Лекции 15-16 «**Проектирование интеграционных решений**»

Овчинников П.Е.
МГТУ «СТАНКИН»,
ст.преподаватель кафедры ИС

27.03.2019

Дальше

Лекция 15
«Справочники и классификаторы»

Овчинников П.Е.
МГТУ «СТАНКИН»,
ст.преподаватель кафедры ИС

27.03.2019

Терминология: классификация

Классифика́ция, также ***классифици́рование*** (от лат. *classis* «разряд» и *facere* «делать») — понятие в науке (в философии, в формальной логике и др.), обозначающее разновидность **деления объёма** понятия по **определённому основанию (признаку, критерию)**, при котором объём родового понятия (класс, множество) делится на виды (подклассы, подмножества), а виды, в свою очередь делятся на подвиды и т.д.

Правила:

- классификацию необходимо проводить **только по одному** конкретному основанию
- необходимо соблюдать **соразмерность деления**, т.е. сумма членов классификации должна равняться объёму родового понятия (класса, множества), возможные ошибки при несоблюдении данного правила:
  - неполная (узкая) классификация
  - классификация с лишними видовыми понятиями.
- члены классификации должны **взаимно исключать** друг друга.
- подразделение на подклассы должно быть **непрерывным**

Классификация (Википедия)

Терминология: классификация

Распознавание образов — это **отнесение исходных данных** к **определенному классу** с помощью выделения **существенных признаков**, характеризующих эти данные, из общей массы несущественных данных

Классическая постановка задачи распознавания образов

1. Дано множество объектов, относительно которых необходимо провести классификацию
2. Множество представлено подмножествами, которые называются классами.
3. Заданы:
   - информация о классах
   - описание всего множества и
   - описание информации об объекте, принадлежность которого к определенному классу неизвестна
4. Требуется по имеющейся **информации о классах** и **описании объекта** установить - к какому классу относится этот объект

Распознавание образов (Википедия)

Терминология: множество

Мно́жество — одно из ключевых понятий математики; это математический объект, сам являющийся набором, совокупностью, собранием каких-либо объектов, которые называются **элементами** этого множества и обладают общим для всех их характеристическим свойством

Изучением общих свойств множеств занимаются теория множеств, а также смежные разделы математики и математической логики

Множество может быть:

- пустым и
- непустым

- упорядоченным и
- неупорядоченным

- конечным и
- бесконечным, бесконечное множество может быть
  - счётным или
  - несчётным

Множество (Википедия)

Объектно-ориентированный подход (инженерия знаний)

Базовыми видами отношений для представления фреймов в виде семантической сети являются:

• отношение **AKO** (англ. a kind of), являющееся отношением **классификации** и позволяющее строить иерархические связи между объектами, реализующие основные принципы **наследования** свойств объектов

общее <- (AKO) – частное (только абстракции)

• отношение **HasPart** (англ. has part), являющееся отношением **вхождения** и позволяющее **декомпозировать** сложные объекты на их составляющие

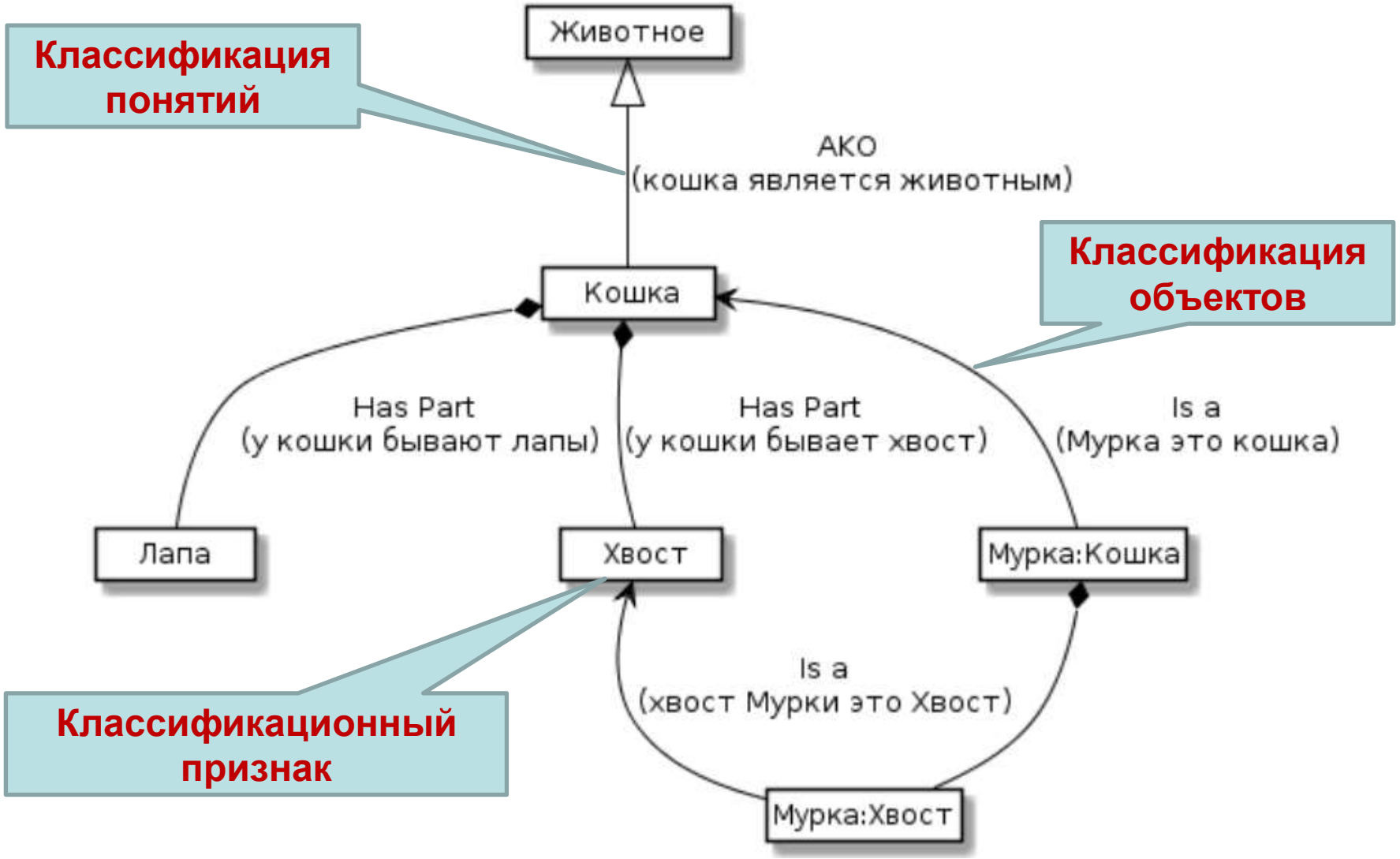
целое <- (HasPart) – часть (только абстракции)

• отношение **ISA** (англ. is a), являющееся отношением между фреймом-**образцом** и фреймом-**экземпляром** , появляющимся **в результате классификации** (отнесения конкретного объекта к абстрактному классу)

абстрактное <- (ISA) – конкретное

Семантическая сеть Минский М. Фреймы для представления знаний. М.: Мир, 1979.

Объектно-ориентированный подход (UML)



Семантическая сеть

Терминология ООП: классы и объекты

Класс (class) - категория вещей, которые имеют общие **атрибуты** и **операции**

Объект (object) -

- конкретная **материализация** абстракции
- сущность с хорошо определенными границами, в которой **инкапсулированы состояние и поведение**
- **экземпляр** класса

Объект **уникально идентифицируется** значениями атрибутов, определяющими его состояние в данный момент времени

Атрибут класса (class attribute) -
именованное свойство класса, описывающее **множество значений**, которые могут принимать экземпляры этого свойства

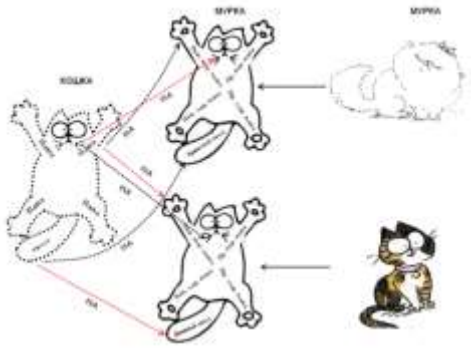
Операция класса (class operation) -
метод или функция, которая может быть выполнена экземпляром класса или интерфейсом

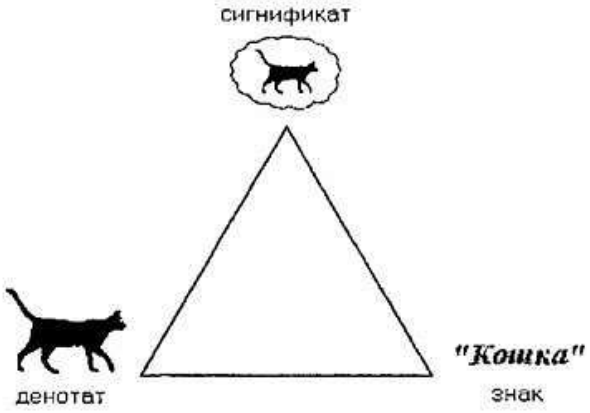
Терминология ООП: смысл и значение

Смысл — сущность феномена в более широком контексте реальности. Смысл феномена оправдывает существование феномена, так как определяет его место в некоторой целостности, вводит отношения «часть-целое», делает его необходимым в качестве части этой целостности

Смыслом также называют мнимое или реальное предназначение каких-либо вещей, слов, понятий или действий, заложенное конкретной личностью или общностью

