February - 2016

ИСПОЛЬЗОВАНИЕ ОНТОЛОГИЙ ДЛЯ СОЗДАНИЯ СЦЕНАРИЕВ ОБРАБОТКИ ДАННЫХ

Prof. Larysa Globa
National Technical University of Ukraine
“Kiev Polytechnic Institute”

План

1. Big Data, пути преодоления

2. Модели и методы представления и обработки данных
   - 2.1. Онтологические модели
   - 2.2. Метаописания
   - 2.3. Обнаружение web-сервисов в мульти-онтологической среде
   - 2.4. Динамическое формирование и выполнение сценариев обработки заданий пользователя

3. Комплекс инструментальных средств построения сценариев обработки данных

4. Заключение

BIG DATA

Увеличение : V_4

$V_1 \times V_2 \times V_3 \times V_4$,

V_1 - *объем данных;*

V_2 - *скорость прироста как объемов данных, так и их обработки;*

V_3 - *многообразие данных(как структурированных, так и полу-, неструктурированных как семантически, так и синтаксически;*

V_4 – *стоимость (ценность) данных;*

BIG DATA

Объем неструктурированных данных увеличивается лавинообразно, включает:

high-definition video

неподвижные изображения,

получаемые от постоянно *растущего числа мобильных устройств с камерами*

Главные причины увеличения объемов:

- *больше информации распределяется между бизнес-партнерами*;
- *увеличивается число каналов и источников сбора информации от клиентов*, таких как мобильные apps;
- *больше Интернет-связанных устройств*;
- *больше онлайн источников информации*, напр. social media.

