Методы, модели и средства организации хранения и использования научных данных

ЮРЧЕНКО АНДРЕЙ ВАСИЛЬЕВИЧ, ФИЦ ИВТ (НОВОСИБИРСК)

Введение

Мотивация

- **Цифровые данные – ключевой драйвер современной науки**
- Различные детекторы, датчики, сенсоры, их сети, компьютерное моделирование, также как и просто действия пользователей компьютерных сетей и подключенных к ней устройств **генерируют большие объемы данных**, которые используются или могут быть использованы в научных исследованиях
- Качественная и целеустремленная **организация хранения и использования** (в том числе – переиспользования) научных данных **необходима для повышения «отдачи»** от них, в том числе, позволяя **ставить и решать новые научные задачи** и проблемы.

поэтому

Вопросы выбора, адаптации и создания средств, методов, моделей для работы с научными данными, формирования на их основе комплексных решений и систем, **организации** с их помощью **хранения и использования научных данных – актуальная проблема современной науки**

Научные данные и их особенности

Под ***научными данными*** будем понимать любые данные в цифровом виде, генерируемые в ходе научных исследований и / или используемые в них.

Особенности научных данных:

- множественность источников данных
- неоднородность данных и их форматов
- разное качество данных
- большие объемы данных и необходимость использования высокопроизводительных ресурсов при работе с ними
- необходимость обмениваться и делиться данными
- разнообразие и постоянное развитие методов и средств для анализа и обработки данных
- потребность в интеграции разнородных данных

Контекст проблемы

- Задачи организации работы с научными данными соответствуют целям Национальных проектов «Цифровая экономика» и «Наука».
- Они направлены на повышение качества и увеличение числа передовых научных исследований путем «цифрового» упорядочивания их процессов и результатов.

Контекст проблемы

- На подходы к решению задач организации хранения и использования научных данных непосредственно влияет идеология *Открытой науки*, активно развивающаяся, в первую очередь – в Европе.
- *Открытая наука* предполагает, что научные данные и инструменты их обработки и анализа также должны становиться открытыми, что позволит:
  - верифицировать получаемые результаты, повышая объективность исследований
  - проводить новые основанные на данных исследования за пределами группы, получившей эти данные

