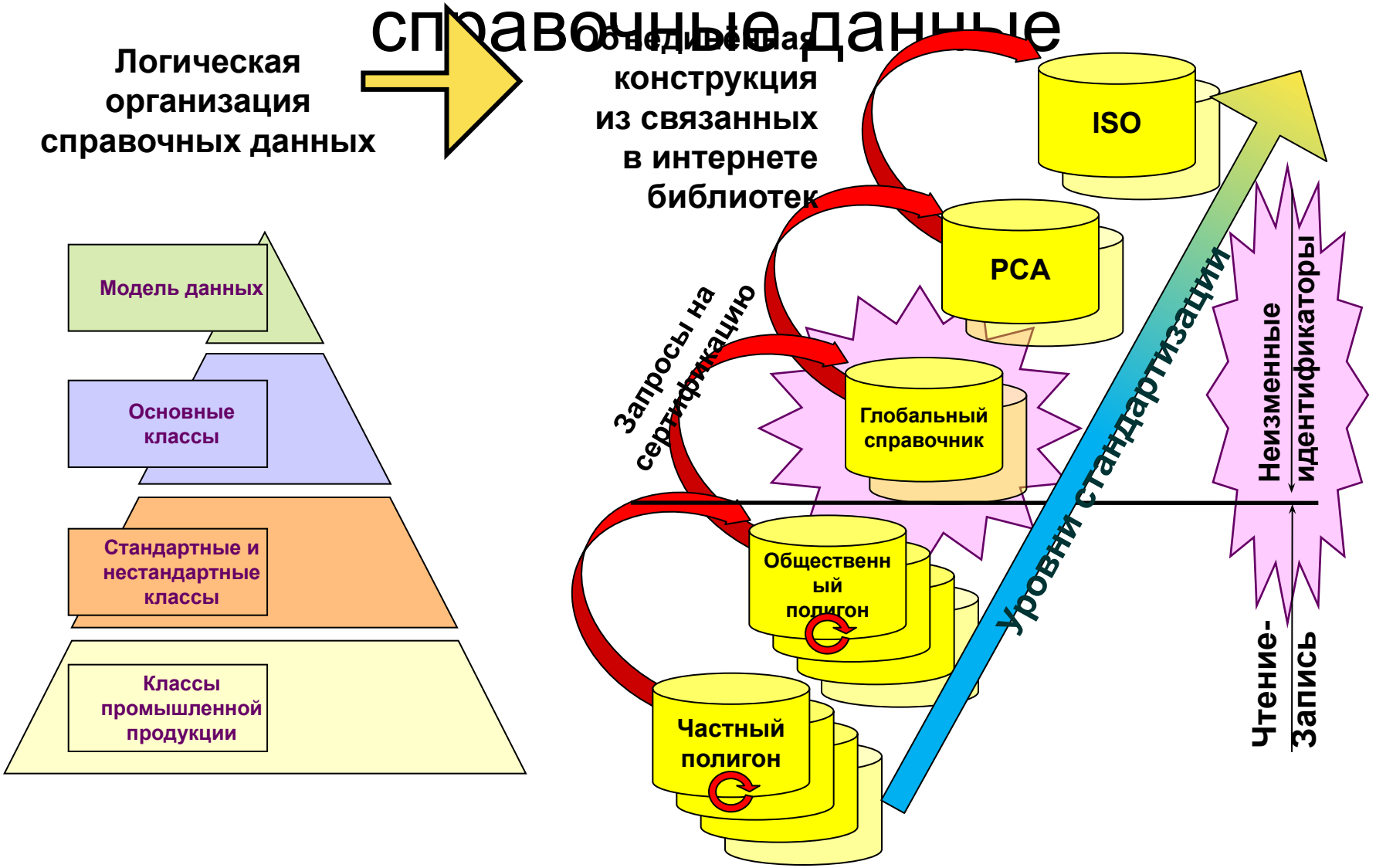
Если мы используем семантический веб, мы, наверное, можем автоматизировать тут ещё больше? То есть, “*технологии iRING*”

Поможет ли вам, если я расскажу, как я использовал данные?
То есть,*“образцы использования и шаблоны”*

Хорошо, давайте хотя бы договоримся использовать одинаковые термины.
То есть, *“общий словарь”*

