СИСТЕМА СУБЪЕКТ-ОБЪЕКТНЫХ ОТНОШЕНИЙ ПРИ ОРГАНИЗАЦИИ ХРАНЕНИЯ И ИСПОЛЬЗОВАНИЯ НАУЧНЫХ ДАННЫХ В ИДЕОЛОГИИ FAIR

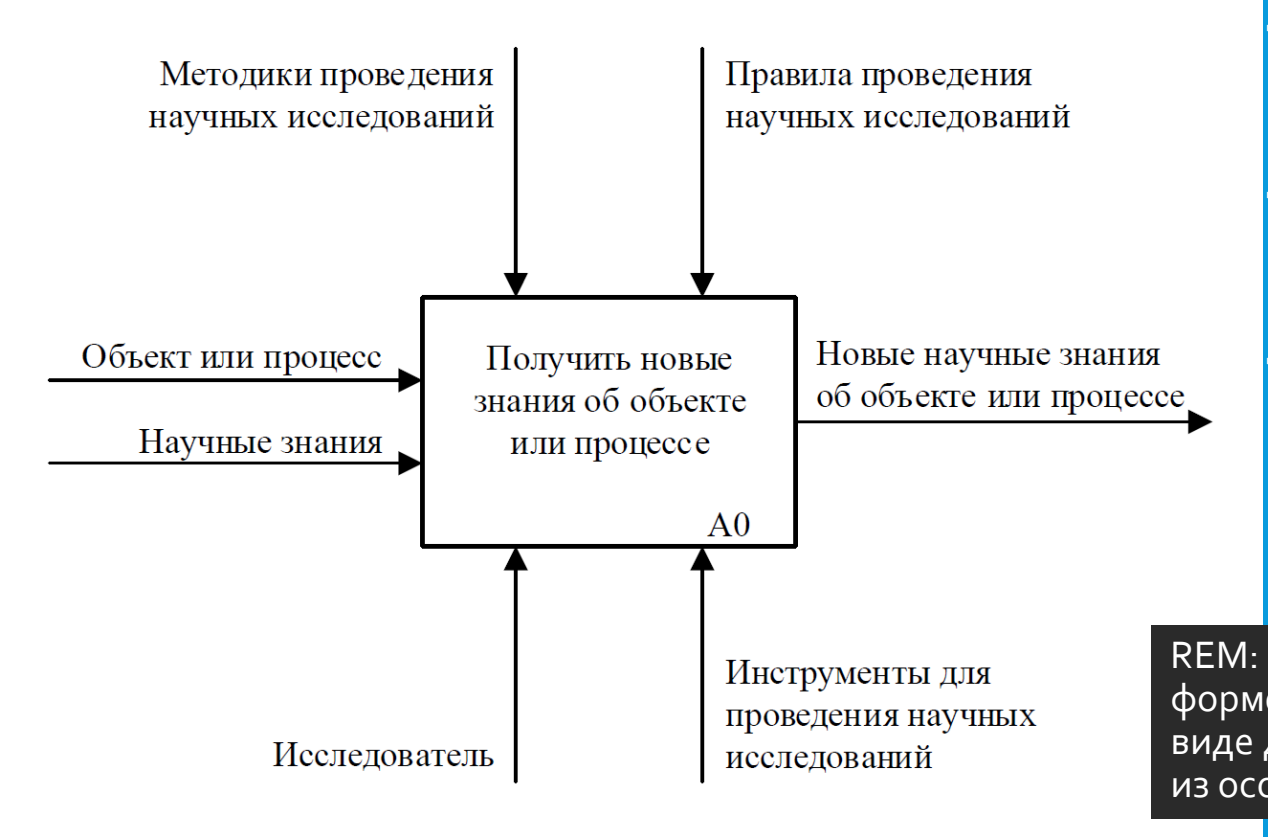
Юрченко Андрей Васильевич

Федеральный исследовательский центр информационных и вычислительных технологий

КОНТЕКСТ: МИССИЯ – ЦЕЛЬ – ОБЪЕКТ – ПРЕДМЕТ ИССЛЕДОВАНИЙ

- Миссия работы -
  обеспечение необходимыми инструментами исследований, основанных на интенсивном использовании данных
- Целью настоящего исследования является
  разработка методологических основ, организационных решений и инструментальных средств для организации хранения и использования научных данных
- Объект исследования –
  научные данные и процессы работы с научными данными
- Предмет исследования –
  методы, технологии и средства организации научных данных, их хранения и использования, управления ими, включая нормативную базу и инфраструктуру научных исследований, основанных на данных

КОНТЕКСТ: ОПИСАНИЕ ИССЛЕДОВАТЕЛЬСКОГО ПРОЦЕССА – ВЕРХНИЙ УРОВЕНЬ



- Цель научных исследований – получение новых научных знаний об изучаемых объектах
- Формально, приступая к исследованию мы пользуемся всем накопленным багажом знаний современной науки
- Аналогично, при выполнении исследований мы пользуемся сложившимися правилами (в том числе для описания результатов) и инструментами в рамках научного метода

REM: научные знания в их объективизированной форме, т.е. записанные в виде текстов, оформленные в виде документов или иным способом, являются одним из особых видов научных данных