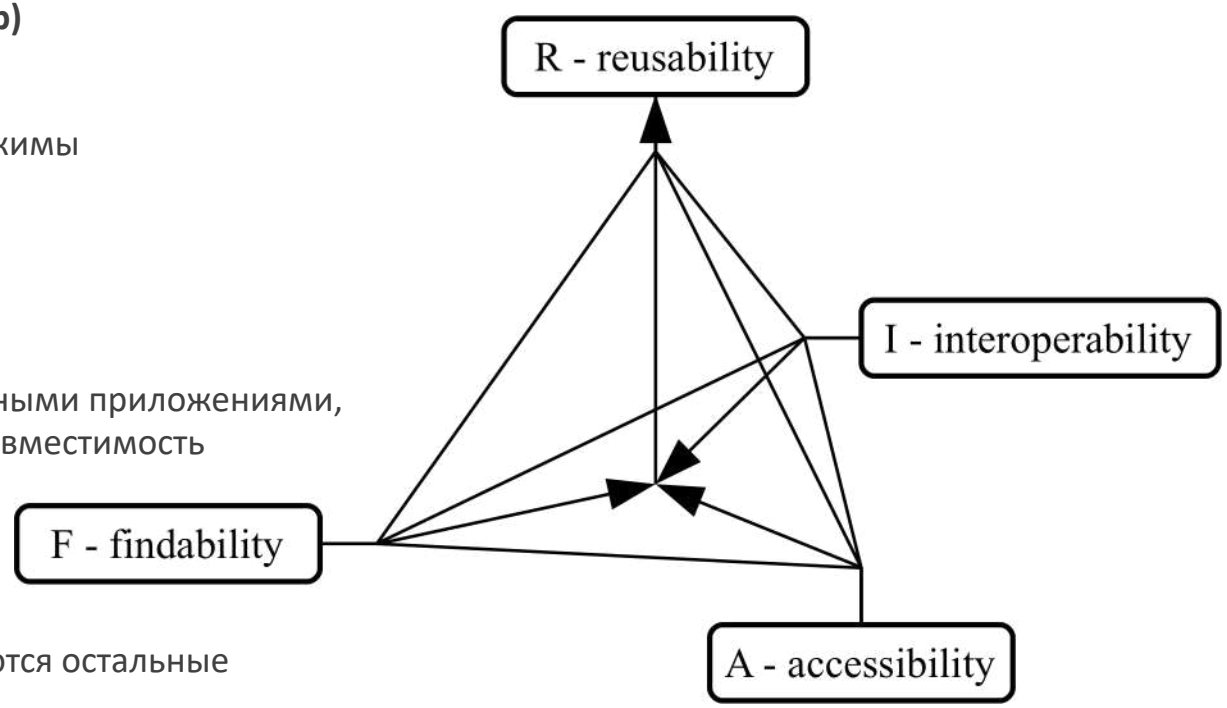
Инициативы в ЕС, США и глобальные

- Spatial Data Infrastructure – Инфраструктура пространственных данных
  - USA - Circular A-16 (1953, переработан в 1967, 1990, 2002) -> В 1994 основана Национальная инфраструктура пространственных данных (NSDI) -> Веха – в 2013 принята «Политика открытых данных» (Open Data Policy) и запущена инициатива «Общедоступные сервисы геопространственных данных» (Geospatial Shared Services) -> В 2016 принята «Стратегическая программа развития NDSI» (National Spatial Data Infrastructure (NSDI) Strategic Framework)
  - EU - DIRECTIVE 2007/2/EC (2007) -> Запущена The European INSPIRE initiative (Европейская инициатива создания инфраструктуры пространственной информации) -> Сроки реализации – 2007-2021
- EU – e-Infrastructure Reflection Group (e-IRG) – Основано в 2003 г. как сообщество для развития киберинфраструктуры научных исследований в Европе -> В 2009 опубликовало отчет e-IRG Report on Data Management – «Отчет об управлении данными»
- EU – Knowledge Exchange (с 2004 г.) – партнерство CSC (Финляндия), CNRS (Франция), the Danish Agency for Science and Higher Education (Дания), DFG (Германия), Jisc (Великобритания) и SURF (Нидерланды): “Мы работаем вместе, чтобы поддержать развитие цифровых инфраструктур, обеспечивающих открытые исследования. Мы информируем национальные и международные структуры и продвигаем общие подходы, чтобы позволить исследованиям пересекать национальные границы”
- EU - OpenAIRE - DRIVER I&2 в 2006-2009 -> OpenAIRE в 2008-2011 -> OpenAIRE+ в 2012-2015 -> OpenAIRE2020&OpenAIRE-Connect в 2015-2018 -> OpenAIRE Advance в 2018-2020 – Миссия: Сдвиг научной коммуникации в сторону открытости и прозрачности и содействие инновационным способам общения и мониторинга исследований
- OECD – В 2007 опубликованы OECD Principles and Guidelines for Access to Research Data from Public Funding (Принципы и руководства по доступу к научным данным, полученным при государственном финансировании)
- Worldwide – ICSU (International Science Council) – World Data System (с 2008 г.), цель - долгосрочное сопровождение и обеспечение универсального и справедливого доступа к научным данным и услугам, продуктам и информации гарантированного качества, идеология и поддержка реестра доверенных сервисов по работе с данными для глобальной науки (Trusted Data Services for Global Science)

FAIR Data Principles

В марте 2016 года 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



Go-FAIR инициатива

- В апреле 2016 Европейская комиссия обнародовала свои планы сделать все данные, полученные из финансируемых ЕС исследовательских проектов, доступными, совместимыми и повторно используемыми (FAIR).
- Тогда же. По оценкам Комиссии, 2 миллиарда евро на финансирование Horizon 2020 будет выделено на ее так называемую «Европейскую облачную инициативу» (European Open Science Cloud) для обеспечения исследований, основанных на данных.
- На саммите G20 в Ханчжоу в 2016 году лидеры G20 выступили с заявлением, в котором одобрили применение принципов FAIR к исследованиям.
- Создана инициатива Go-FAIR (go-fair.org).
- К 2020 г. Ученые, научные организации и сообщества, финансирующие науку фонды в Европе, США и в других странах мира разрабатывают и внедряют решения, направленные на реализацию FAIR Data Principles.

Исследования и разработки в ФИЦ ИВТ

- Начала:
  - *Шокин Ю.И., Жижимов О.Л., Пестунов И.А., Смирнов В.В., Синявский Ю.Н.* Распределенная информационно-аналитическая система для поиска, обработки и анализа пространственных данных // Вычислительные технологии. - 2007. - Т.12. - № спецвыпуск № 3. - С.108-115.
  - *Федотов А.М., Шокин Ю.И., Жижимов О.Л., Молородов Ю.И.* Служба директорий LDAP как единая информационная среда // Открытое и дистанционное образование. - 2007. - № 4(28). - С.31-41. - ISSN 1609-5944.
  - *Шокин Ю.И., Федорук М.П., Чубаров Д.Л., Юрченко А.В.* Высокопроизводительные вычисления в ИВТ СО РАН // Вычислительные технологии. - 2006. - Т.11. - № Избранные доклады X Российской конференции "Распределенные информационно-вычислительные ресурсы“ (DICR-2005), Новосибирск 6-8 октября 2005 г. - С.17-26.
  - *Жижимов О.Л., Мазов Н.А., Федотов А.М., Шокин Ю.И.* Сервер ZooPARK как сервер для построения распределенных информационных систем // Информационные технологии в высшем образовании. - 2005. - Т.2. - № 1. - С.53-67. - ISSN 1811-3346.