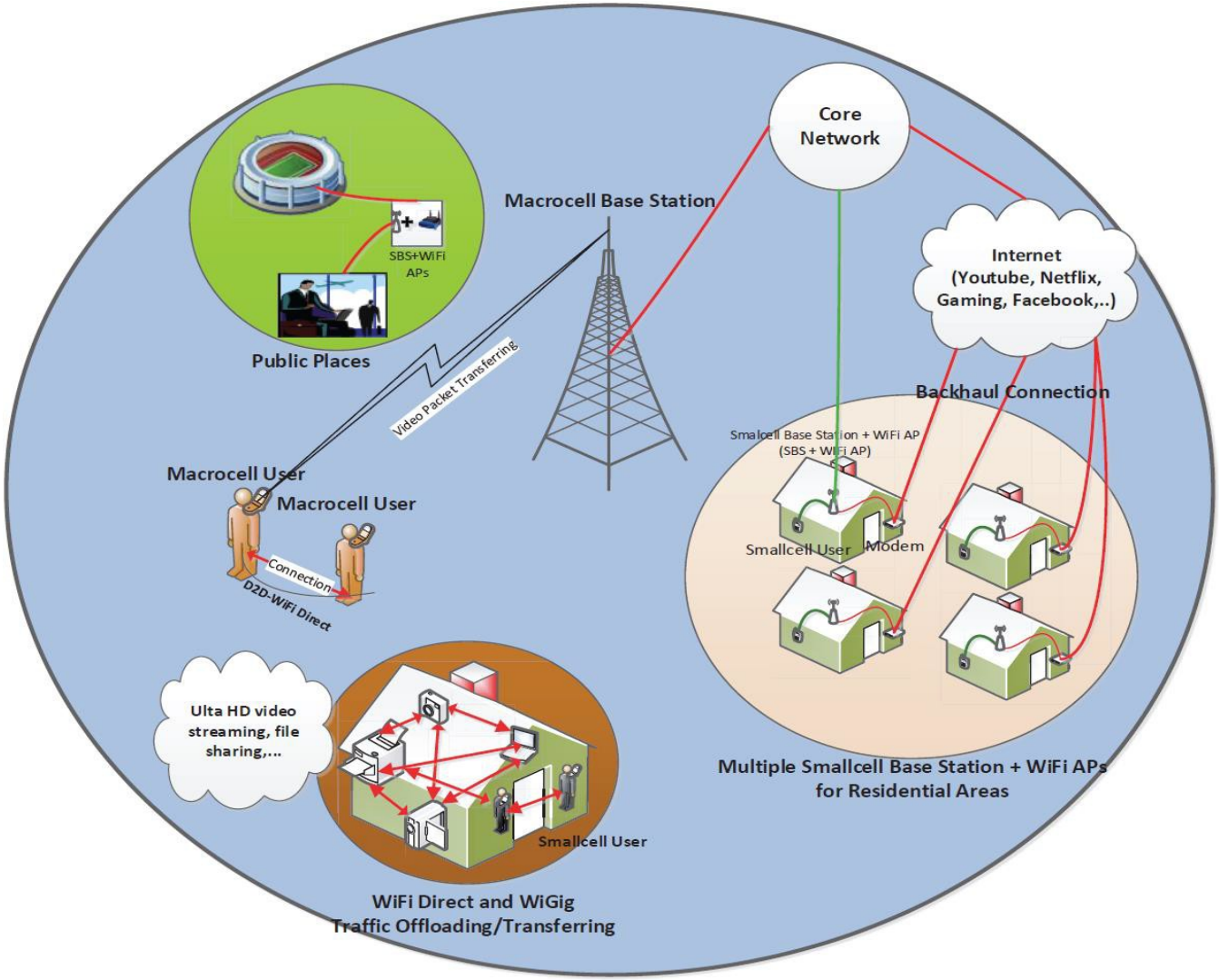
DATBIGA

ДАННЫЕ

ИНФОРМАЦИЯ

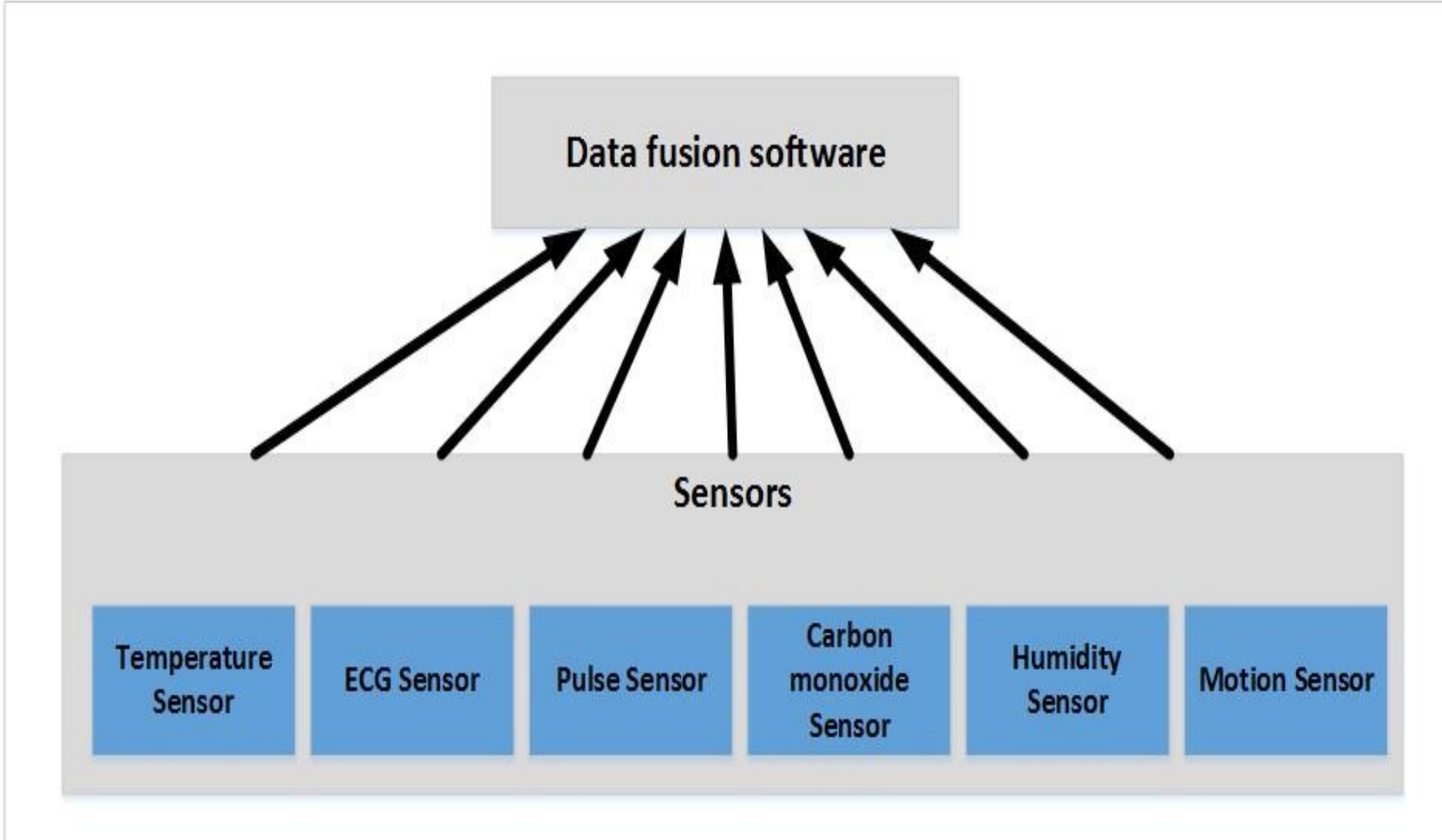
ЗНАНИЯ
(ПРОДУЦИРОВАНИЕ НОВЫХ ЗНАНИЙ)

BIG DATA: Все технологи и сети работают вместе!

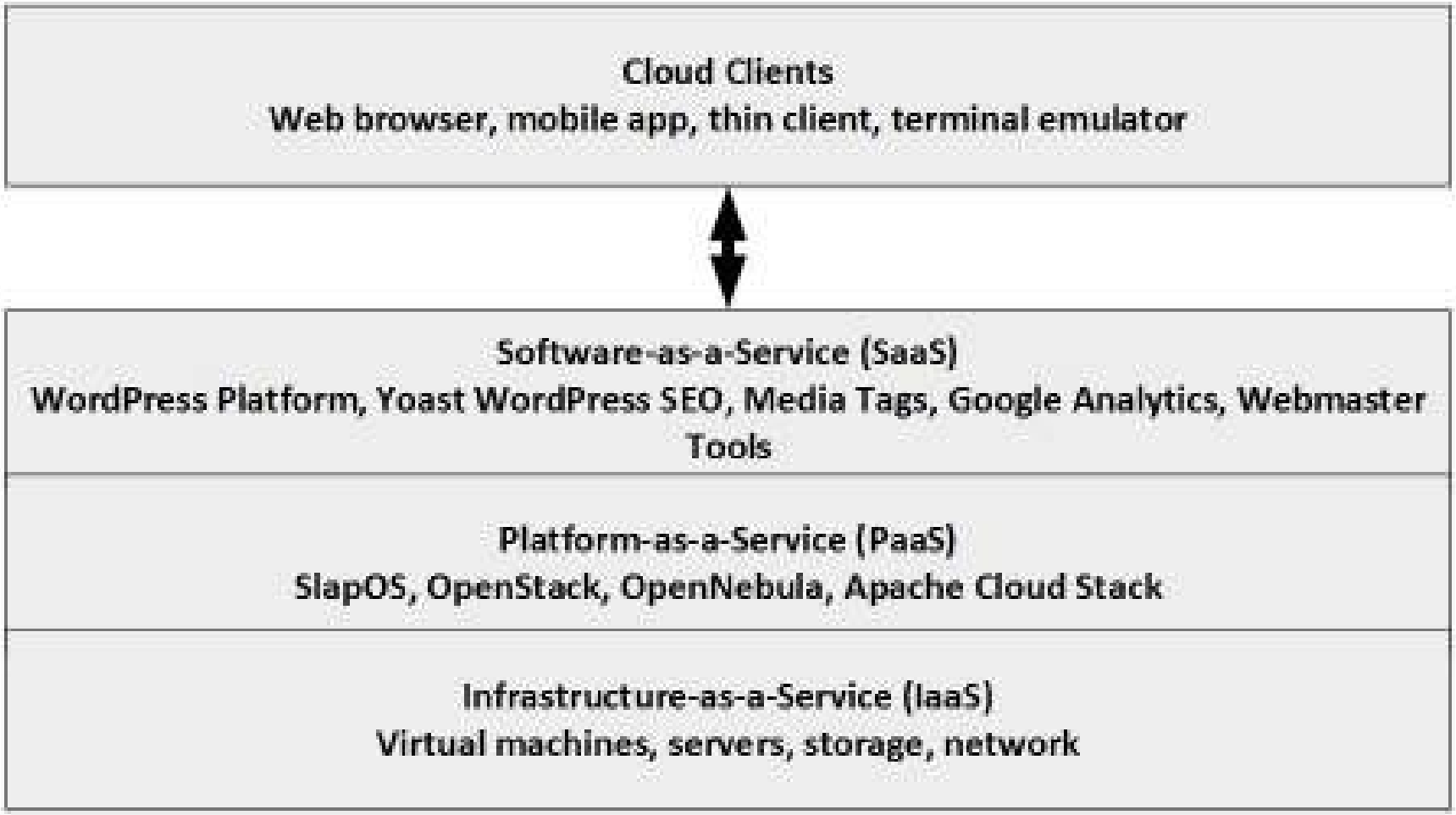


РАСПРЕДЕЛЕННЫЕ ДАТА ЦЕНТРЫ

(обычно Cloud Data Centers, M2M, D2D и т.д.…)