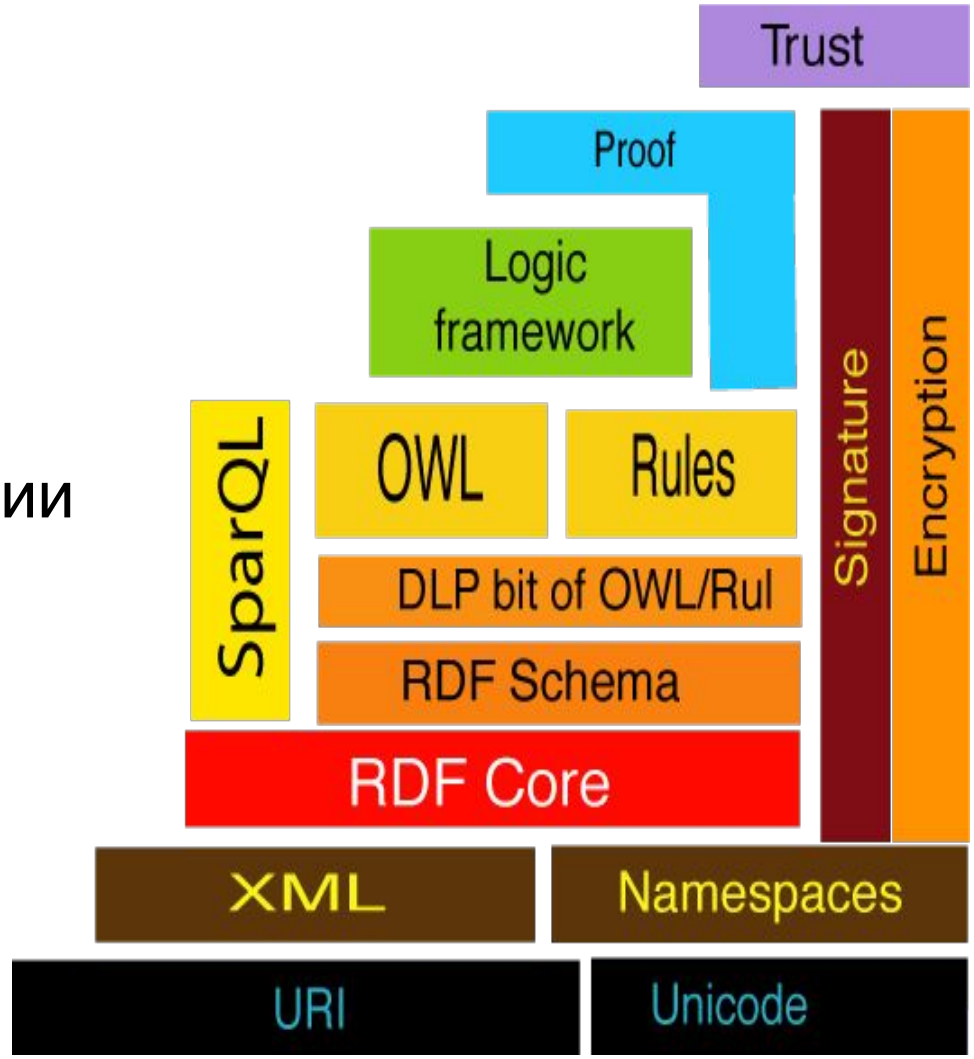
Я просто дам вам кое-какие данные. Вы как-нибудь разберётесь. То есть, *”это не моя проблема”*

15926 и объединённые справочные данные



Технологии представления знаний

С 2004 года язык представления онтологий OWL рекомендован консорциумом W3C в качестве основного средства описания онтологий. Тем же консорциумом W3C рекомендован стандарт представления информации RDF, как основа компьютерного описания знаний о мире в проектах, призванных объединить накопленные в интернете знания в единый семантический интернет



Триплет

- N-Triple
- Turtle
- Сериализация в XML

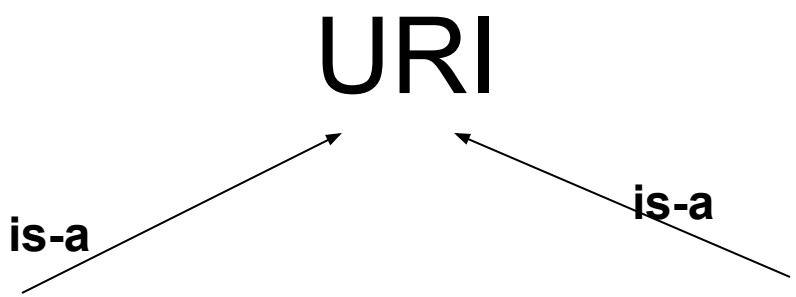
Суть одна: каждое утверждение – это триплет (triple) вида:

RDF

- Тройки <Subject, Predicate, Object>
- Направленный помеченный граф
- URI – уникальные обозначения
  - Пространства имён
- RDF Schema (RDFS) – набор фиксированных элементов, могущих быть S,O,P
- Форматы сериализации – XML, N3, Turtle

Идентификация

- Уникальные идентификаторы ресурсов: **URI** (Unified Resource Identificator)
- **Unicode** – способ представления строк в национальных кодировках (= нет привязки к латинице)
- **URI** + поддержка **Unicode** в идентификаторах ресурсов = **IRI**: (International Resource Identificator)



URL – Uniform resource Locator

URN – Uniform Resource Name

http://www.amazon.com/Foundations-Semantic-Technologies-Textbooks-Computing/dp/142009050X