Г. Фреге в своей статье «О смысле и значении» (1892) противопоставил смысл (нем. *Sinn*) и значение (нем. *Bedeutung*, денотат), хотя в немецком языке эти слова иногда использовались как синонимы. Если значение (денотат) — это сам обозначаемый предмет, то смысл — это информация о предмете (сигнификат или десигнат)





Смысл (Википедия)

Терминология ООП: отображение (mapping)

Мапирование (англ. *data mapping*, иногда маппинг, маппирование, мэппинг, но не путать с маппингом игровых уровней) — определение **соответствия данных** между потенциально **различными семантиками** одного объекта или разных объектов

Термин понимается очень широко от отображения одной последовательности элементов на другую последовательность до банальной конвертации файлов

Рассматриваемому термину по смыслу ближе всего англоязычный термин data mapping

Например, один объект (база данных) для обозначения элементов использует обозначения «да» и «нет», а другой объект (другая база данных) для обозначения этих же объектов использует обозначения «1» и «0», то есть коды одной базы данных мапируются в соответствии с кодами другой базы данных

Мапирование (Википедия)

Терминология ООП: отображение (mapping)

```
1. <xsd:element name="Кошка">
2.         <xsd:complexType>
3.                 <xsd:sequence>
4.                 <xsd:element name="Имя" type="xsd:string"/>
5.                 <xsd:element name="Лапы">
6.                         <xsd:element name="Шерсть" type="xsd:string"/>
7.                         <xsd:element name="Когти" type="xsd:boolean"/>
8.                 </xsd:element>
9.                 <xsd:element name="Морда">
10.                        <xsd:element name="Шерсть" type="xsd:string"/>
11.                        <xsd:element name="Усы" type="xsd:positiveInteger"/>
12.                        <xsd:element name="Глаза" type="xsd:string"/>
13.                </xsd:element>
14.                <xsd:element name="Хвост">
15.                        <xsd:element name="Шерсть" type="xsd:string"/>
16.                        <xsd:element name="Длина" type="xsd:positiveInteger"/>
17.                </xsd:element>
18.                <xsd:element name="Кисточки на ушках" type="xsd:boolean"/>
19.                </xsd:sequence>
20.        </xsd:complexType>
21.</xsd:element>
```

```
1. <xsd:element name="Кошка">
2.         <xsd:complexType>
3.                 <xsd:sequence>
4.                 <xsd:element name="Имя" type="xsd:string"/>
5.                 <xsd:element name="Шерсть">
6.                         <xsd:element name="Лапы" type="xsd:string"/>
7.                         <xsd:element name="Морда" type="xsd:string"/>
8.                         <xsd:element name="Хвост" type="xsd:string"/>
9.                 </xsd:element>
10.                <xsd:element name="Усы" type="xsd:positiveInteger"/>
11.                <xsd:element name="Глаза" type="xsd:string"/>
12.                <xsd:element name="Когти" type="xsd:boolean"/>
13.                <xsd:element name="Длина хвоста" type="xsd:positiveInteger"/>
14.                <xsd:element name="Кисточки на ушках" type="xsd:boolean"/>
15.                </xsd:sequence>
16.        </xsd:complexType>
17.</xsd:element>
```

Мапирование (Википедия)

Терминология: интероперабельность

ГОСТ Р 55062-2012 Информационные технологии (ИТ). Системы промышленной автоматизации и их интеграция. Интероперабельность. Основные положения

интероперабельность (interoperability) - **способность** двух или более информационных систем или компонентов к **обмену** информацией и к **использованию** информации, полученной в результате обмена

семантическая интероперабельность (semantic interoperability) - способность любых взаимодействующих в процессе коммуникации информационных систем **одинаковым** образом понимать **смысл информации**, которой они обмениваются.

барьер интероперабельности (interoperability barrier) - **несовместимость сущностей**, которая препятствует обмену информацией с другими сущностями, использованию сервисов или общему пониманию обмененных элементов

глоссарий интероперабельности (glossary) - термины и определения, используемые в области интероперабельности с толкованием, иногда переводом на другой язык, комментариями и примерами.

ГОСТ Р 55062-2012

Терминология: справочники и классификаторы

Спра́вочник — издание практического назначения, с кратким изложением сведений в систематической форме, в расчёте на выборочное чтение, на то, чтобы можно было быстро и легко навести по нему справку

Классифика́тор, или (от лат. *classis* — *разряд* и *facere* — *делать*) — систематизированный перечень наименованных объектов, каждому из которых в соответствие дан уникальный код

Классификация объектов производится согласно правилам распределения заданного множества объектов на подмножества (**классификационные группировки**) в соответствии с установленными **признаками** их различия или сходства

Единая система классификации и кодирования информации (ЕСКК) -совокупность общероссийских классификаторов технико-экономической и социальной информации; средств ведения классификаторов; нормативных и методических документов по их разработке, ведению и применению

Классификатор - нормативный документ, содержащий систематизированный свод наименований и кодов классификационных группировок и (или) объектов классификации

Справочник (Википедия) *Классификатор (Википедия)*

Терминология: справочники и классификаторы

ОБЩЕРОССИЙСКИЙ КЛАССИФИКАТОР ИНФОРМАЦИИ ОБ ОБЩЕРОССИЙСКИХ КЛАССИФИКАТОРАХ

ОК 026-2002

- Общероссийский классификатор информации об общероссийских классификаторах. ОК 026-2002
  - Предисловие
  - Введение
  - 1. Общероссийские классификаторы
  - 2. Фасеты общероссийских классификаторов
    - 006 Общероссийский классификатор информации по социальной защите населения (ОКИСЗН)
      - Раздел I. Пенсионное обеспечение
      - Раздел II. Социальная защита граждан, подвергшихся воздействию радиации вследствие чернобыльской катастрофы и других радиационных и техногенных катастроф
      - Раздел III. Пособия, компенсации, льготы и другие меры социальной поддержки и социальной помощи
      - Раздел IV. Социальное обслуживание
      - Раздел V. Медико-социальная экспертиза, реабилитация и абилитация инвалидов
      - Раздел VI. Меры социальной поддержки граждан в области занятости населения
      - Раздел VII. Обеспечение по обязательному социальному страхованию
    - 016 Общероссийский классификатор профессий рабочих, должностей служащих и тарифных разрядов (ОКПДТР)
    - 018 Общероссийский классификатор информации о населении (ОКИН)
    - 031 Общероссийский классификатор видов грузов, упаковки и упаковочных материалов (ОКВГУМ)
    - 036 Общероссийский классификатор трансформационных событий (ОКТС)
- Приложение А. Международные (региональные) классификации и стандарты, используемые в общероссийских классификаторах
- Приложение Б. Межгосударственные классификаторы, с которыми гармонизированы общероссийские классификаторы
- Приложение В. Объекты классификации и структура кодов общероссийских классификаторов

ОК 026-2002

Классификация

Классификация - разделение множества объектов на подмножества по их сходству или различию в соответствии с принятыми методами

Задача классифика́ции — формализованная задача, в которой имеется множество объектов (ситуаций), разделённых некоторым образом на классы

Задано конечное множество объектов, для которых известно, к каким классам они относятся, это множество называется выборкой

Классовая принадлежность остальных объектов неизвестна

Требуется построить алгоритм, способный классифицировать (см. ниже) произвольный объект из исходного множества.

Классифици́ровать объект — значит, указать номер (или наименование) класса, к которому относится данный объект.

Классификация
Классификация (задача)

Классификация: методы

В зависимости от специфики представления информации, целей и способов взаимодействия с ней, используются два основных метода классификации объектов: **иерархический** и **фасетный**

В самом широком смысле **иерархический** метод классификации является частным случаем фасетного метода и отличается от него только возможностью задания строгой зависимости классификационных признаков более низкого уровня от классификационных признаков более высокого уровня

Фасетный метод классификации применяется для классификации сложно организованных множеств объектов, а также для классификации множеств объектов с неопределенными, не поддающимися формализации или динамичными границами

Основными целями проведения фасетной классификации являются:
- получение и кодирование заданного набора показателей (преимущественно оценочного типа),
- отнесение каждого объекта классификации к одному из заданного набора показателей.

Классификация: иерархический метод

При проведении классификации **иерархическим методом** все исходное множество объектов классификации последовательно разбивается на соподчиненные (вложенные друг в друга) подмножества:

1. все исходное множество делится на классы в зависимости от выбранного для него классификационного признака

2. полученные классы образуют классификационные группировки первого уровня классификации

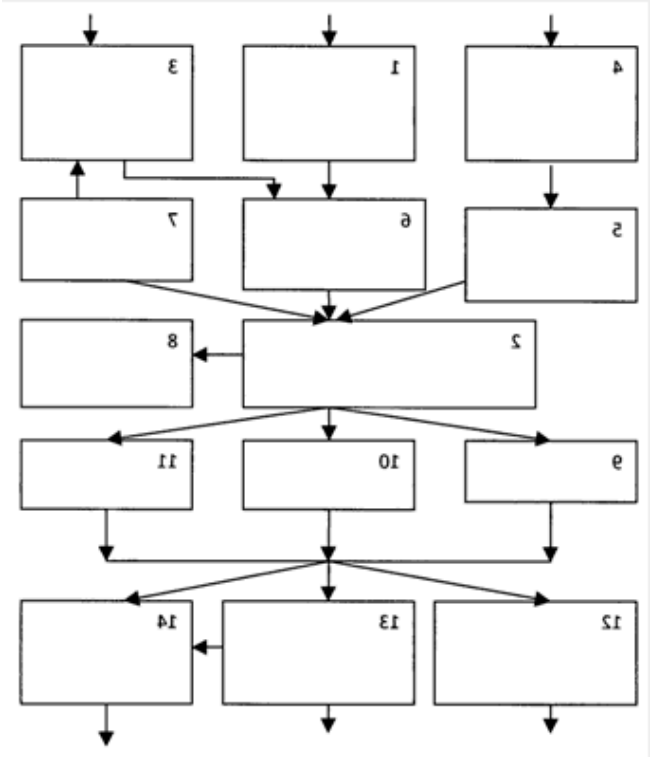
3. каждый класс очередного уровня делится на подклассы, которые образуют классификационные группировки следующего уровня классификации.