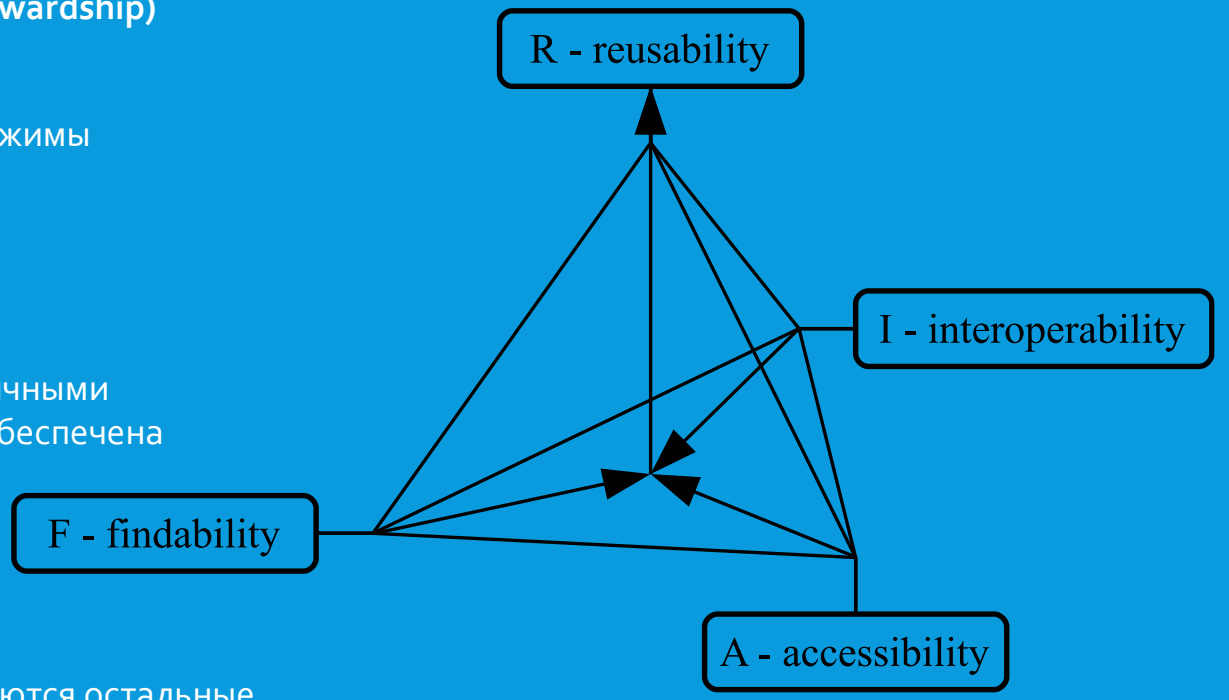
КОНТЕКСТ: НАУЧНЫЕ ДАННЫЕ

- Под научными данными будем понимать любые данные в цифровом виде, генерируемые в ходе научных исследований и / или используемые в них
- Основными особенностями научных данных, заставляющими искать новые подходы к организации их хранения и использования, являются:
  - множественность источников данных,
  - неоднородность данных и их форматов,
  - разное качество данных,
  - большие объемы данных,
  - необходимость обмениваться и делиться данными,
  - разнообразие и постоянное развитие методов и средств для анализа данных, в том числе – многообразие форм их визуального представления
  - потребность в интеграции разнородных данных,
  - необходимость использования высокопроизводительных ресурсов.