urn:isbn:978-1-4200-9050-5

Идентификатор конкретной книги по её адресу в он-лайн магазине Amazon

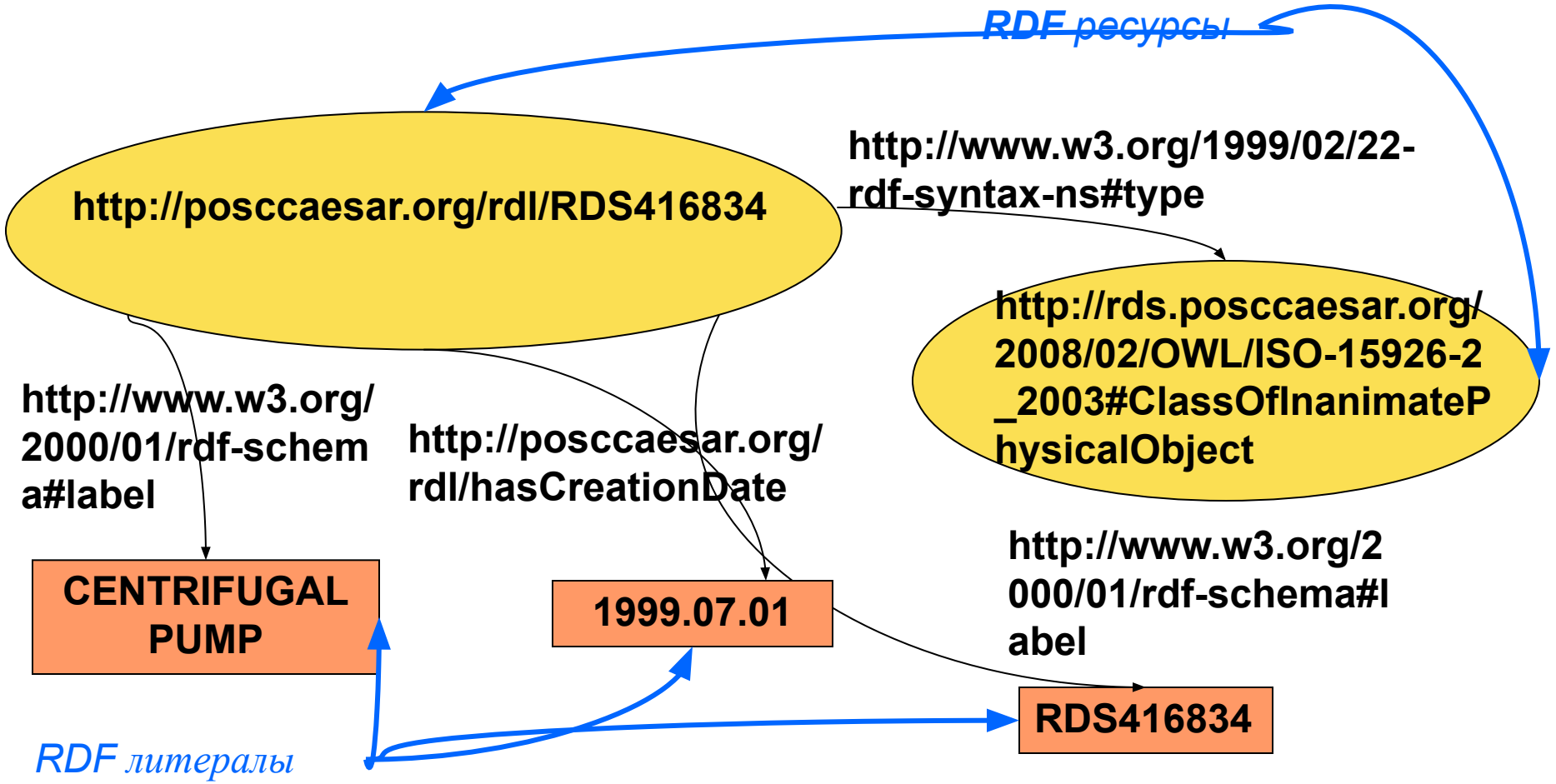
Идентификатор конкретной книги по ISBN (где находится сама книга - неизвестно)

Итак, RDF

- RDF – Resource Description Framework – Среда описания ресурсов (в Сети)

  Сеть моделируется как гиперграф (точнее, Giant Global Graph, GGG), вершинами которого являются ресурсы произвольной природы (в том числе, литералы), а дугами – связи (или ссылки) между ресурсами

Пример RDF графа



В описании дуг используются как специальные словари, созданные для моделей данных в промышленности (https://www.posccaesar.org/wiki/Rds/), так и общие словари, например, словарь описания ресурсов в сети, RDF Schema

RDF на Turtle

```
@prefix part2    <http://rds.posccaesar.org/2008/02/OWL/ISO-15926-2_2003#/>.
@prefix rdl  <http://posccaesar.org/rdl/> .
@prefix rdfs      <http://www.w3.org/2000/01/rdf-schema#/> .
@prefix rdf  <http://www.w3.org/1999/02/22-rdf-syntax-ns#/> .
@prefix lib   <http://www.deri.ie/library/0.1/>
@prefix dc   <http://purl.org/dc/elements/1.1/>
```

Субъекты	*Предикаты*	*Объекты*
rdl:RDS416834	rdl:hasIdPCA	“RDS416834”.
rdl:RDS416834	rdl:hasCreationDate	“1999.07.01”.
rdl:RDS416834	rdfs:label	“CENTRIFUGAL PUMP”.
rdl:RDS416834	rdf:type	part2:ClassOfInanimatePhysicalObject.