Разбиение каждого отдельно взятого класса на подклассы выполняется либо в соответствии со своим оригинальным значением классификационного признака более высокого уровня, либо в соответствии со своим оригинальным классификационным признаком, что позволяет использовать независимые классификационные признаки для разных ветвей иерархической структуры.

Классификация: фасетный метод

При проведении классификации **фасетным методом**:

1. определяются классификационные признаки объектов классификации (фасеты)
2. для каждого фасета устанавливается набор конкретных значений показателей (терминов)
3. заданное множество объектов классификации рассматривается в произвольном порядке, при этом:
   - для каждого из заданного множества объектов классификации выявляются необходимые классификационные признаки
   - для каждого заданного объекта принимается решение о его отнесении к определенному значению показателей (термину) фасета либо о его исключении из классификационной группировки



Патент RU 108179

Классификация: фасетный метод

Принятие решения осуществляется следующим образом:

1. при наличии у заданного объекта классификации необходимых классификационных признаков и соответствии их значений значениям, определенным для одного из конкретных значений (терминов) фасета - объект относится к указанному термину фасета
2. при отсутствии у заданного объекта классификации необходимых классификационных признаков либо набора их значений, позволяющих однозначно отнести его к какому-либо конкретному значению (термину) фасета - объект исключается из классификационной группировки.

Главное требование при заполнении фасета - исключение возможности повторения одних и тех же значений классификационных признаков в различных фасетах.

Фасетная система классификации позволяет при группировке объектов выбирать классификационные признаки независимо друг от друга, что придает ей большую гибкость относительно других методов классификации и практически неограниченного добавления числа фасетов, в частности:

- расширения состава значений (терминов) в отдельных фасетах,
- группировки заданного множества по любому сочетанию и числу фасетов.

Терминология: системы кодирования

Система кодирования - строго определенный порядок присвоения условных обозначений единицам информации. Таким образом, все коды строятся по определенным правилам (системам).Используемые для этих целей системы построения кодов подразделяются на:

- линейные (одномерные)
- шахматные.

Линейные системы кодирования

В линейных кодах условное обозначение соответствует только одной единице информации. По способу построения различают следующие линейные системы кодирования:

- порядковые,
- серийные,
- позиционные (разрядные, или десятичные),
- повторения,
- смешанные (комбинированные).

Шахматная система кодирования

В шахматных используются двухпозиционные коды, одновременно отражается характеристика двух информационных единиц (по строке и столбцу)

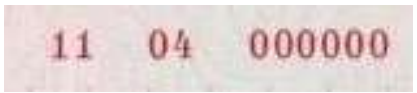
Терминология: системы кодирования

Порядковая система кодирования предполагает последовательное присвоение условных обозначений кодируемым единицам информации.

Специальной классификации информации, как правило, не требуется.
Последовательность кодов задается, прежде всего, хронологией возникновения информационных единиц, но чаще всего объектом кодирования выступает информация, упорядоченная (систематизированная) по алфавиту.

Серийная система кодирования ориентирована на разделение классифицируемого множества по какому-либо признаку на отдельные части (серии). За каждой серией закрепляется своя группа условных обозначений (чисел, называемых номерами).

При этом номера единиц информации последующих серий не продолжают последовательно номера уже имеющихся единиц предыдущей серии, в результате создается определенный разрыв номеров, используемый в качестве резерва для последующего расширения (в случае необходимости) множества кодируемых позиций в каждой серии без нарушения общей логики построения списка.

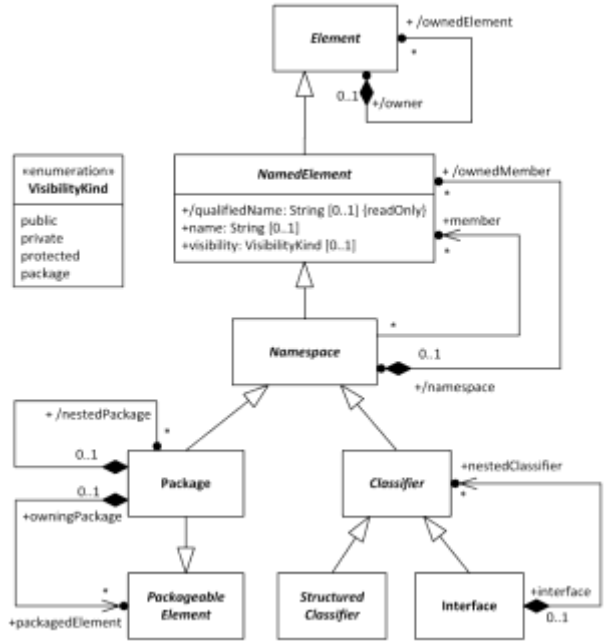


Терминология: системы кодирования

Позиционная (разрядная, десятичная) система кодирования предполагает иерархическую структуру представления информации или разделение ее по нескольким соподчиненным признакам.

Сущность данной системы заключается в том, что каждый уровень (или признак) классификации обеспечивается своей нумерацией в пределах всего уровня или признака (группы информации). При этом устанавливается предел разрядности группы и выбирается ее некоторая кратность.

Позиционная система применяется для кодирования сложных составных (иерархических) номенклатур, в которых, как правило, каждый реквизит, характеризующий низший уровень классификации, получает ряд характеристик, отражающих его принадлежность к более высокому уровню классификации.



[Namespace]

Терминология: переходные ключи

Гармонизация общероссийского классификатора: Приведение общероссийского классификатора в соответствие с международной (региональной) классификацией, межгосударственным классификатором или международным (региональным) стандартом по классификации установленными путями гармонизации.

Переходной ключ: Таблица, устанавливающая соответствие каждой группировки или объекта классификации общероссийского классификатора одной или нескольким группировкам или объектам сопоставляемой классификации.

Код		Дробная единица[с 1]	Наименования валют[с 2]		Наименования государств и территорий[с 3]
буквенный	цифровой	разряды	ОКВ	ISO 4217	в соответствии с ISO 4217 и/или ОКВ
Денежные единицы, включённые в действующую редакцию ОКВ					
AED	784	2	Дирхам (ОАЭ)	UAE Dirham	ОАЭ
AFN	971	2	Афгани	Afghani	Афганистан
ALL	008	2	Лек	Lek	Албания
AMD[а 1]	051	2	Армянский драм	Armenian Dram	Армения
ANG	532	2	Нидерландский антильский гульден	Netherlands Antillean Guilder	Кюрасао Синт-Мартен
AOA	973	2	Кванза	Kwanza	Ангола
ARS	032	2	Аргентинское песо	Argentine Peso	Аргентина

ПР 50.1.024-2005
ОКВ

НСИ, MDM

ГОСТ 34.003-90 Информационные технологии. Комплекс стандартов на автоматизированные системы. Автоматизированные системы. Термины и определения

Нормативно-справочная информация (НСИ) автоматизированной системы - Информация, заимствованная из нормативных документов и справочников и используемая при функционировании

Управление основными данными, **управление мастер-данными** (англ. *Master Data Management*, MDM) — совокупность процессов и инструментов для постоянного определения и управления основными данными компании (в том числе справочными).

Мастер-данные — это данные с важнейшей для ведения бизнеса информацией о:

- клиентах
- продуктах
- услугах
- персонале
- технологиях
- материалах и так далее.

Они относительно редко изменяются и не являются транзакционными.

MDM

НСИ, MDM

РД 50-34.698-90 Автоматизированные системы. Требования к содержанию документов.

5.3. Описание информационного обеспечения системы

Документ содержит разделы:

1. состав информационного обеспечения;
2. организация информационного обеспечения;
3. организация сбора и передачи информации;
4. построение системы классификации и кодирования;
5. организация внутримашинной информационной базы;
6. организация внемашинной информационной базы.

5.6. Описание систем классификации и кодирования

Документ содержит:

- перечень применяемых в АС зарегистрированных классификаторов всех категорий по каждому классифицируемому объекту
- описание метода кодирования, структуры и длины кода
- указания о системе классификации и другие сведения по усмотрению разработчика.

РД 50-34.698-90

НСИ, MDM

РД 50-34.698-90 Автоматизированные системы. Требования к содержанию документов.

5.3.3. В разделе "Организация информационного обеспечения" приводят:

1. принципы организации информационного обеспечения системы
2. обоснование выбора носителей данных и принципы распределения информации по типам носителей
3. описание принятых видов и методов контроля в маршрутах обработки данных при создании и функционировании внемашинной и внутримашнной информационных баз с указанием требований, на соответствие которым проводят контроль
4. описание решений, обеспечивающих информационную совместимость АС с другими системами управления по источникам, потребителям информации, по сопряжению применяемых классификаторов (при необходимости), по использованию в АС унифицированных систем документации

РД 50-34.698-90

НСИ, MDM

РД 50-34.698-90 Автоматизированные системы. Требования к содержанию документов.