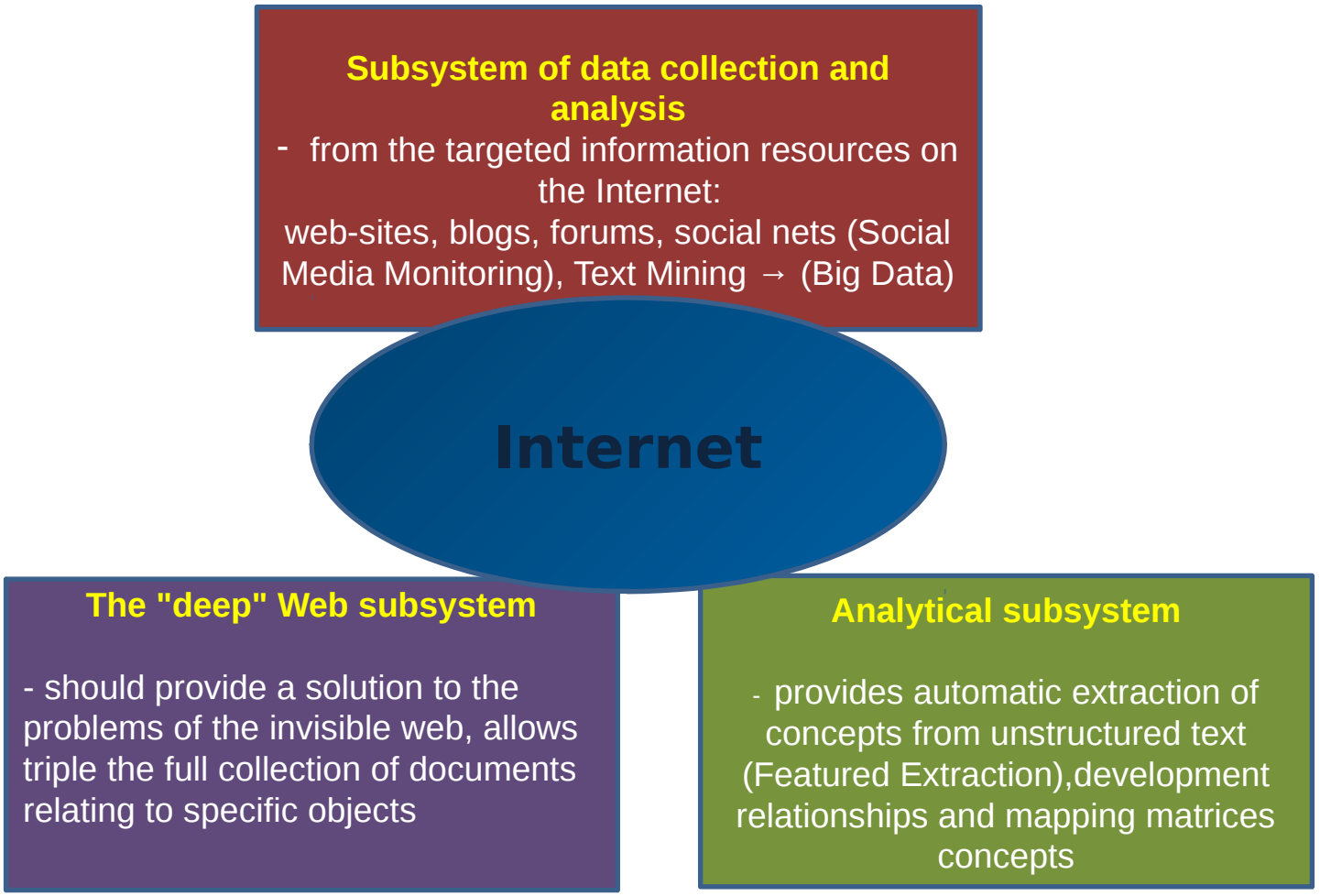


ПРИМЕР: Компоненты платформы ускоренной обработки



РАСПРЕДЕЛЕННЫЕ ДАТА ЦЕНТРЫ

(обычно Cloud Data Centers, M2M, D2D и т.д.…)



ЗАДАЧИ:

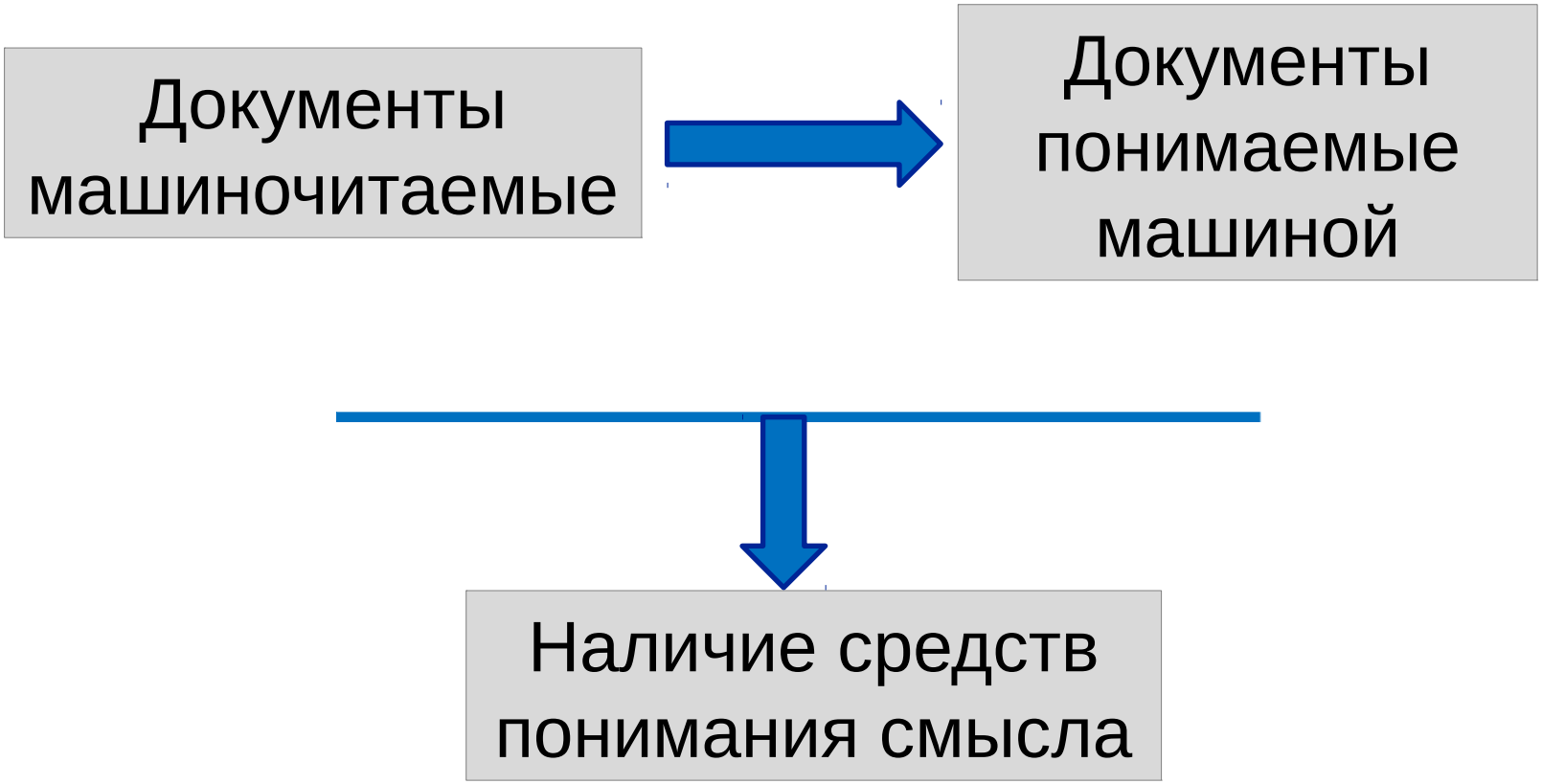
Для создания виртуальных распределённых Дата Центров необходимо реализовать:

- *программно-определяемые хранилища,* которые поддерживают виртуализацию *ресурсов хранения;*
- *ПО*, которое *управляет распределенными хранилищами*, *отделено* от разрозненного аппаратного обеспечения;
- *технологии Semantic Web,* способные создать Современную Систему Управления Знаниями;
- *использование Онтологий* для создания IT-технологий получения информации и знаний из данных;
- *технологии формирования сценариев обработки*

Предлагается: 3 группы математических моделей и методов

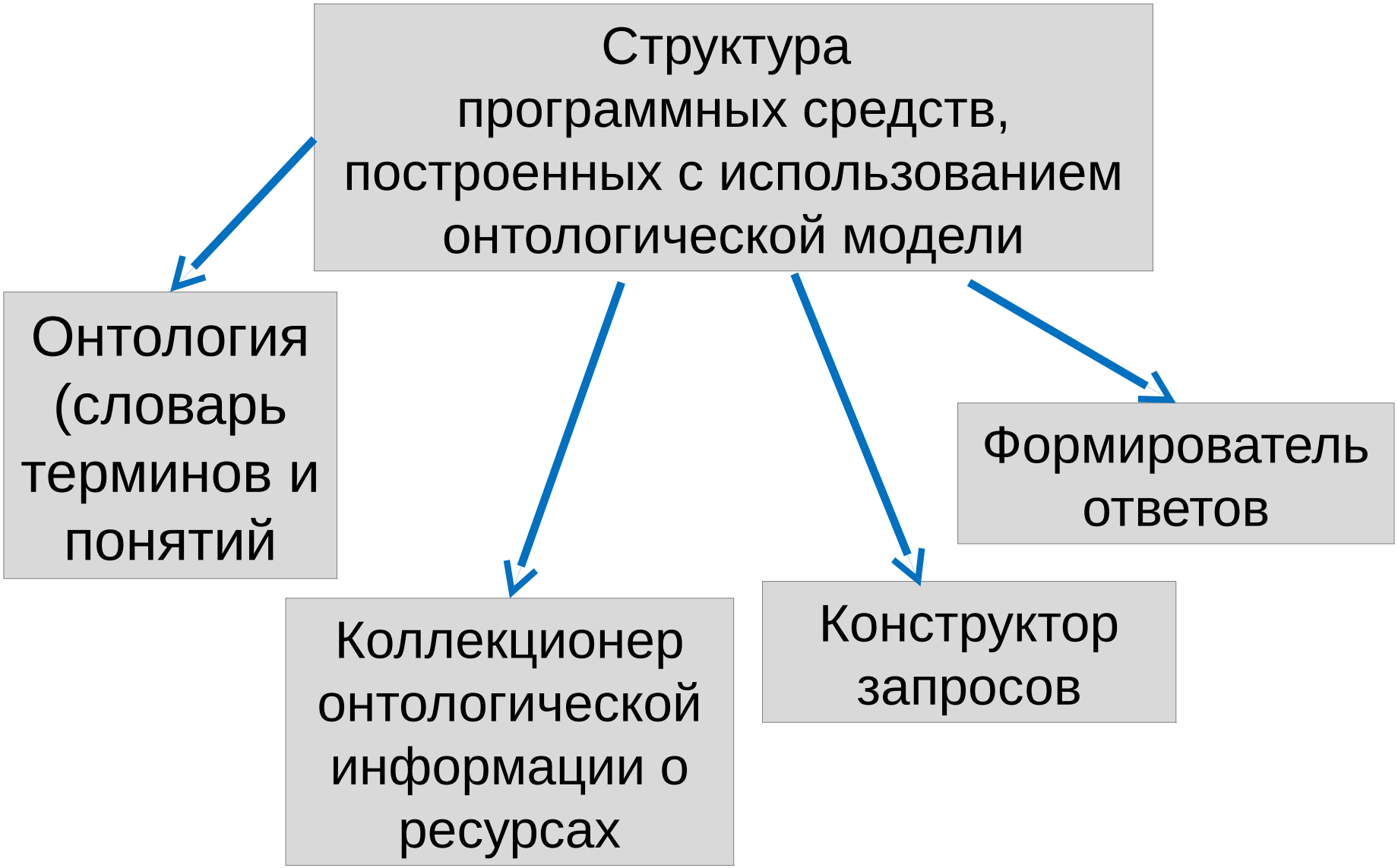
- *онтологические модели* как средство описания наборов данных в сети Интернет;
- *теория метаграфов* для формирования баз знаний и работы с ними;
- *нечеткая логика* как средство обработки больших объемов нечеткой информации и извлечения знаний из них;