Символ « . » - конец трипла

RDF в XML

```
<?xml version="1.0" encoding="utf-8">
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
      xmlns:rdl="http://posccaesar.org/rdl/">

<rdf:Description
    rdf:about="http://rds.posccaesar.org/2008/06/OWL/RDL#RDS416834">
    <rdl:hasIdPCA>RDS416834</rdl:hasIdPCA>
    <rdl:hasCreationDate>1999.07.01</rdl:hasCreationDate>
    <rdfs:label>CENTRIFUGAL PUMP</rdfs:label>
</rdf:Description>
</rdf:RDF>
```

Ещё более детальная типизация ресурсов: язык OWL

OWL = Web Ontology Language

Язык разработан для более детального описания групп ресурсов в сети

Разработан так, чтобы по исходной, частичной, классификации некоторой группы ресурсов можно было получить (с помощью логического машинного вывода!) полную классификацию этой группы ресурсов

Используется везде - в науке, в бизнес-приложениях, при описании ресурсов в Интернет (Web.2.0, Semantic Web, Web of Data...), когда нужно точно описать семантику ресурса

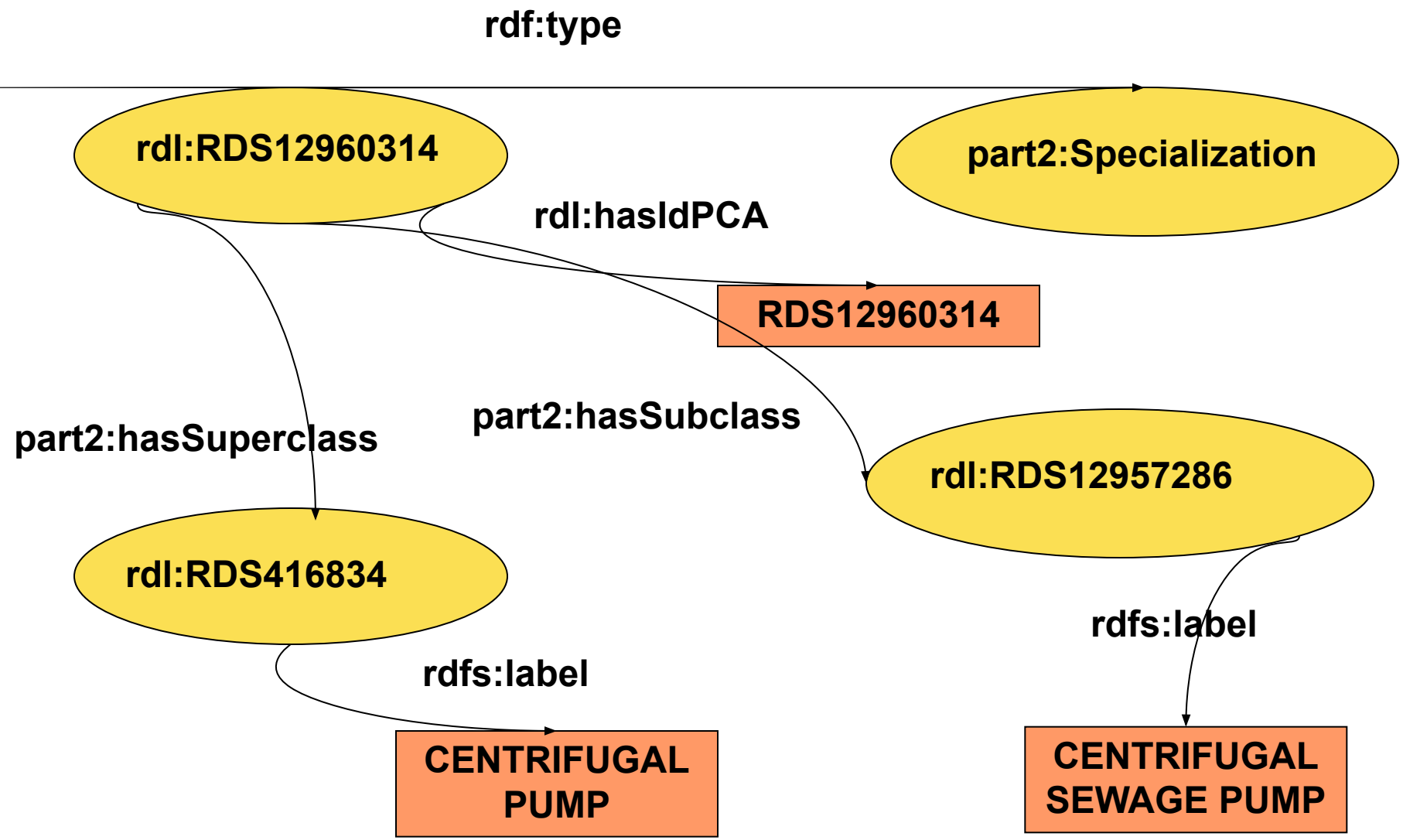
Стандартное пространство имён для OWL

xmlns:owl = “http://www.w3.org/2002/07/owl#”

Что можно “сказать” в OWL по сравнению с RDF Schema

- Что есть класс ресурсов, к которому можно применить логический машинный вывод owl:Class
- Два (или более класса) ресурсов
  - Эквивалентны owl:equivalentClass
  - Не имеют общих элементов owl:disjointWith
  - Объединяются/пересекаются в новый класс owl:unionOf / owl:intersectionOf
- Данный класс ресурсов задаётся перечислением его экземпляров owl:oneOf
- Объявить сложный безымянный класс owl:Restriction
- Объявить, что объект связи (в трипле <субъект, предикат, объект>) должен быть непременно ресурсом (owl:objectProperty) или непременно литералом (owl:datatypeProperty)
- Объявить, что количество ресурсов, участвующих в связи, равно (owl:cardinality), больше (owl:minCardinality) или меньше (owl:maxCardinality) определенного числа

Специализация для CENTRIFUGAL PUMP: RDF-граф



RDF хранилища

RDF triplestore (RDF хранилище, хранилище триплов) – база данных, (грубо) состоящая из двух таблиц:

1) таблица целочисленных идентификаторов для всех используемых URI
   URI (Code int not null, URI uri)
2) таблица квадов
   Quad (Graph int not null, Subject int not null,
   Predicate int not null, Object any not null)
3) индексы GSPO, PGOS, OGPS, SPGS
4) view, связывающий таблицу квадов с таблицей идентификаторов URI и возвращающий квады в читабельном виде.