5.3.4. В разделе "Организация сбора и передачи информации" приводят:

1. перечень источников и носителей информации с указанием оценки интенсивности и объема потоков информации;
2. описание общих требований к организации сбора, передачи, контроля и корректировки информации.

5.3.5. В разделе "Построение системы классификации и кодирования" приводят:

1. описание принятых для применения в АС классификации объектов во вновь разработанных классификаторах и в тех действующих классификаторах, из которых используется часть кода;
2. методы кодирования объектов классификации во вновь разработанных классификаторах.

РД 50-34.698-90

НСИ, MDM

http://solutions.1c.ru/catalog/mdm/features

PDM

PDM-система (англ. *Product Data Management* — система управления данными об изделии) — организационно-техническая система, обеспечивающая управление всей информацией об изделии.

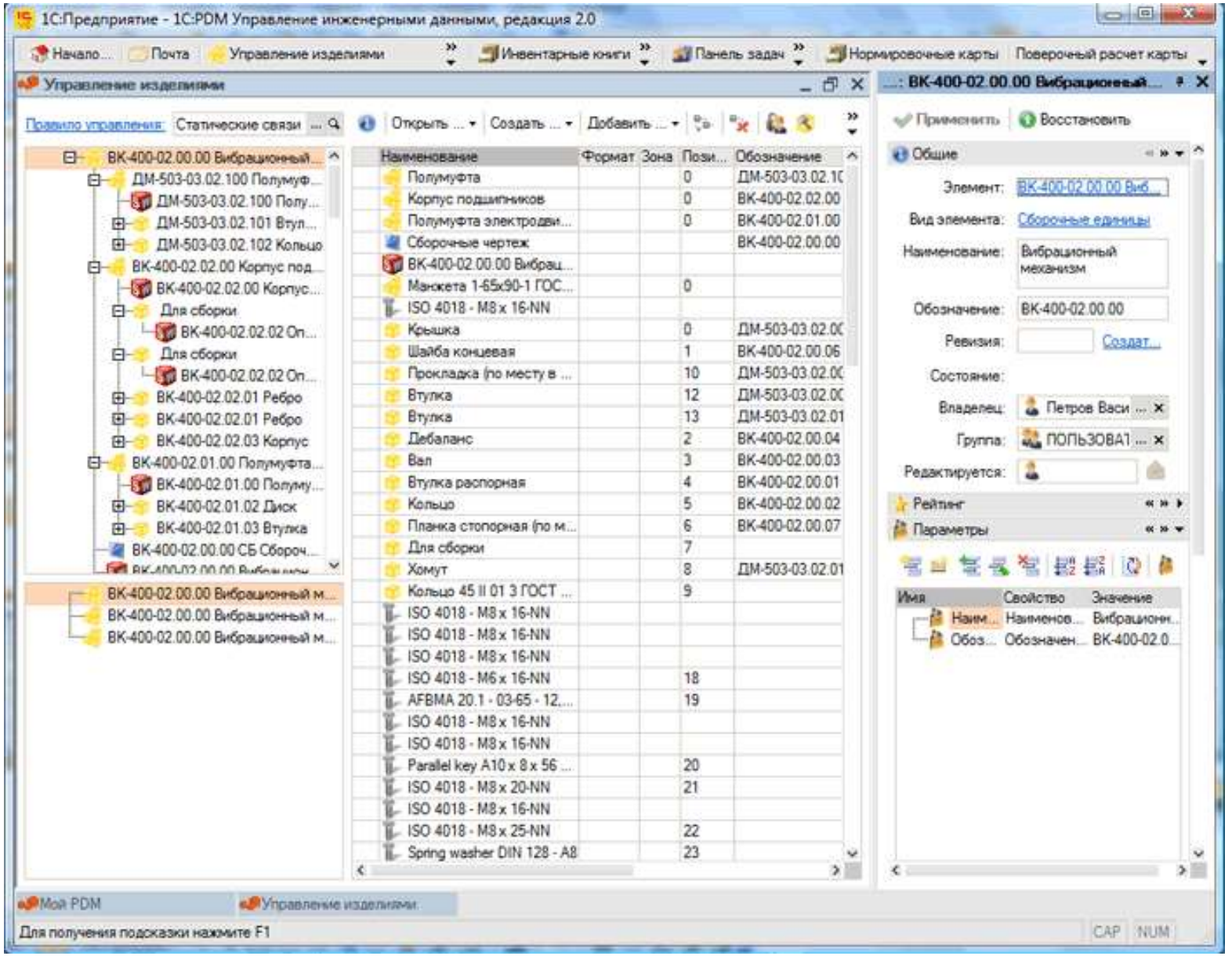
При этом в качестве изделий могут рассматриваться различные сложные технические объекты (корабли и автомобили, самолёты и ракеты, компьютерные сети и др.). PDM-системы являются неотъемлемой частью *PLM*-систем.

В PDM-системах обобщены такие технологии, как:

- управление инженерными данными (engineering data management — **EDM**)
- управление документами
- управление информацией об изделии (product information management — **PIM**)
- управление техническими данными (technical data management — **TDM**)
- управление технической информацией (technical information management — **TIM**)
- управление изображениями и манипулирование информацией, всесторонне определяющей конкретное изделие.

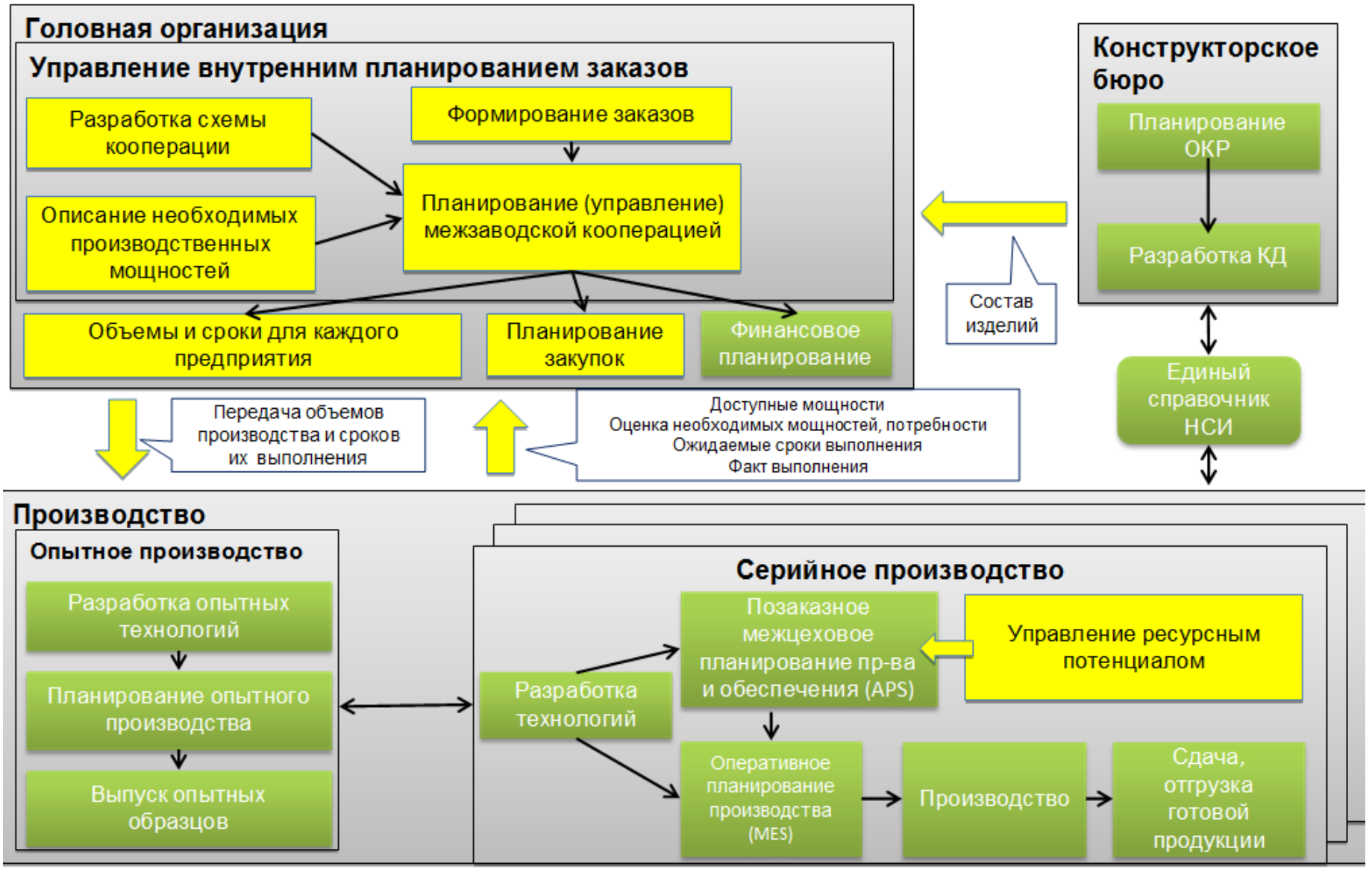
PDM
https://www.youtube.com/watch?v=68_VxQYyK5g