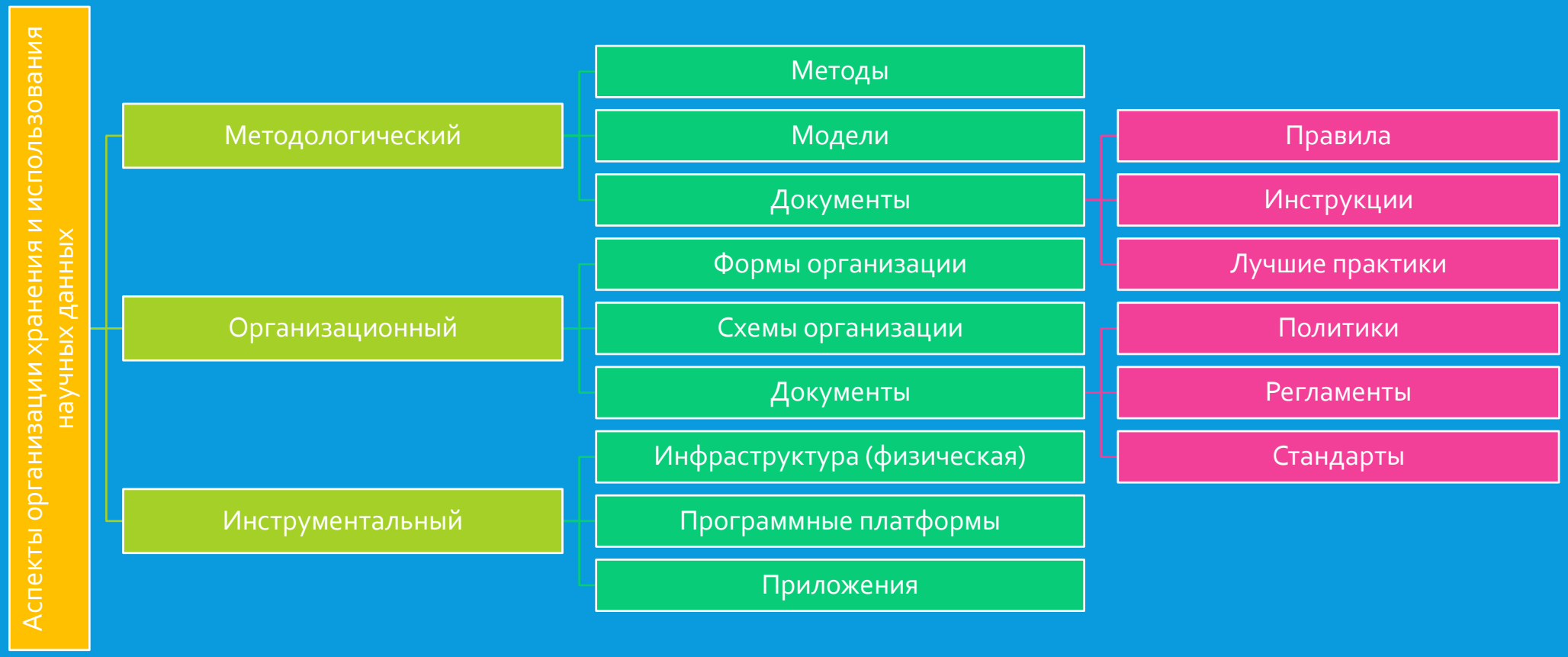
КОНТЕКСТ: FAIR DATA PRINCIPLES

В 2016 году консорциумом ученых и организаций опубликованы **«Руководящие принципы FAIR для управления научными данными и их сопровождения» (The FAIR Guiding Principles for scientific data management and stewardship)**

- F – Findable: для того, чтобы использовать существующие данные, их нужно найти, они должны быть обнаружимы
- A – Available: данные и метаданные должны быть достижимы и доступны для получения
- I – Interoperable: данные объединяются (интегрируются), анализируются, хранятся и обрабатываются различными приложениями, сервисами, системами, должна быть обеспечена совместимость
- R – Reusable: основная цель FAIR – оптимизация переиспользования, т.е. повторного использования данных, этому принципу «переиспользуемости» и подчиняются остальные принципы FAIR