Т.е. хранятся не триплы, а квады (quads, “четвёрки”)!

По структуре триплстора благодарность Ивану Михайлову,
http://forum.semanticfuture.net/viewtopic.php?id=74

SPARQL

- SPARQL – это SPARQL Query Language for RDF – язык запросов для RDF
- Описан здесь
- http://www.w3.org/TR/rdf-sparql-query/
- Похож на SQL

ABBYY Compreno

AABBYY Syntactic and Semantic Parser выполняет точный и подробный анализ текстов на русском и английском языках, создавая прочный фундамент для решения главной задачи приложения на высоком уровне.

JОбласть применения

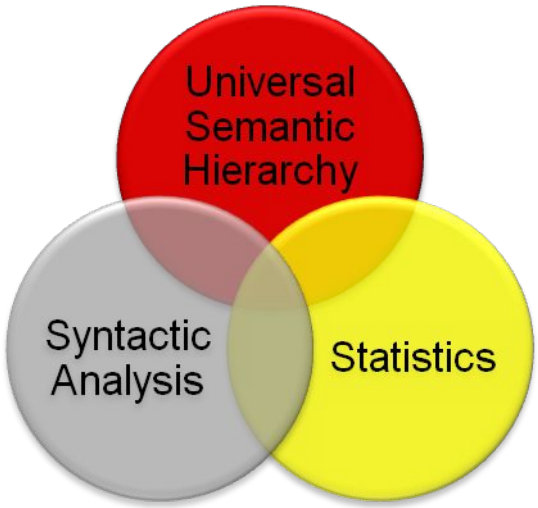
●- Интеллектуальный корпоративный поиск

●- Автоматическое реферировании документов

●- **Извлечение фактов из больших объемов информации**

●- Мониторинг СМИ и социальных сетей с последующим анализом тональности найденных сообщений

●- Другие приложения, включающие анализ текстов



.15926 Editor: инструментарий ISO 15926

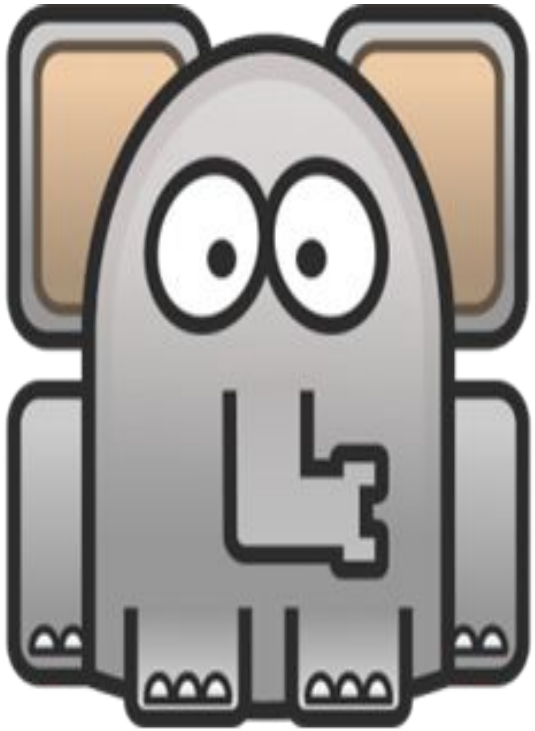
Система онтологического программирования:

- просмотр, создание, поиск и преобразование данных в форматах стандарта ISO 15926;
- поддержка множественности неймспейсов, работа с серверами SPARQL;
- консоль онтологического программирования на языке Python;
- распознавание онтологических паттернов;