PDM

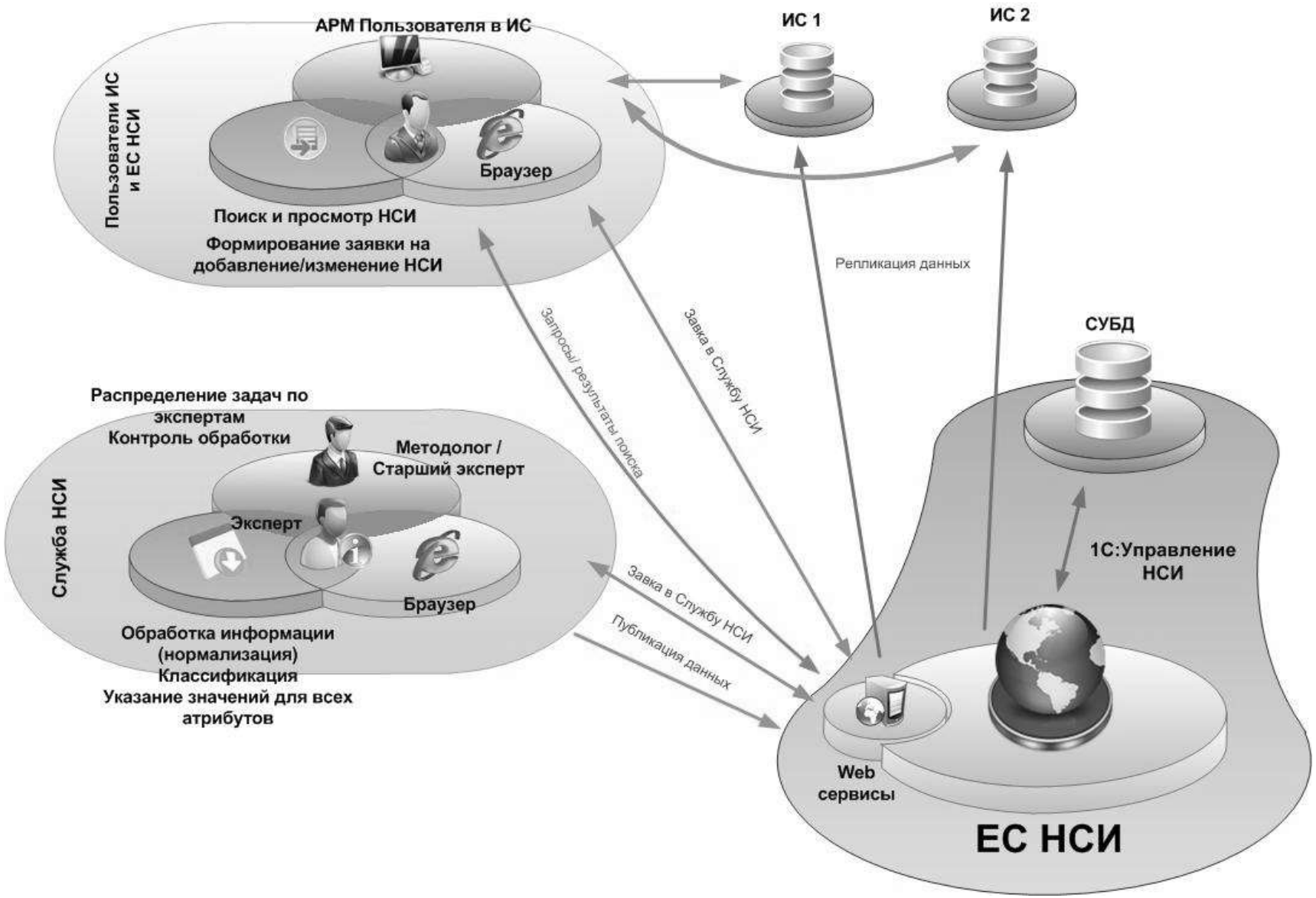


http://solutions.1c.ru/catalog/pdm/features

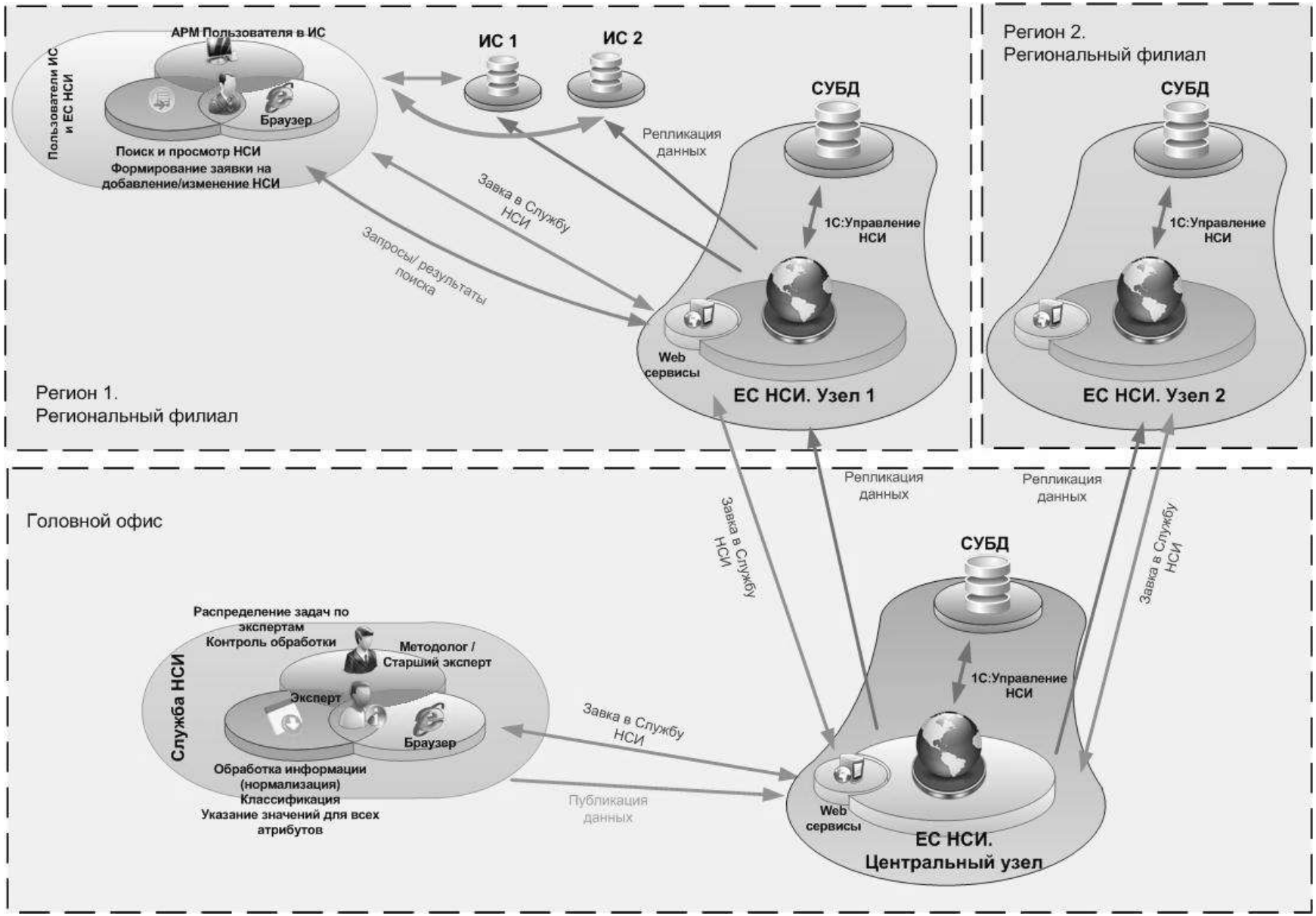
Типовые схемы интеграции



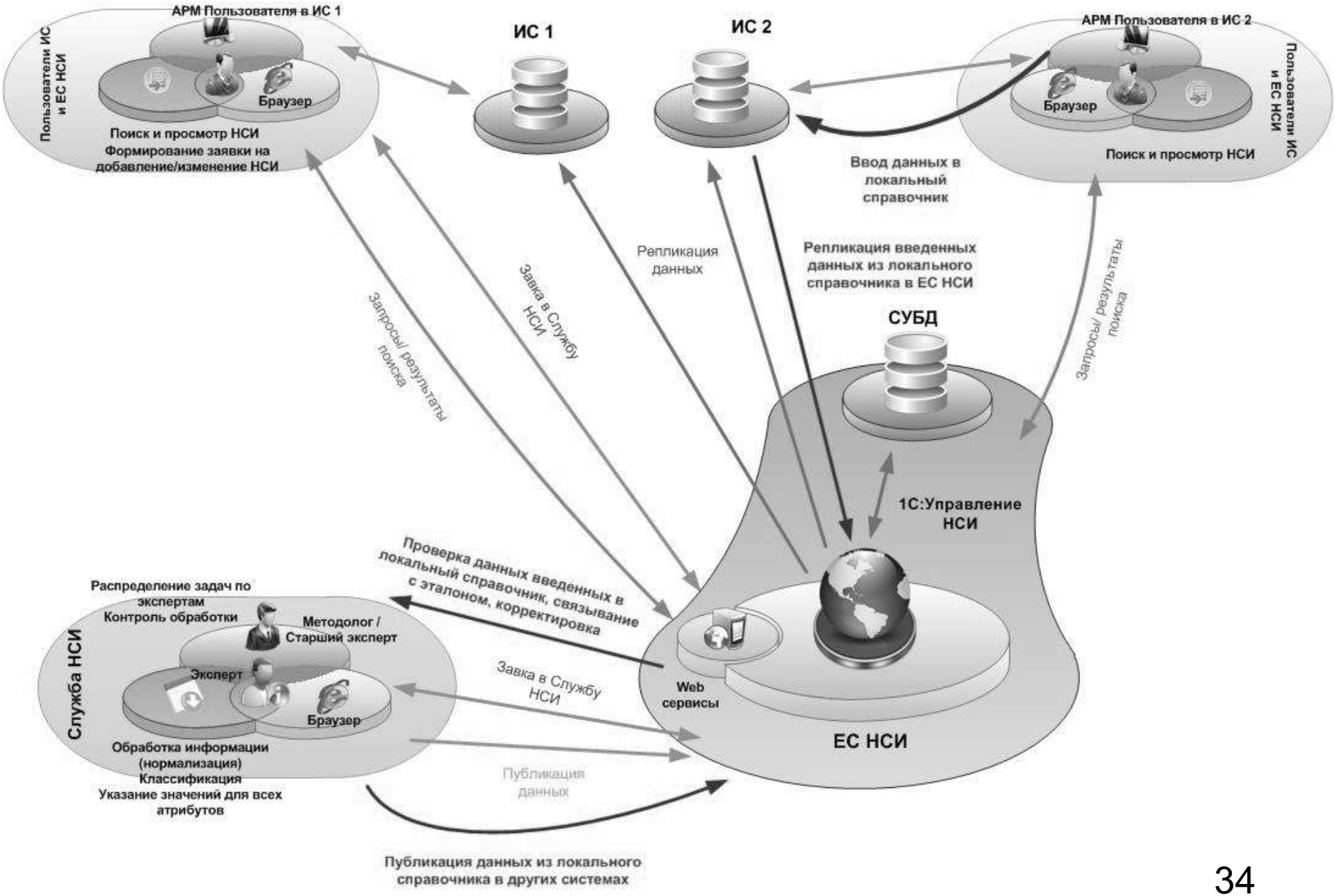
Типовые схемы интеграции



Типовые схемы интеграции



Типовые схемы интеграции



Лекция 16
«**Извлечение и очистка данных**»

Овчинников П.Е.
МГТУ «СТАНКИН»,
ст.преподаватель кафедры ИС

27.03.2019

Терминология: качество данных

ГОСТ Р ИСО 8000-2-2014 Качество данных. Часть 2. Словарь

метаданные (metadata): Данные, которые описывают и определяют другие данные

данные (data): Символическое представление чего-либо, частично зависящего в своем значении от метаданных

точность данных (data accuracy): Точность соответствия между значением свойства и истинным значением

истинное значение (true value): Значение параметров характеристики какого-либо объекта в определенных условиях

авторитетный источник данных (authoritative data source): Владелец процесса, производящего данные

словарь данных (data dictionary): Совокупность вводимых в словарь данных, которые можно найти по идентификатору объекта

основные данные (master data): Данные, находящиеся во владении организацией и описывающие основные объекты этой организации. На эти данные следует ссылаться при составлении транзакций

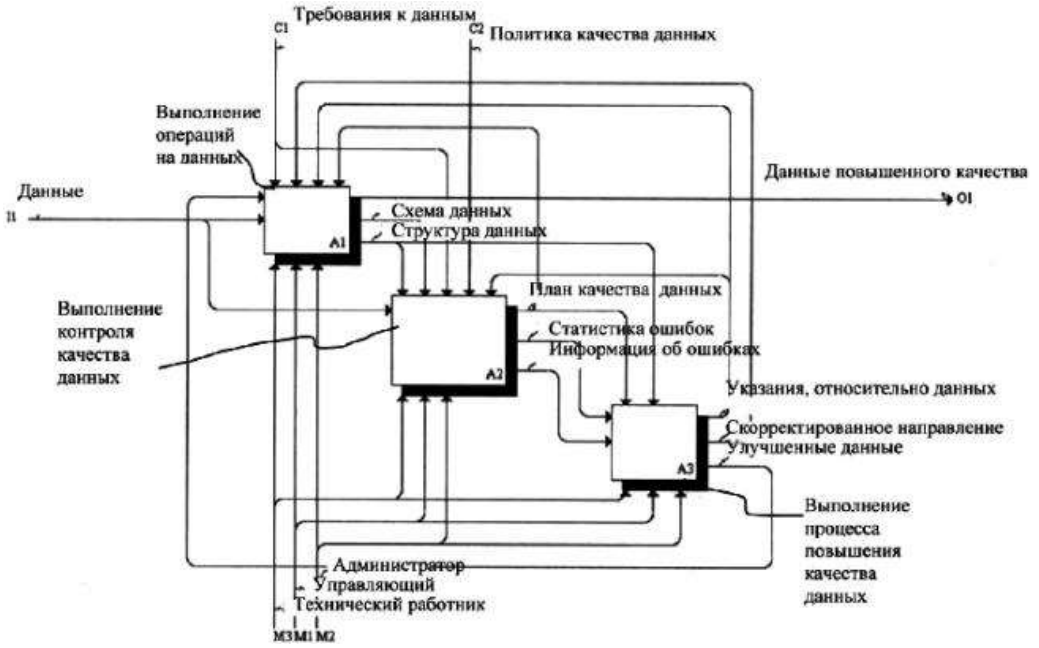
ГОСТ Р ИСО 8000-2-2014

Терминология: нормализация НСИ

Нормализация нормативно-справочной информации представляет собой приведение к стандартному виду всех данных, содержащихся в справочниках

В процессе нормализации первоначальная информация в справочниках **разбирается** и **структурируется** в соответствии с созданными правилами, одновременно выполняется **множественная классификация** элементов справочников

ГОСТ Р 56215-2014/ISO/TS 8000-150:2011 Качество данных. Часть 150. Основные данные. Структура управления качеством



DataReon ГОСТ Р 56215-2014

Пример: нормализация адресов

Для использования «КЛАДР» - Классификатор адресов Российской Федерации на сайте, мы получаем актуальные данные Государственного реестра адресов ФНС России.

Актуальность базы: **2019.03.21**

Адрес: Москва Город -> Вадковский Переулок