КОНТЕКСТ: ДЕКОМПОЗИЦИЯ ПРОБЛЕМЫ НА МЕТОДОЛОГИЧЕСКИЙ, ОРГАНИЗАЦИОННЫЙ И ИНСТРУМЕНТАЛЬНЫЙ АСПЕКТЫ

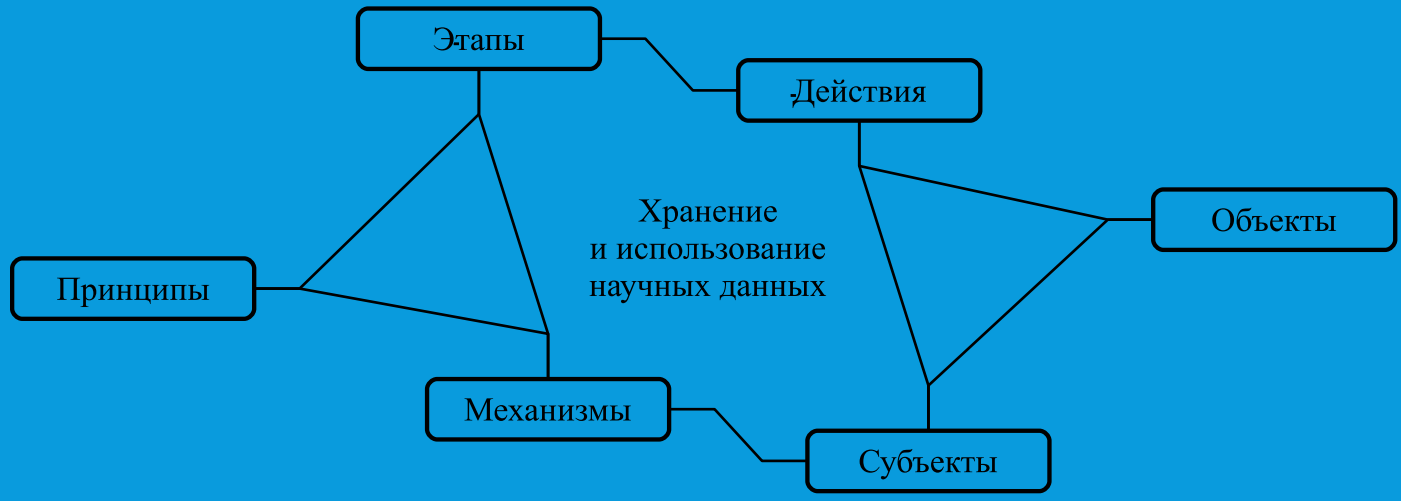


КОНТЕКСТ: РЕЗЮМЕ

- **Ключевой целью при организации хранения и использования научных данных определено повышение эффективности их использования**
- Эффективность использования научных данных должна выражаться в активном **переиспользовании** их различными исследовательскими группами, в решении новых научных задач, индикатором чего является рост числа соответствующих публикаций
- Способы достижения цели разделяются на три основных уровня-аспекта: методологический, организационный и инструментальный
- Основная задача исследования заключается в систематизации проблем, возникающих при организации хранения и использования научных данных, формировании комплексов соответствующих им целей, определении (разработке, комплексировании, …) способов их достижения путем разработки и построения соответствующих моделей, методов и технологий, инфраструктурных и инструментальных решений

КОРНИ ИЕРАРХИИ КЛАССОВ-ПОНЯТИЙ – БАЗОВЫЕ ПОНЯТИЯ

Основой иерархии понятий предметной области «Организация хранения и использования научных данных» являются две тройки классов:



- В рамках тройки *объект – субъект – действие*
  - субъект совершает действие (над объектом)
  - объект подвергается воздействию, т.е. действию со стороны субъекта
- В рамках тройки *принцип – механизм – этап*
  - принцип реализуется через механизм
  - механизм применяется на этапе

ОБЪЕКТЫ

В рамках проблемы организации хранения и использования научных данных ключевыми объектами являются ресурсы (МВ: из винограда искрящееся шампанское) – это

- научные данные в форме отдельных объектов данных и коллекций,
- методы, реализуемые в форме различных инструментальных средств;

Кроме ресурсов – вспомогательные объекты

- описания или объекты метаданных,
- связи и отношения между объектами и субъектами,
- классы и свойства ресурсов и связей / отношений.

ДЕЙСТВИЯ - ОБЪЕКТЫ

Объекты данных предполагают в отношении себя пять базовых действий