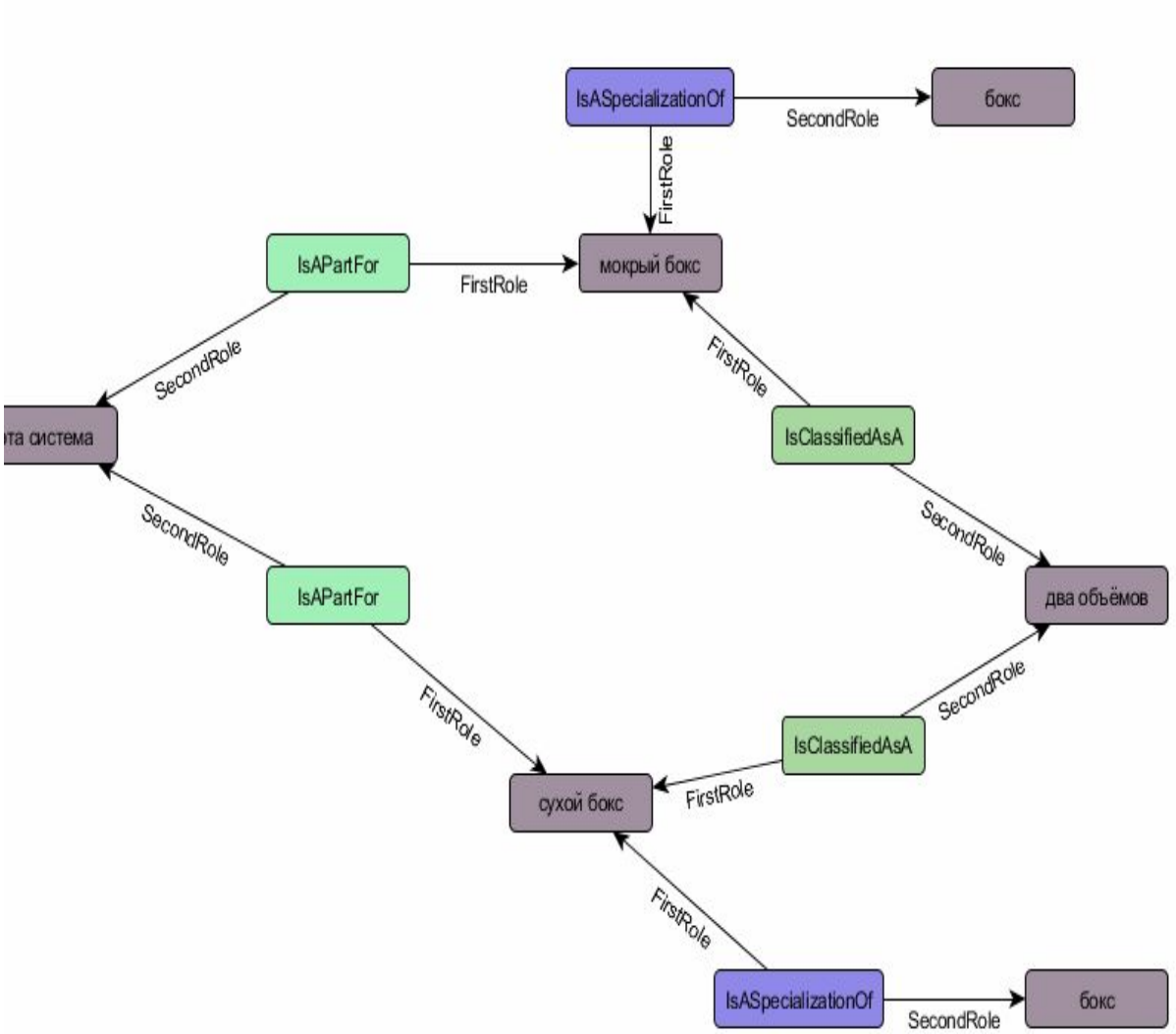
Работа будет продолжаться в направлениях:

- разработка адапторов для различных инженерных (и не только инженерных) применений;
- интерфейсная поддержка exploratory programming;
- развитие возможностей онтологического программирования (подъем уровня языка работы с онтологическими данными, разработка верификаторов, reasoners, средств эволюции онтологий и т.д.).

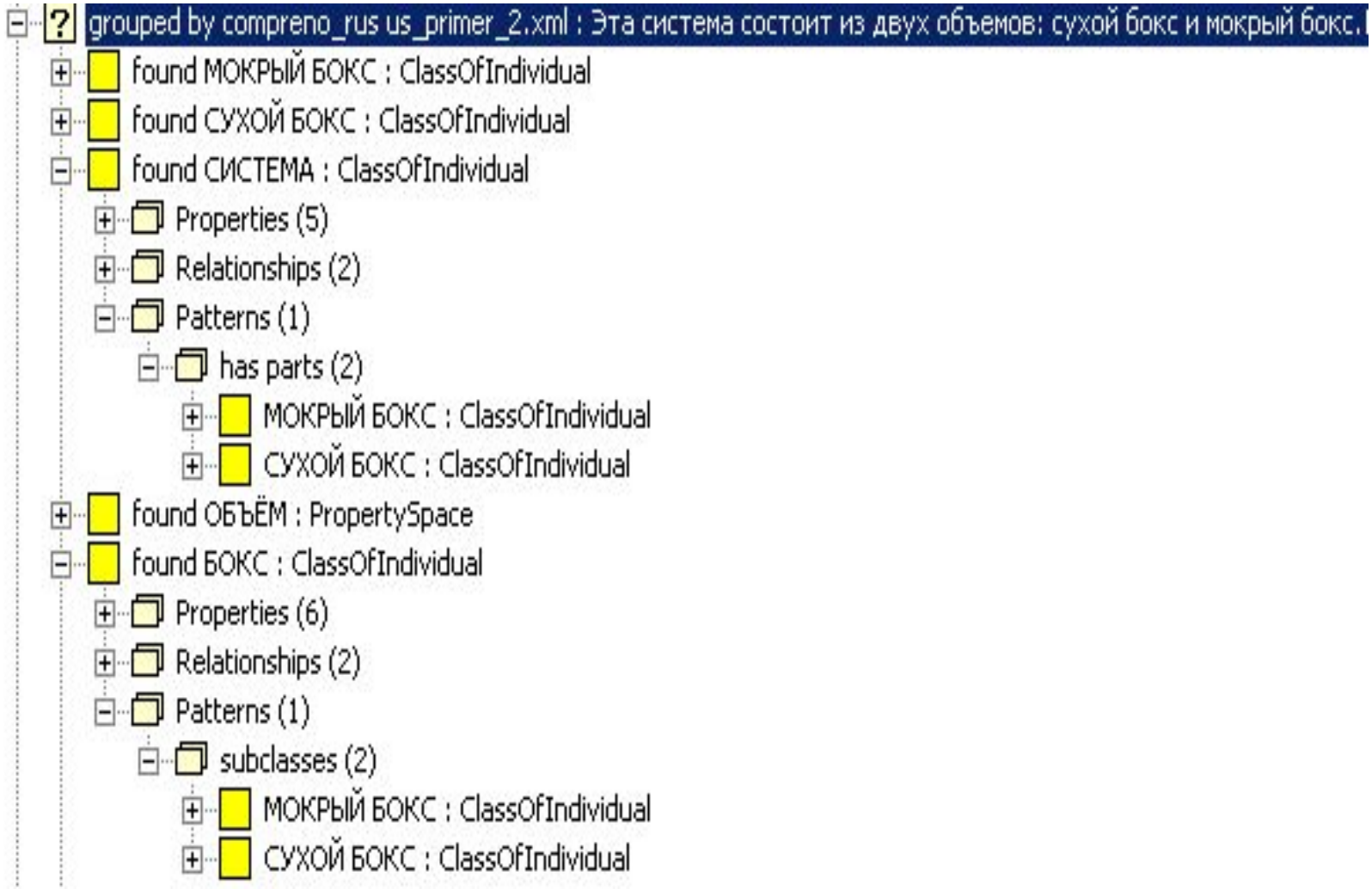
Скачать с http://techinvestlab.ru/dot15926Editor/