Код КЛАДР: ***77000000000053400***

Код региона	Почтовый индекс	Код окато	Код налоговой
77	127055	45286585000	7707

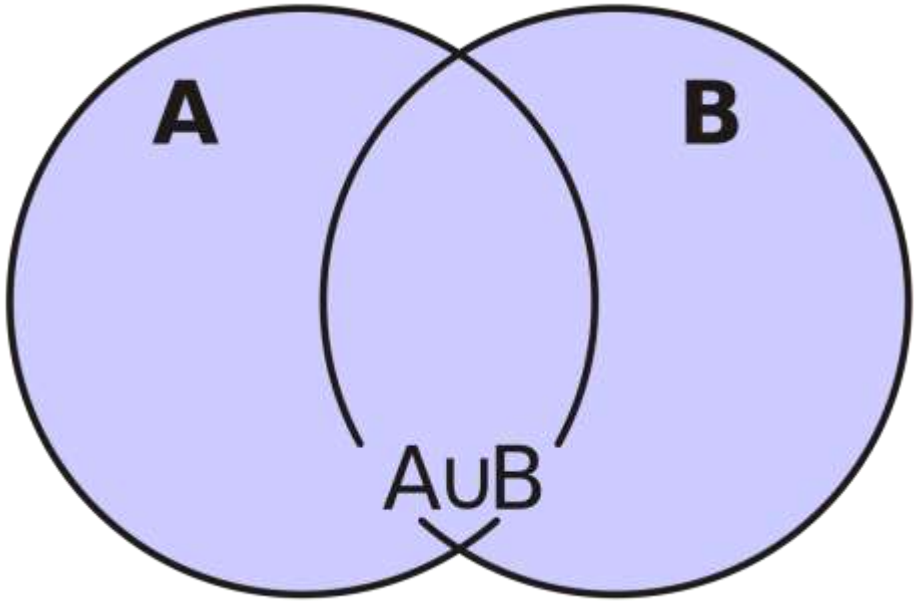
Интервал домов	Почтовый индекс	Код окато	Код налоговой
1,10стр1,10стр13,10стр2,12,16,18стр1	127055	45286585000	7707
18стр10,18стр1А,18стр4,18стр5,18стр7	127055	45286585000	7707
18стр8,18стр9,20стр1,20стр2,24/35стр1,3	127055	45286585000	7707

КЛАДР

Терминология: дедупликация

Объедине́ние мно́жеств (тж. **су́мма** или **соедине́ние**) в теории множеств — множество, содержащее в себе все элементы исходных множеств

Объединение двух множеств {A} и {B} обычно обозначается {A} ∪ {B}, но иногда можно встретить запись в виде суммы {A+B}



Объединение множеств – потенциальный источник дубликатов!

Сложение (объединение) множеств

Терминология: дедупликация

В языке SQL операция **UNION** применяется для объединения двух наборов строк, возвращаемых SQL-запросами

Оба запроса должны возвращать одинаковое число столбцов, и столбцы с одинаковым порядковым номером должны иметь совместимые типы данных

Результат получает структуру (названия и типы столбцов) первого (левого) запроса, то есть операция не является симметричной

```
(SELECT * FROM sales2005)
UNION
(SELECT * FROM sales2006);
```

Без дупликатов

```
(SELECT * FROM sales2005)
UNION ALL
(SELECT * FROM sales2006);
```

С дупликатами

Оператор UNION (Википедия)

Терминология: очистка данных

Очистка данных (англ. *Data cleansing*) — процесс выявления и исправления ошибок, несоответствий данных с целью улучшения их качества, иногда классифицируется как составная часть интеллектуального анализа данных

Очистка данных выполняется с определенными наборами данных в базах данных или файлах

Необходимость в очистке данных чаще всего возникает при интеграции различных информационных систем (хранилища данных, системы управления ресурсами предприятия, системы управления взаимодействием с клиентами)

Источники данных в различных системах часто находятся в разрозненном виде и в различных состояниях. Преобразования выполняются автоматически (в соответствии с набором правил) либо вручную (в интерактивном режиме).