2. Модели и методы представления и обработки данных



Процесс интеллектуальной обработки требует автоматизированного выполнения ряда этапов:

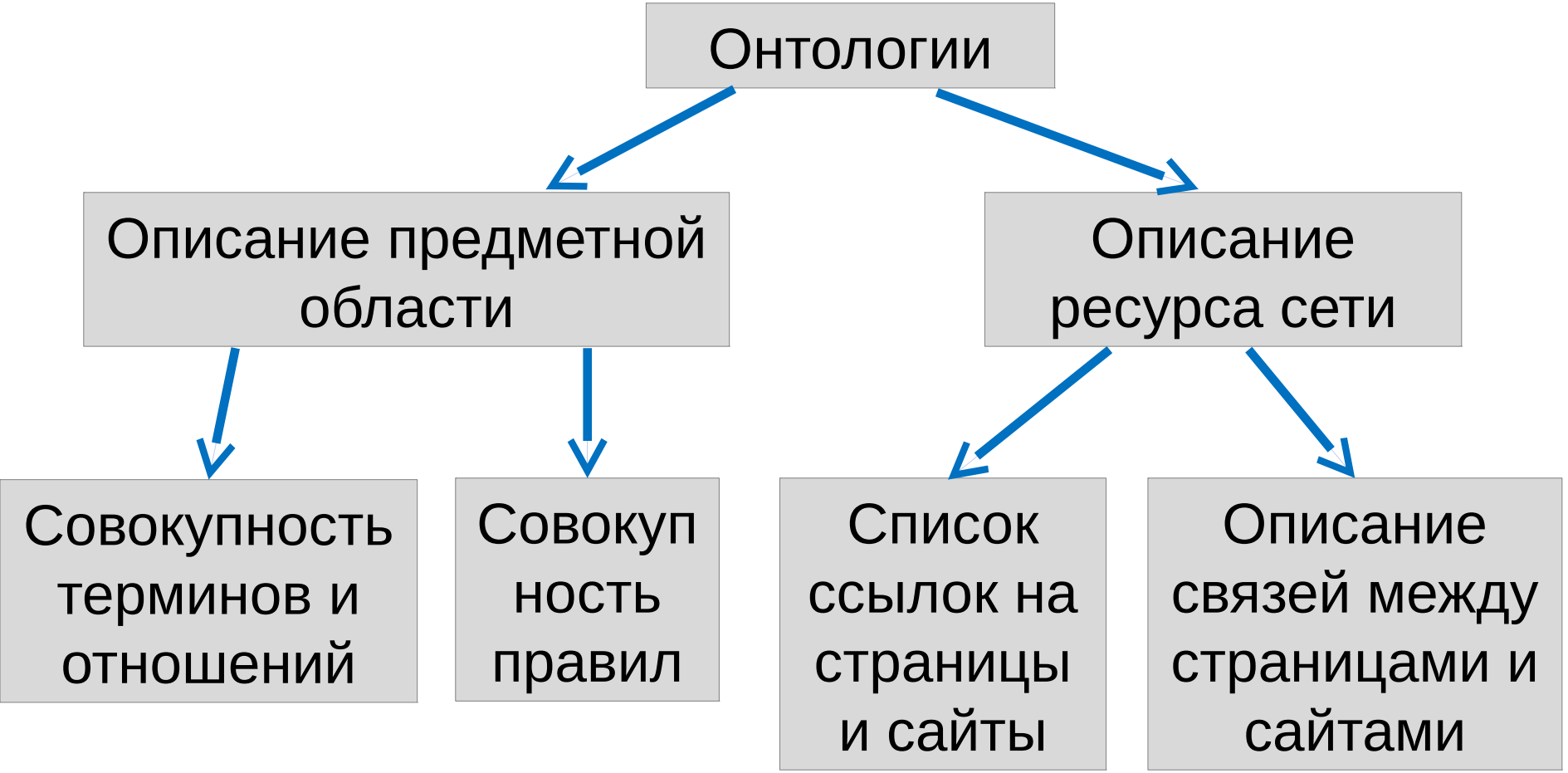
- *создания онтологий,* связывающих ресурсы под конкретные задачи на основании онтологических моделей;
- *формирования поисковых запросов «на лету»* с учетом возможной *неточности метаописаний* ;
- их *эффективного исполнения* на основании *автоматизированных рабочих процессов* (*сценариев обработки информации*, учитывая альтернативные).


Структура программных средств, построенных с использованием онтологической модели
Онтология (словарь терминов и понятий
Коллекционер онтологической информации о ресурсах
Конструктор запросов
Формирователь ответов

2.1 Онтологические модели

Онтологические модели можно использовать для:

- моделирования доменов программных сред;
- формирования автоматизированных рассуждений и получения выводов;
- автоматического программирования и удовлетворения спецификаций;
- построения программных систем, основанных на моделях;
- технологий семантики (онтологий и семантического web).


Онтологии
Описание предметной области
Описание ресурса сети
Совокупность терминов и отношений
Совокуп ность правил
Список ссылок на страницы и сайты
Описание связей между страницами и сайтами

Основные преимущества использования онтологических моделей:

- Онтологии делают систему *предметно-ориентированной*, т.к. пользователь задает поисковый запрос в хорошо известных терминах предметной области;
- Онтологии - интеллектуальное средство *поиска ресурсов* в сети Интернет, используют *методы представления и обработки знаний и запросов*;
- Онтологии - точно и эффективно описывают *семантику данных* для предметной области, решают проблему *несовместимости и противоречивости* данных;
- Онтологии имеют *собственные средства обработки (логического вывода)* и соответствующие *модели и методы семантической обработки данных*.