Извлечение онтологической информации "Эта система состоит из двух объемов: сухой бокс и мокрый бокс" (ABBYY Compreno).

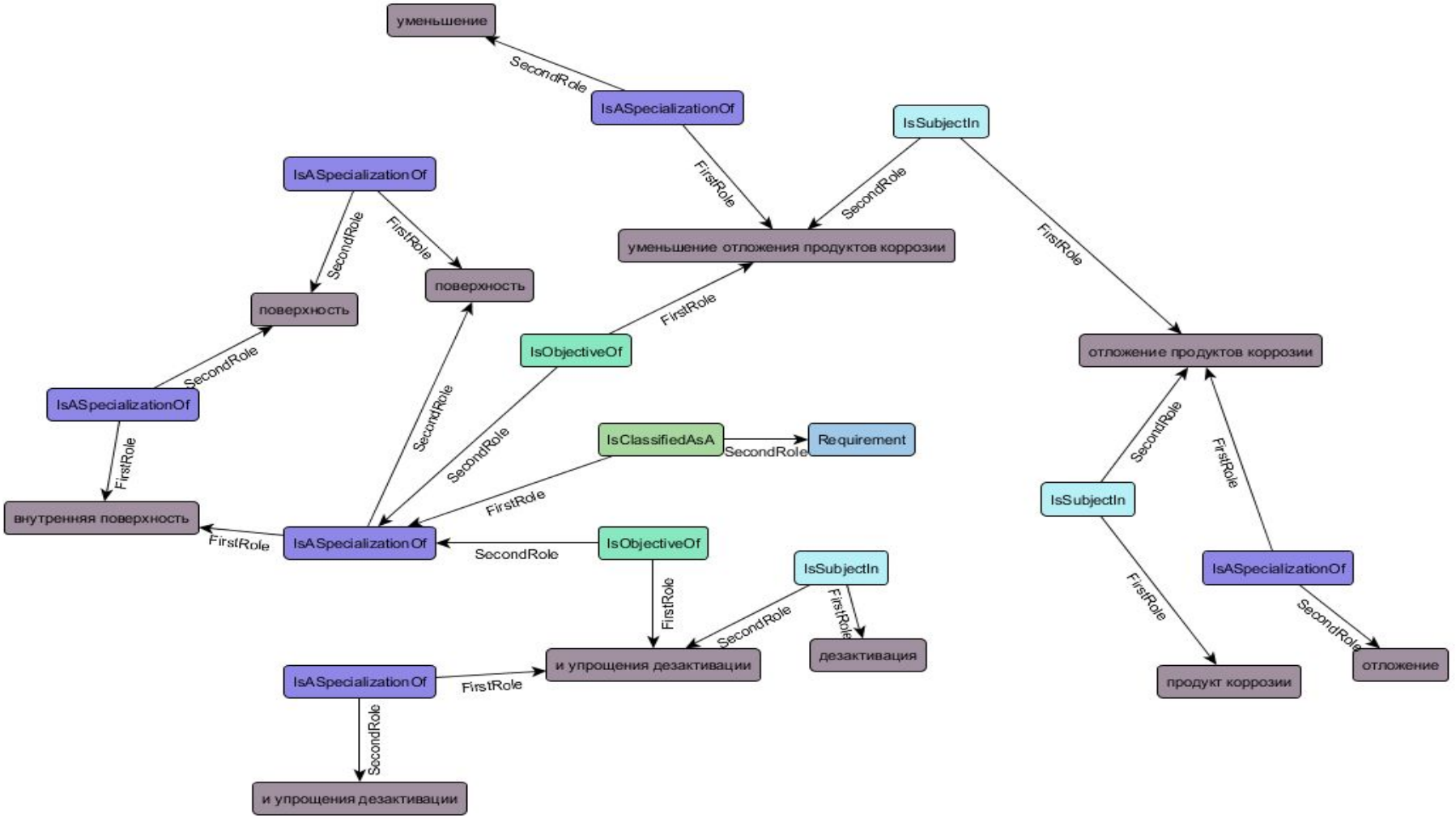


Формирование справочных данных на основе онтологического разбора (.15926 Editor)