- создание,
- (со)хранение,
- получение / передачу,
- изменение,
- использование,
- удаление

Дополнительные виды действий в отношении объектов данных

- описание
- построение связей
- оценка (верификация, ранжирование, оценка качества)
- классификация

...

ДЕЙСТВИЯ - СУБЪЕКТЫ

Действие, как основная форма субъект-объектных отношений, может быть инициировано:

- человеком, например, пользователем информационной системы;
- программно по расписанию либо при выполнении другого внутреннего критерия (а не только достижения заданного момента времени);
- программно, как реакция на внешнее воздействие / фактор

Таким образом, субъектом отношений в рамках предметной области организации работы с научными данными может быть и человек, и программа. Выделим ряд ролей субъектов:

- создатели ресурсов, владельцы ресурсов, администраторы ресурсов, регуляторы, консультанты-стюарты («библиотекари данных»), пользователи ресурсов.

АКТ

- Любое действие реализуется (выполняется) с помощью какого-либо механизма или инструмента на основе определенного метода, инициируясь субъектом и имея в качестве результата набор объектов (в том числе, пустой). Этот процесс в его завершенной форме назовем **актом**.
- Акт определяется на следующих множествах:
  - множество всех объектов $O = D \cup M \cup E \cup R \cup C$, где
  - $D = \{d_i\} \equiv D_O \cup D_C$ – объединение множеств объектов данных $D_O = \{d_{O,i}\}$ и коллекций данных $D_C = \{d_{C,i}\}$,
  - $M = \{m_i\}$ – множество методов,
  - $E = \{e_i\}$ – множество описаний / метаданных,
  - $R = \{r_i\}$ – множество связей,
  - $C = \{c_i\}$ – множество классов и свойств
  - множество субъектов $B = \{b_i\} \equiv B_M \cup B_H$, где
  - $B_M = \{b_{M,i}\}$ – программно-аппаратные системы
  - $B_H = \{b_{H,i}\}$ – люди
  - множество целевых входов и выходов $F = F_I \cup F_O$, где
  - $F_I = \{f_{I,i}\}$ – входящие потоки данных
  - $F_O = \{f_{O,i}\}$ – исходящие потоки данных
- Дополнительно введем множества $W = B \cup O \cup F$ и $\hat{O} = O \cup F$.