2.2. Метаописания

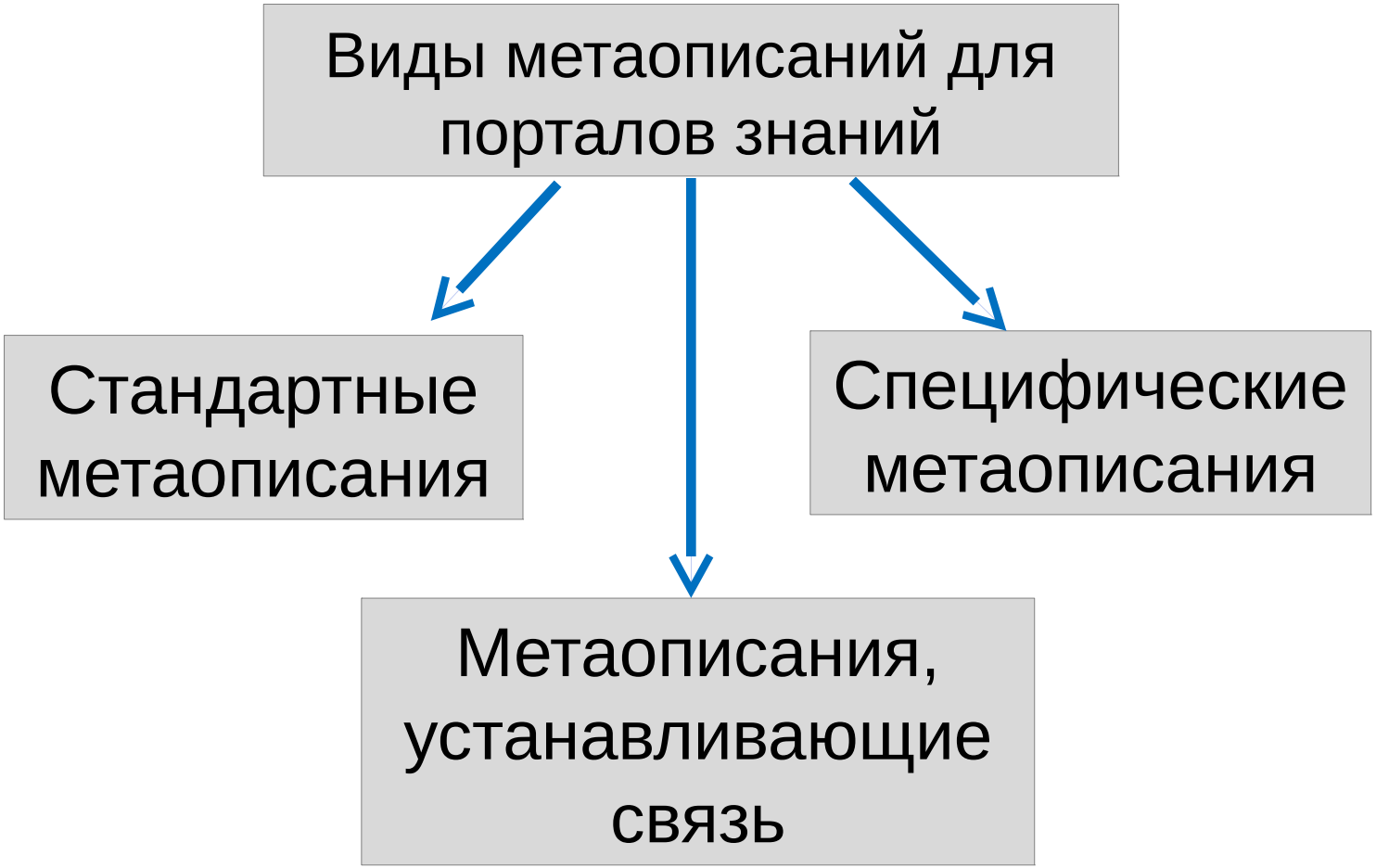
Метаописания устанавливают связи между объектами.

Метаописание – объект для смыслового определения элементов системы, информационных и вычислительных ресурсов.

Метаописания однозначно идентифицируют элементы и характеризуются двойкой:

$$M_i = (A, V)$$

где A – атрибут, V – значение атрибута.


Виды метаописаний для порталов знаний
Стандартные метаописания
Метаописания, устанавливающие связь
Специфические метаописания

Правило объедения для установления связей между метаописаниями :

«Если два элемента имеют одинаковые метаописания, то они связаны».

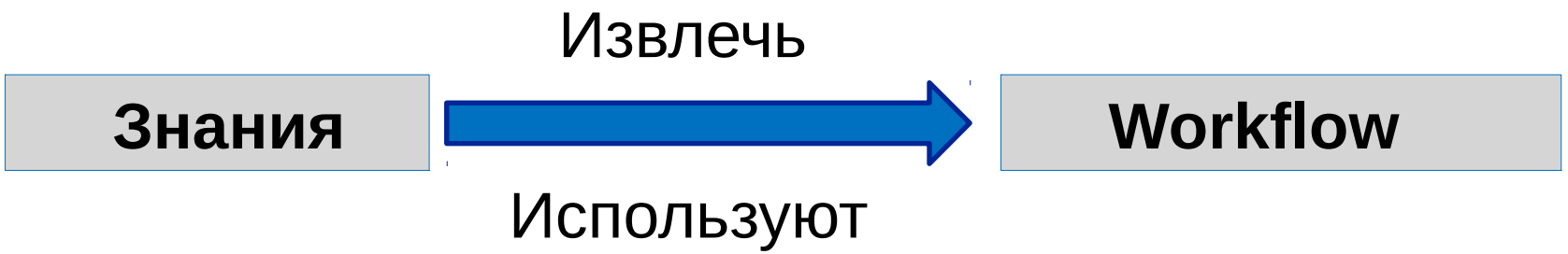
Существует два вида связности:

1) элементы компонуются для объединения в общий элемент;
2) один из элементов включается как составной элемент в исходный.

Сравнение по метаописаниям не является точным или хотя бы числовым требует

принятия решения о принадлежности элементов к тем или иным классам (решение задачи классификации на нечеткой области значений)

2.3. Обнаружение web-сервисов в мульти-онтологической среде



Workflow = **U Web-сервисов**

Проблема неоднозначности языковых конструкций при обнаружении web-сервисов в Интернет

требует

Создания web-служб поиска на основе метаописаний онтологических понятий

Специализированный инструментарий поиска web-сервисов в среде Интернет:

- DAML (DARPA (Agent Markup Language);
- RDF-графов (Resource Description Framework);
- LARKS (Language for Advertisement and Request for Knowledge Sharing;
- LSD (Learning Source Descriptions).

Релевантность обнаружения web-сервисов
предлагается использовать
семантическую близость с учетом *входов и выходов операций*

Подход к поиску web-сервисов в нескольких онтологических средах состоит из трех последовательных этапов:

- *Создание «шаблона сервиса»*, базируясь на запросе пользователя;
- *Сравнение «шаблона сервиса» с несколькими web-сервисами*, которые были определены как сервисы-кандидаты;
- *Возвращение web-сервисов*, удовлетворяющих минимально приемлемой оценке сходства с запрошенными пользователем в упорядоченном списке.

Семантический шаблон сервиса описывает запрос пользователя и рассматривается как промежуточная абстракция - прокси web-сервис.