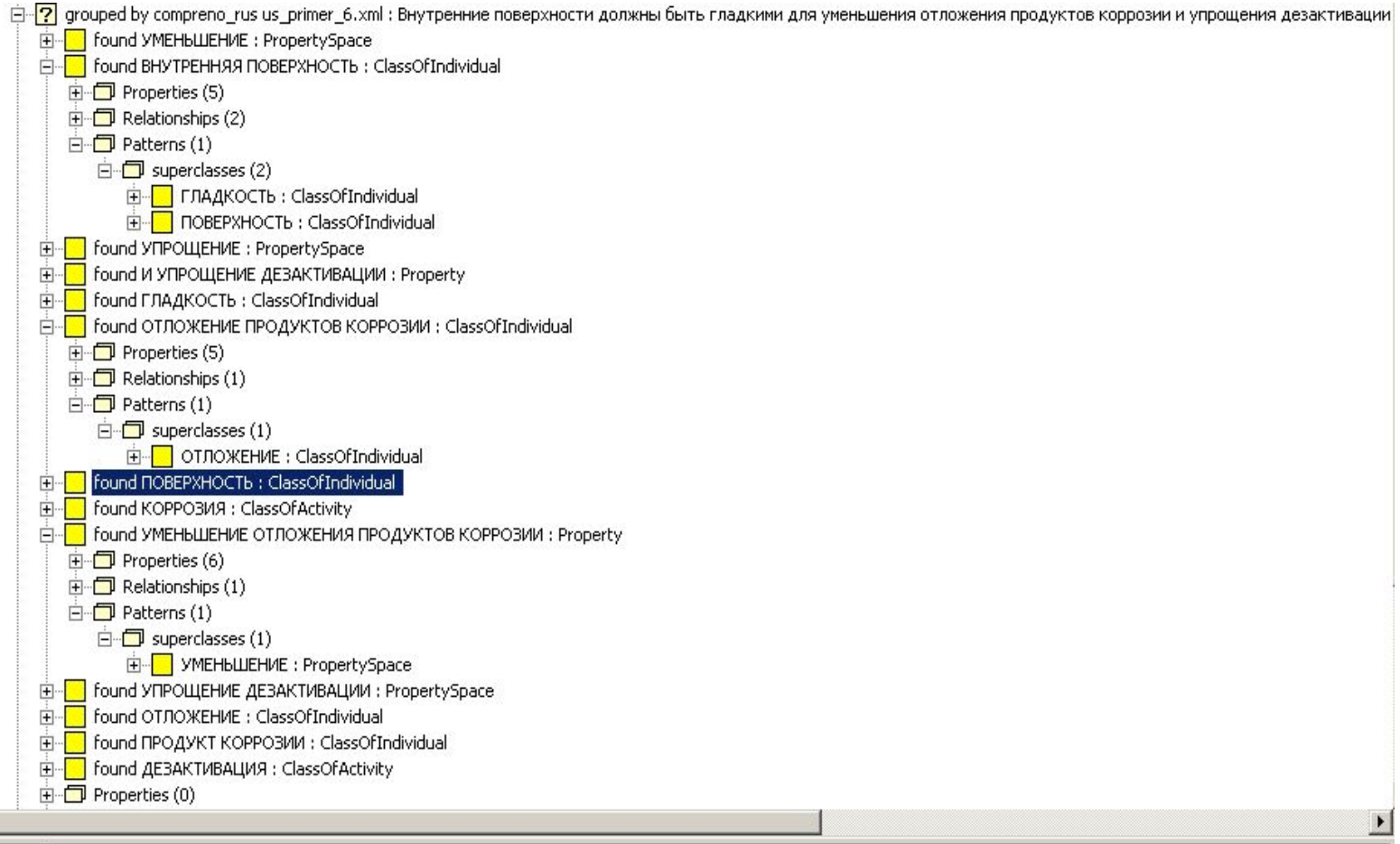
*

Извлечение онтологической информации "Внутренние поверхности должны быть гладкими для уменьшения отложения продуктов коррозии и упрощения дезактивации" (ABBYY Compreno).



*

Формирование справочных данных на основе онтологического разбора (.15926 Editor)



Property	Value
Source name	abbyyrus.rdf
URI	http://example.org/rdl#id1221a9b6-6dd7-4abc-aa8f-60b5373bcf95
Name	ПОВЕРХНОСТЬ
annCreationD	2012-11-15 03:03:33.906000

*

Спасибо за внимание!

Анатолий Левенчук,
http://ailev.ru
ailev@asmp.msk.su
Президент Русского отделения INCOSE
Член исполкома Русского отделения SEMAT

Виктор Агроскин
vic5784@gmail.com
Член экспертной группы ISO TC184/SC4/WG3

ТехИнвестЛаб.ру
(POSC Caesar member)
+7 (495) 748-5388

.15926 Editor http://techinvestlab.ru/dot15926Editor