МЕТОДЫ

- Метод является многоместным отображением из декартовой степени множества элементов W в декартову степень множества элементов W:

$$m: W^n \to W^k, n \geq 0, k \geq 0.$$

- Не всякий метод применим к произвольному набору элементов W, поэтому

– метод – лишь частичное отображение $m: (W^n) \to W^k$;

– каждый метод определяет класс $W|_m$ элементов W, к которым он применим, за исключением методов с $n = 0$;

– каждый метод, за исключением методов с $k = 0$, определяет и класс $W|^m$ элементов W, которые были спродуцированы его применением.

ПРИМЕРЫ ОСОБЫХ КЛАССОВ МЕТОДОВ

- Методы, устанавливающие связи m: $W^n \rightarrow R,\ n > 1$
- Методы, формирующие комбинации элементов m: $W^n \rightarrow W,\ n > 1$
  - методы, формирующие коллекции объектов научных данных m: $D^n \rightarrow D_C$
  - методы, формирующие «рабочие процессы» или «потоки работ» (Workflows) m: $M^n \rightarrow M$
- Обучающие методы, которые в обобщенном виде представимы, как m: $O^n \rightarrow M$, а в наиболее распространенном случае – m: $D_C \rightarrow M$, например нейросеть с алгоритмом ее обучения на наборе данных

АКТ

- В рамках акта выделяются создаваемые элементы, инициировавший акт субъект и метод. Методы на настоящем этапе будем считать единой сущностью с механизмами и инструментами их реализации.
- Акт, как отношение, можно записать в виде

$$a: B \times W^n \times M \to W^k, n \geq 0, k \geq 0.$$

- Акты могут завершаться успешно или неуспешно, с различными причинами неуспешности: недопустимость, аварийный останов, внешне инициированный (субъектом) останов, это будет одним из свойств акта. Еще одним свойством акта будут две временные метки: начала и завершения акта.
- Обозначим множество совершенных актов $\{a_i\} \equiv A_R$, оно будет пополняемым / расширяемым множеством регистрации изменений во множестве W, а его элементы a_i, однажды возникнув, будут неизменяемы.
- Происхождение объектов в O за пределами актов – недопустимо , а возникновение новых субъектов в B и потоков из F обязательно сопровождается актом их регистрации путем создания объекта метаданных $e \in E$.

АКТ

- Фиксация (т.е. запись в читаемой / распознаваемой форме) актов позволяет накапливать информацию о происхождении и эволюции объектов и других элементов W, их взаимосвязях, что, в свою очередь, дает возможность формировать систему знаний об элементах W и строить на ее основе рекомендательную систему по их возможному использованию для решения новых исследовательских задач. При этом важно, что фиксируется информация и об использованном методе (инструменте), и об инициировавшем акт субъекте.

ОСНОВНЫЕ ВИДЫ АКТОВ

- Основными видами актов будут акты первичной генерации объектов данных и акты преобразования объектов и коллекций данных в новые объекты или коллекции данных.
- Акт a первичной генерации объекта данных связан с потоком данных f и инициировавшим его субъектом b, итогом его будет объект данных d_O, формируемый с помощью метода m, и комплекс связей $\boldsymbol{r} \in R^l,\ l \geq 0$ между участвующими в акте элементами W.
  В этом случае применение метода можно записать, как $d_O = m(f)$, а сам акт будет шестеркой $a =< b, f, m, d_O, \boldsymbol{r}, p >$, где p – набор свойств акта, таких как успешность выполнения и метки времени начала и завершения.
- Акт a преобразования объекта (или коллекции) данных d_{inp} в новый объект (коллекцию) d_{out} также записывается как шестерка $a =< b, d_{inp}, m, d_{out}, \boldsymbol{r}, p >$, а применение метода – как $d_{out} = m(d_{inp})$.
- Представление акта в виде таких шестерок можно признать общим, записав его в виде

$$a =< b, \boldsymbol{w}_{inp}, m, \boldsymbol{w}_{out}, \boldsymbol{r}, p >, \text{ где } \boldsymbol{w}_{inp} \in W^n, \boldsymbol{w}_{out} \in W^k, \boldsymbol{r} \in R^l.$$