Шаблон сервиса может быть определен как:

$$ST = \langle N_{ST}, D_{ST}, OPs_{ST}(N_{OP}, D_{OP}, O_{OP}, I_{OP}) \rangle$$

где N_{ST} - имя web-сервиса,

D_{ST} - текстовое описание web-сервиса,

OPs_{ST} - множество операций web-сервиса.

Каждая из операций web-сервиса в свою очередь определяется с помощью: N_{OP} - имени операции, D_{OP} - текстовое описания операции, O_{OP} и I_{OP} - входных и выходных параметров операции.

Шаблон сервиса сравнивается с набором сервисов -кандидатов (CS) - шаблонов сервисов.

При сравнении используют для оценки синтаксическое и семантическое сходство, определяемое как:

$$\Theta(ST, CS) = \frac{w_H \cdot H(ST,CS) + w_\Phi \cdot \Phi(ST,CS)}{w_H + w_\Phi}$$, где

$\Theta(ST, CS)$ - обобщенная оценка,

$H(ST, CS)$ - синтаксическое сходство,

$\Phi(ST, CS)$ - семантическое (функциональное) сходство,

w_i - весовой коэффициент,

соответствующий каждому типу сходства, предназначен для более гибкого управления критерием сравнения.

2.4. Динамическое формирование и выполнение сценариев обработки заданий пользователя

Проектирование последовательности элементов сценариев обработки заданий пользователя (Сложный сценарий, workflow)

Состоит из

Упорядоченное объединение соответствующих частичных элементов сценария (подрасчет решения конкретной задачи, web-сервис)

Формальное описание сценариев Сэ:

1) множество сложных сСэ:

$R^{з} \ni n_k^{з}, n_k^{з} = (T_k^{з}, p_{kj}^{з})$, где

$R^{з}$ - множество сложных сценариев сСэ,

$n_k^{з}$ - *k*-й сценарий Сэ из множества сложных сценариев сСэ,

$T_k^{з}$ - название *k*-ого сценария Сэ из множества сложных сценариев сСэ,

$p_{kj}^{з}$ - *p*-й параметр *k*-ого сценария Сэ из множества сложных сценариев сСэ.

2) множество частичных сценариев чСэ:

$R^{ч} \ni n_l^{ч}, n_l^{ч} = (T_l^{ч}, p_{lq}^{ч})$, где

$R^{ч}$ - множество частичных сценариев чСэ,

$n_l^{ч}$ - *l*-й сценарий Сэ из множества частичных сценариев чСэ,

$T_l^{ч}$ - название *l*-ого сценария Сэ из множества частичных сценариев чСэ,

$p_{lq}^{ч}$ - *q*-й параметр *l*-ого сценария Сэ из мн-ва частичных сценариев Сэ.

Дерево сценария - упорядоченное,
с корневым узлом и
заданным порядком прохождения дочерних узлов

Последовательность выполнения частичных сценариев Сэ соответствует
введенному порядку на дереве и
выполняется в процессе построения дерева
(на «лету»)

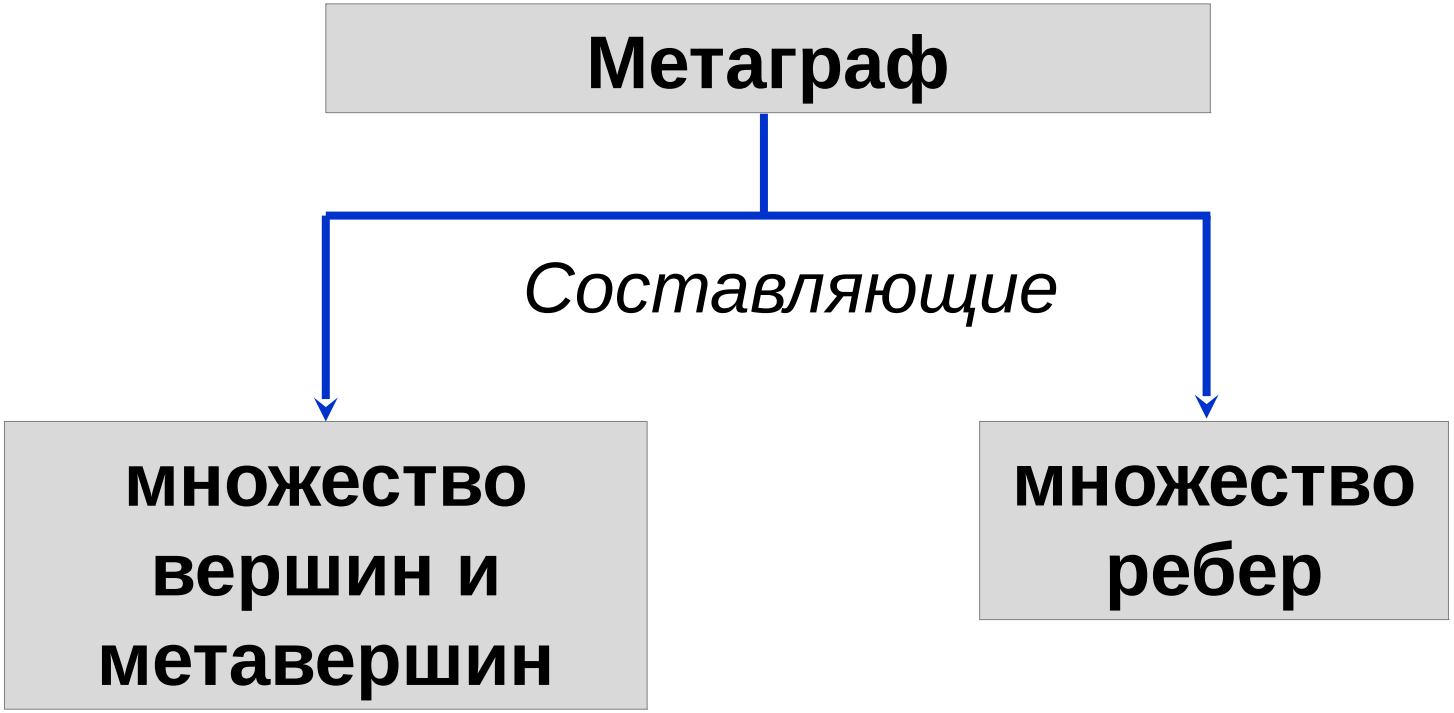
Этапы метода динамического формированию сложного сценария:

Этап 1. На первом этапе происходит отбрасывание из множества $R^{ч}$ сценариев, в которых ни один $p^{з}_{kj}$ не равняется никакому $p^{ч}_{lq}$

Этап 2. На этом этапе происходит сравнение значений сценариев, отобранных на этапе 1, для отсечения подмножества $p^{ч}_{lq}$ сценариев, которые имеют общие параметры, но множества их значений не пересекаются

Этап 3. На этапе 3 проводится упрощение формулы частичного сценария Сэ.