Наиболее типичные предметные области, подлежащие очистке и исправлению в корпоративных информационных системах — сведения о лицах и организациях, адресная и контактная информация, также подлежит очистке любая справочная информация, вносимая вручную в текстовом виде.

Очистка данных (Википедия)

Терминология: обогащение данных

Процесс насыщения данных новой информацией, которая позволяет сделать их более ценными и значимыми с точки зрения решения той или иной аналитической задачи

Существует два основных метода обогащения данных - **внешнее** и **внутреннее**

Внешнее обогащение предполагает привлечение дополнительной информации из источников, которые находятся вне информационной системы предприятия

Практически источником информации для обогащения данных могут быть любые организации, которые в процессе своей деятельности собирают, структурируют и хранят сведения, связанные с их деятельностью

Внутреннее обогащение не предполагает привлечения какой-либо внешней информации. Оно обычно связано с получением и включением в набор данных полезной информации, которая отсутствует в явном виде, но может быть тем или иным способом получена с помощью манипуляций с имеющимися данными.

Затем, эта информация встраивается в виде новых полей или даже таблиц в хранилище данных и может быть использована для дальнейшего анализа

Обогащение данных (Мнение)

НСИ, MDM

http://solutions.1c.ru/catalog/mdm/features

НСИ, MDM

http://solutions.1c.ru/catalog/mdm/features

Терминология:ETL

ETL (от англ. *Extract, Transform, Load* — дословно «извлечение, преобразование, загрузка») — один из основных процессов в управлении хранилищами данных, который включает в себя:

- извлечение данных из внешних источников;
- трансформацию и очистку данных
- загрузку в хранилище данных

С точки зрения процесса ETL, архитектуру хранилища данных можно представить в виде трёх компонентов:

- источник данных: содержит структурированные данные в виде таблиц, совокупности таблиц или просто файла (данные в котором разделены символами-разделителями);
- промежуточная область: содержит вспомогательные таблицы, создаваемые временно, и, исключительно для организации процесса выгрузки.
- получатель данных: хранилище данных или база данных, в которую должны быть помещены извлечённые данные.

Перемещение данных от источника к получателю называют потоком данных

ETL (Википедия)

Терминология:ETL

Извлечение данных в ETL

Начальным этапом процесса ETL является процедура извлечения записи из источников данных и подготовка их к процессу преобразования

При разработке процедуры извлечения данных, в первую очередь необходимо определить частоту выгрузки данных из OLTP-систем или отдельных источников

Выгрузка данных занимает определённое время, которое называется окном выгрузки.

Процедуру извлечения данных можно реализовать двумя способами:

- извлечение данных с помощью специализированных программных средств;
- извлечение данных средствами той системы, в которой они хранятся.

После извлечения данные помещаются в так называемую «промежуточную область», где для каждого источника данных создаётся своя таблица или отдельный файл, или и то и другое

ETL (Википедия)

Терминология:ETL

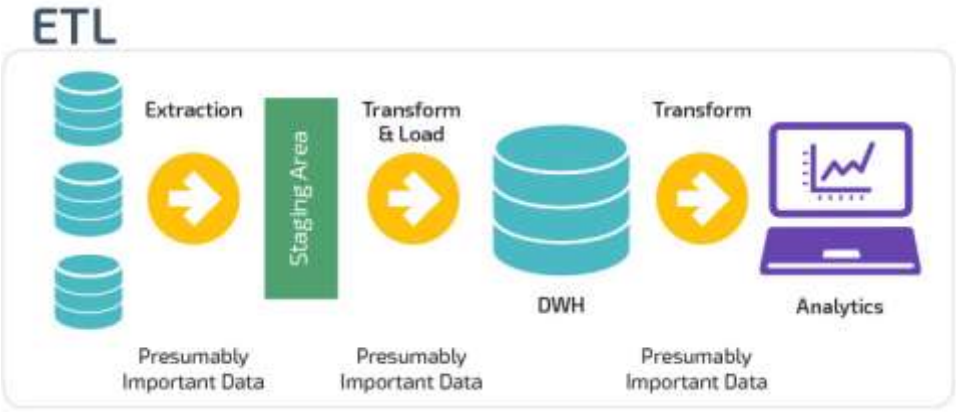
Преобразование данных

Цель этого этапа — подготовка данных к размещению в хранилище данных и приведение их к виду более удобному для последующего анализа

При этом должны учитываться некоторые, выдвигаемые аналитиком, требования, в частности, к уровню качества данных. Поэтому в процессе преобразования может быть задействован самый разнообразный инструментарий, начиная с простейших средств ручного редактирования данных и заканчивая системами, реализующими сложные методы обработки и очистки данных

В процессе преобразования данных в рамках ETL чаще всего выполняются следующие операции:

- преобразование структуры данных
- агрегирование данных
- перевод значений
- создание новых данных
- очистка данных



ETL (Википедия)

Терминология:ETL

Загрузка данных

Процесс загрузки заключается в переносе данных из промежуточных таблиц в структуру хранилища данных. При очередной загрузке в хранилище данных переносится не вся информация из источников, а только та, которая была изменена в течение промежуточного времени, прошедшего с предыдущей загрузки

При этом выделяют два потока:

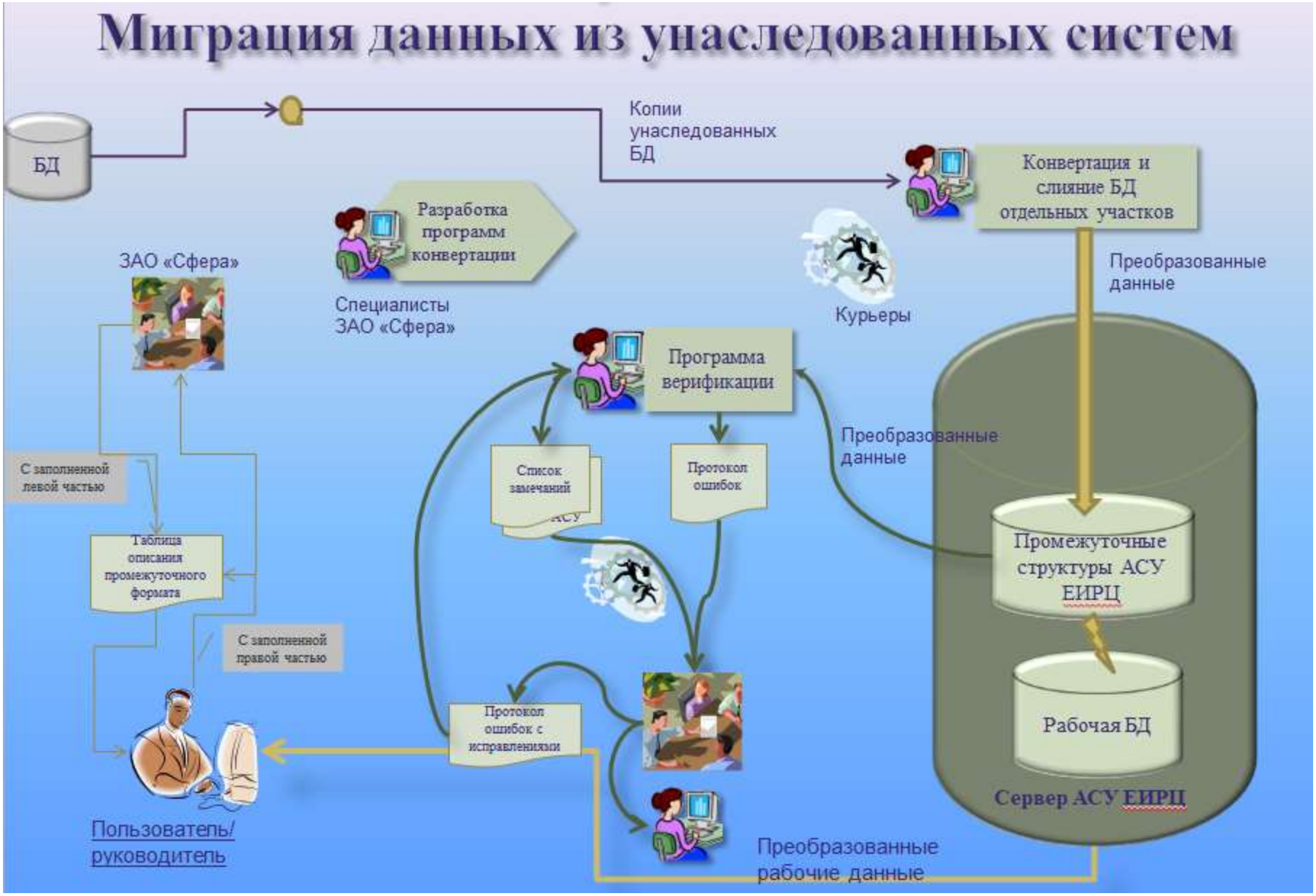
- поток добавления — в хранилище данных передается новая, ранее не существовавшая информация;
- поток обновления (дополнения) — в хранилище данных передается информация, которая существовала ранее, но была изменена или дополнена.