СИСТЕМА СУБЪЕКТ-ОБЪЕКТНЫХ ОТНОШЕНИЙ

- Принимая в качестве носителя множество W построим систему субъект-объектных отношений предметной области хранения и использования научных данных

$$S \subset < B \times W^n \times M > \times < W^k \times R^l \times A_R > (n, k, l \geq 0).$$

- Система будет динамической и открытой:

– W будет расширяться новыми элементами (далее $w_i \in W$ будем называть элементами, элементы частного подмножества $o_i \in O$ продолжим называть объектами), как извне, через входящие потоки $f_I \in F_I$, так и в результате внутренних процессов и преобразований объектов $o \in O$;

– система будет взаимодействовать с внешней средой через исходящие потоки $f_O \in F_O$;

– вводится начальное состояние системы через соответствующий носитель W_0, который содержит, как минимум, описания базовых субъектов $b_i \in B_0$ и базовых методов $m_i \in M_0$, к которым относятся методы регистрации субъектов и потоков данных от источников и направление данных за пределы системы, создания (в системе) новых методов.

К СИСТЕМЕ СУБЪЕКТ-ОБЪЕКТНЫХ ОТНОШЕНИЙ

- Всякий элемент, за исключением элементов начального состояния W_0, порожден в результате совершения какого-либо акта, т.е.

$$\forall w \in W \setminus W_0 \; \exists a \in A_R : a =< b, \boldsymbol{w}_{inp}, m, \boldsymbol{w}_{out}, \boldsymbol{r}, p > \; \& \; w \in \boldsymbol{w}_{out}.$$

- У каждого акта $a \in A_R$ определяется такое свойство, как момент завершения его выполнения $t_{fin}(a)$, поэтому для системы в целом можно определить понятие состояния на момент времени t, которое задается в виде:

$$A_R^t = \{a_i \in A_R : t_{fin}(a_i) \leq t\} \subseteq A_R.$$

- Носитель системы на момент времени t определяется следующим образом

$$W^t = \{w \in W : \; \exists a \in A_R^t, a =< b, \boldsymbol{w}_{inp}, m, \boldsymbol{w}_{out}, \boldsymbol{r}, p > \; \& \; w \in \boldsymbol{w}_{out}\}.$$

- Ценность этого объекта заключается в возможности фиксировать выборки-подмножества W при формировании коллекций данных с помощью механизмов типа фильтрации и поиска на всем множестве W.

В ЗАКЛЮЧЕНИЕ

- Система S строится как математическая формализация субъект-объектных отношений, возникающих при организации хранения и использования данных, с упором на применимость к особому классу данных – научным данным
- Так как переиспользуемость является одной из ключевых ценностей научных данных, то формирование вокруг объектов научных данных системы связей с другими объектами данных и не только данных: субъектами, методами и инструментами – путь к их «просветлению», т.е. превращению из понятного только автору набора в источник, доступный и понятный другим исследователям
- Формализация понятия акта, фиксация (запись) актов, формализация состояния системы и ее носителя на момент времени t дают возможность использовать инструменты теории систем для решения задачи построения системы для организации хранения и использования научных данных, ориентированной на их переиспользование, как с целью верификации результатов чужих исследований, так и с целью получения новых знаний путем интеграции данных и методов

СПАСИБО

Юрченко Андрей Васильевич, ФИЦ ИВТ, yurchenko@ict.nsc.ru