Для визуального анализа полученного на «лету» сценария предлагается использовать метаграфы



Множество ребер будет содержать все ребра метаграфа, независимо от того, какие типы узлов метаграфа они соединяют.

Применение метаграфов даёт возможность:

- моделировть бизнес-процессы с использованием диаграмм выполнения работ (workflow);
- визуально представлять и анализировать аномалии в нечетких наборах данных.

Визуальный анализ аномалий в метаграфе полученного workflow может позволить избежать *ошибочных решений*.

Использование онтологий

+

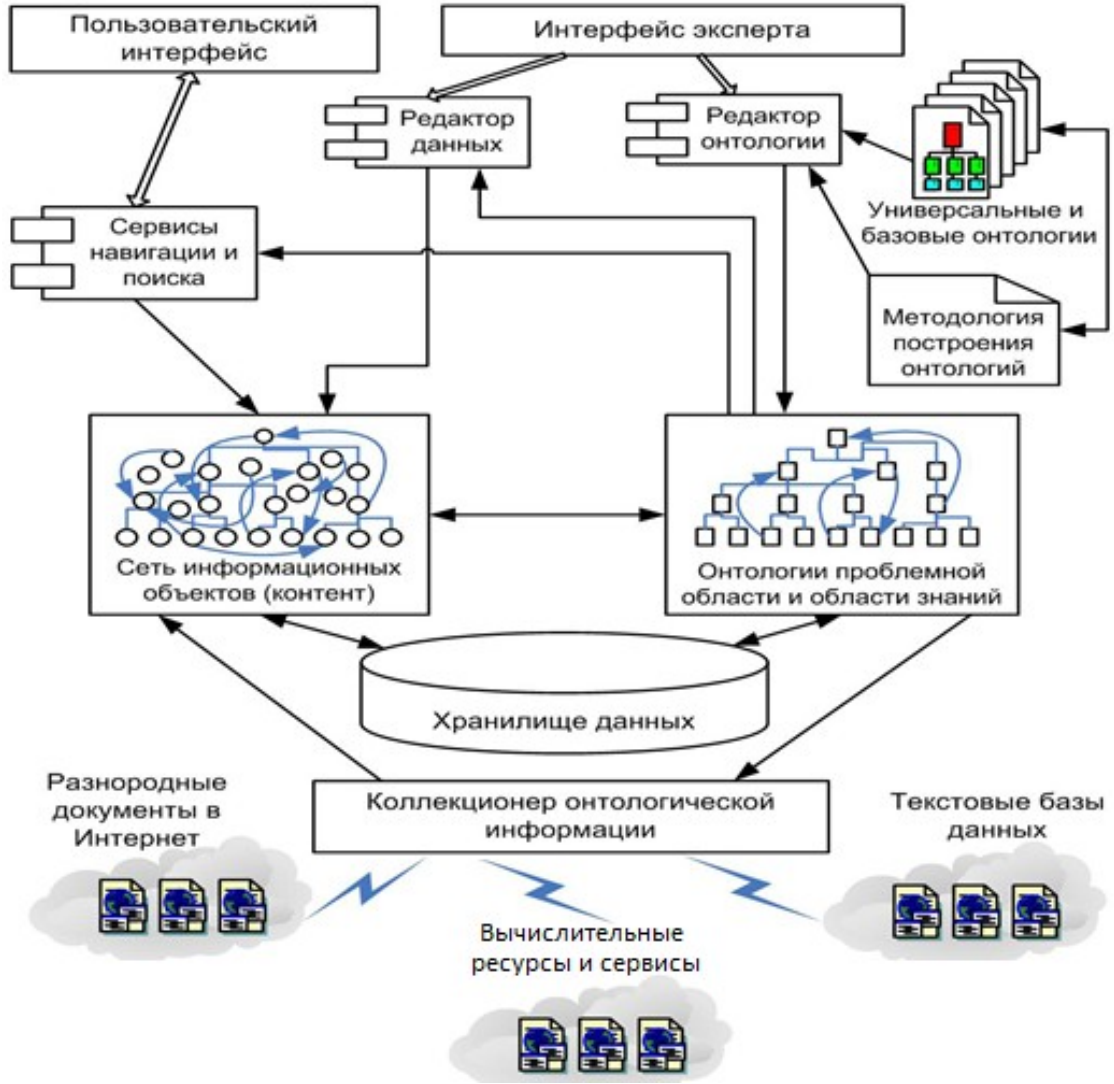
Методы семантического поиска

Целостная методология для создания сценариев (workflow) решения сложных задач в среде Интернет

Визуализация сценариев обработки информации в виде метаграфов повышает релевантность полученных решений

3. Комплекс инструментальных средств построения сценариев обработки данных

Подсистема ведения онтологий на портале знаний



4. Заключение

1. Комплекс инструментальных средств построения бизнес-процессов и обработки больших объемов данных позволяет:

- предоставлять *пользователям доступ к данным*, которые расположенным в Интернет-среде;
- предоставлять *данные группе конечных пользователей* в форме, которая *соответствует их коллективному представлению* о данных;
- *сокращать время ответа* на запрос;
- предоставлять данные, *структурированные в соответствии с требованиями доступа*;
- *упрощать выполнение задач* очистки, загрузки, преобразования, интеграции и анализа данных;
- *формировать сценарии* обработки данных «на лету».

2. Рассмотрен подход к автоматизации построения сценариев обработки данных, отличающийся *использованием онтологических моделей* для описания информационных и вычислительных ресурсов, *теории метаграфов* для визуализации и анализа полученных сценариев

3. Подход эффективен в сложных сферах деятельности, таких как инженерия и научные исследования.

4. *Онтологии* в комплексе *с методами семантического поиска* позволяют сформировать *сценарии* (*workflow*) решения сложных наукоемких задач в среде Интернет.

5. *Визуализация сценариев* обработки информации в виде *метаграфов* позволяет *повысить релевантность* полученных «на лету» сценариев

6. *Использование шаблонов упрощает* выполнение сложных наукоемких расчетов широкому кругу пользователей в различных сферах деятельности.

Вопросы, требующие дополнительного исследования:

- исследование *неоднозначности построенных сценариев* в случае *смысловой нечеткости описания онтологий*, *многозначности описаний* одних и тех же ресурсов в среде Интернет;
- развитие подхода к созданию сценариев обработки больших объемов данных с целью *повышения производительности их обработки*.