Для распределения загружаемых данных на потоке используются средства данных. Они фиксируют состояние данных в некоторые моменты времени и определяют, какие данные были изменены или дополнены

ETL (Википедия)

Терминология: миграция данных

Терминология: миграция данных



Терминология: большие данные

Больши́е да́нные (англ. *big data*, [ˈbɪg ˈdeɪtə]) — обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами и альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence

В качестве определяющих характеристик для больших данных традиционно выделяют «**три V**»:

- **объём** (англ. *volume*, в смысле величины физического объёма),
- **скорость** (*velocity* в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов),
- **многообразие** (*variety*, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных)

в дальнейшем возникли различные вариации и интерпретации этого признака

Большие данные (Википедия)

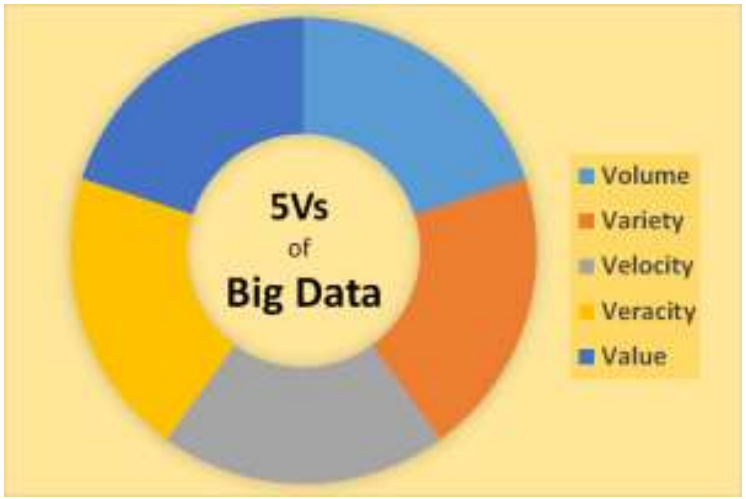
Терминология: большие данные

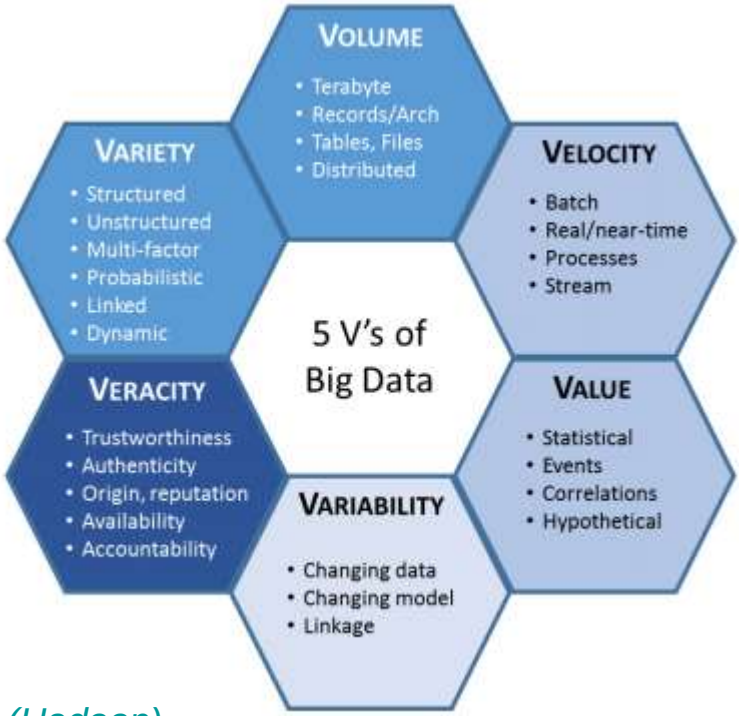
4. Правдивость (Veracity)

Точки данных, которые были собраны и сохранены из различных источников в различных формах, часто имеют дело с неточностью. При этом нам приходится иметь дело с низким качеством данных, в том числе в огромных объемах (например, в сообщениях Twitter с хеш-тегами, опечатками, сокращениями и разговорной речью), которые не являются точными и неопределенными

4V (IBM)

Терминология: большие данные





5. Ценность (Value)
Независимо от того, являются ли данные большими или небольшими, независимо от того, были ли они получены в любом месте и в каком бы то ни было формате, они должны иметь определенную ценность - это означает, что мы можем правильно использовать данные по их правильной причине для их достоверности

Значение, ценность или функциональность данных для тех, кто их потребляет, по-видимому, наиболее важны для различных фирм или организаций. Кроме того, мы знаем, что данные сами по себе не имеют никакого значения или полезности, но все же нам нужны ценные данные для получения информации.

5V (Hadoop)

Пример: технология MapReduce

MapReduce — это фреймворк для вычисления некоторых наборов распределенных задач с использованием большого количества компьютеров (называемых «нодами»), образующих кластер

Работа MapReduce состоит из двух шагов: Map и Reduce, названных так по аналогии с одноименными функциями высшего порядка, map и reduce

На Map-шаге происходит предварительная обработка входных данных. Для этого один из компьютеров (называемый главным узлом — master node):

- **получает** входные данные задачи
- **разделяет** их на части и
- **передает** другим компьютерам (рабочим узлам — worker node) для предварительной обработки

На Reduce-шаге происходит свёртка предварительно обработанных данных.

Главный узел получает ответы от рабочих узлов и на их основе формирует результат — решение задачи, которая изначально формулировалась.

MapReduce (Википедия)

Пример: технологии кластерного анализа

Кластерный анализ (англ. *cluster analysis*) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Кластерный анализ решает следующие основные задачи:

- разработка **типологии** или классификации
- исследование полезных концептуальных **схем** группирования объектов
- порождение **гипотез** на основе исследования данных
- проверка гипотез или **исследования** для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных

Применение кластерного анализа предполагает следующие этапы:

- отбор **выборки** для кластеризации
- определение множества переменных, по которым будут оцениваться объекты в выборке, то есть **признакового пространства**
- вычисление значений той или иной **меры сходства** (или различия) между объектами
- применение метода кластерного анализа для **создания групп** сходных объектов
- проверка **достоверности** результатов кластерного решения

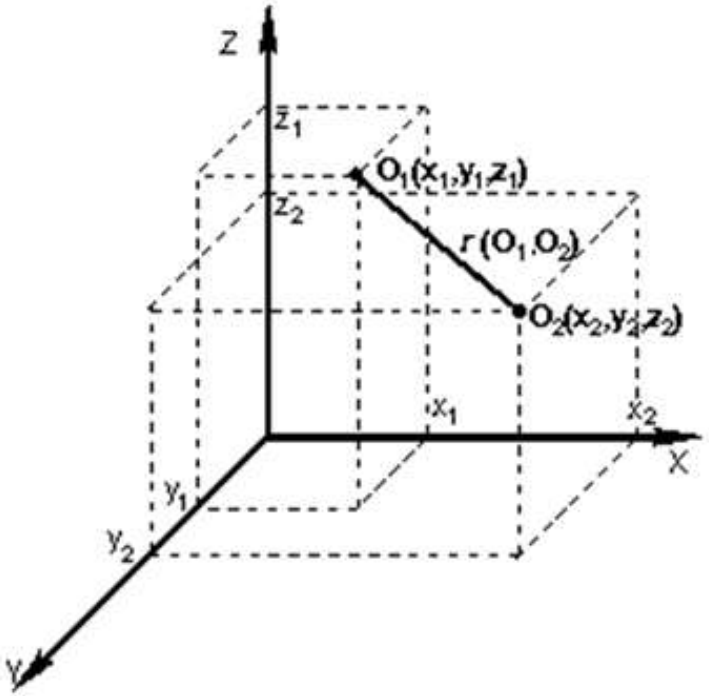
Кластерный анализ (Википедия)

Пример: Эвклидово пространство

Евкли́дово простра́нство (также **эвкли́дово простра́нство**) — в изначальном смысле, пространство, свойства которого описываются аксиомами евклидовой геометрии

Евклидово расстояние между точками находится по формуле

$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, где n – количество измерений



Расстояние между двумя точками в пространстве трех измерений

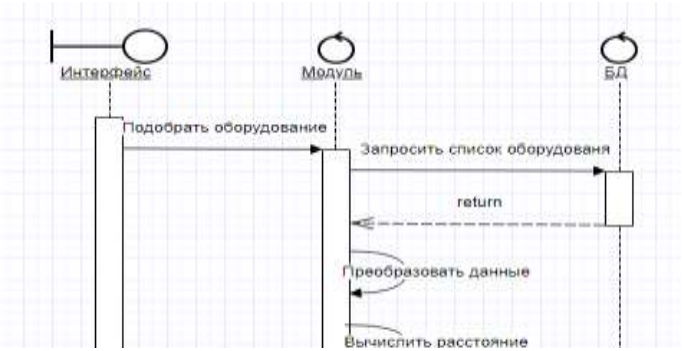
Эвклидово пространство (Википедия)

Пример: расчет расстояний

1. Получаются значения эталона, заданные пользователем, по которым будет выполняться сравнение

2. Из справочника получается список элементов с указанием названия, параметров и их значений

3. Так как параметры могут иметь разные размерности, то для каждого элемента они приводятся к единому виду, где единица – величина, равная значениям эталона.



V1 (эталон)	V2	Расстояние
Число	Пусто	1
Число	Число	$\sqrt[2]{(V1-V2)^2}$
Число	Строка	1
Число	Диапазон	Берется среднее значение диапазона и считается, как «число-число»
Диапазон	Число	
Диапазон	Диапазон	
Строка	Пусто	1
Строка	Число	1
Строка	Строка	0 – совпадают, 1 – не совпадают
Строка	Диапазон	1
Диапазон	Пусто	1
Диапазон	Строка	1

Эвклидово пространство (Википедия)