- К настоящему моменту:
  - *Чл.-корр РАН А.М. Федотов* с коллегами разрабатывал концепцию «Научной информационной системы» - каталогизированной распределённой ИС, позволяющей хранить, обрабатывать, распространять, анализировать, а также организовывать поиск в разнообразных коллекциях электронных (цифровых) документов (электронная библиотека)
  - *Д.ф.-м.н. О.Л. Жижимов* с коллегами развивал технологическую платформу интеграции разнородных распределенных данных ZooSPACE, предназначенной для обеспечения доступа к распределенным разнородным информационным ресурсам по протоколам SRW/SRU, Z39.50.
  - В Кемеровском филиале ФИЦ ИВТ под руководством *д.т.н. В.П. Потапова* осуществлены и ведутся разработки различных специализированных геопорталов, опирающихся на коллекции пространственных данных получаемых в результате дистанционного зондирования Земли из космоса и с наземных станций наблюдения (например, сейсмостанций).

Российские исследования, разработки, центры научных данных

- Данные дистанционного зондирования Земли и другие пространственные данные:
  - ИКИ РАН, Отдел «Технологии спутникового мониторинга», с 1999 года – ЦКП «ИКИ-Мониторинг», УНУ «BS ИКИ-Мониторинг» (Вега-Science), более 2,5 Петабайт данных онлайн
  - ФИЦ ИВТ, Лаборатория аэрокосмического мониторинга и обработки данных, с 2007 года – Центр мониторинга состояния природной среды и социально-экономических процессов, Информационная система спутниковых данных СО РАН, порядка 300 Терабайт данных онлайн
  - ВЦ ДВО РАН – ЦКП «Центр данных ДВО РАН» (на базе ПО «BS ИКИ-Мониторинг»)
- Мировые центры данных (WDS-certified)

  Геофизический центр РАН
  - по Солнечно-Земной физике
  - по физике твердой Земли

  ВНИИ гидрометеорологической информации
  - по метеорологии
  - по океанографии
- Центры хранения и обработки данных БАК и другие
  - Курчатовский институт – Федеральный ЦКП научным оборудованием "Комплекс моделирования и обработки данных исследовательских установок мега-класса« (с 2013 г.), компьютерный центр сети распределенных вычислений БАК (LHC-GRID)
  - Центр научных ИТ-сервисов ФИЦ ИВТ
  - ЦКП «Центр данных ДВО РАН»
  - …

Государственные инициативы РФ

- В рамках Федерального проекта «Информационная инфраструктура» Национального проекта «Цифровая экономика» к 2022 году должна быть создана отечественная **цифровая платформа сбора, обработки, хранения и распространения данных дистанционного зондирования Земли** из космоса, обеспечивающая потребности граждан, бизнеса и власти. Обеспечена реализация проекта «Цифровая Земля»
- В рамках Национального проекта «Наука» к 2022 году должна быть введена в эксплуатацию **цифровая система управления** сервисами научной инфраструктуры коллективного пользования (в том числе ЦКП и УНУ), предоставляющая безбарьерный доступ исследователям к заказу услуг с использованием инфраструктуры, в том числе – к **оцифрованным коллекциям и банкам данных** организаций, выполняющих исследования и разработки, …
  утверждены министерством ТЗ на разработку:
  - «Единой цифровой платформы научного и научно-технического взаимодействия, организации и проведения совместных исследований … » (ЦПСИ)
  - «Цифровой системы управления сервисами научной инфраструктуры коллективного пользования (в том числе ЦКП и УНУ), … » (АС УСНИКП)

Нерешенные проблемы

- Проблемная область работы с научными данными, в том числе, организации их хранения и использования является трудноформализуемой и в настоящее время состоит, преимущественно, из различных узкотематических исследований (case study)
- FAIR Data Principles становятся общепринятыми, однако для их реализации необходимо разрабатывать соответствующие механизмы, детализировать сами принципы, углублять и расширять их, наполняя смыслами
- Множество моделей жизненного цикла научных данных и отсутствие универсальной модели говорит и об отсутствии единого комплексного понимания, как организовывать управление научными данными
  и их сопровождение, и о необходимости исследовать саму возможность построения универсальной модели, а при невозможности – проработать методологию создания моделей под конкретные случаи
- Инфраструктура и инструменты организации работы с научными данными находятся в стадии активной разработки и в Европе, и в США, завершенных и готовых к переносу и внедрению универсальных решений нет
- В РФ пока полностью отсутствует какая-либо политика в отношении организации научных данных (кроме результатов их глубокой переработки в форму научных публикаций, отчетов НИР или патентов, для которых существует система регистрации), соответствующая инфраструктура находится в зачаточном состоянии, нормативных регулирующих актов практически нет

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

Задачи исследования

- **Формализация и постановка задач системного анализа и управления процессами работы с научными данными, организации их хранения и использования**
- **Разработка моделей описания процессов работы с научными данными, организации их хранения и использования**
- **Разработка методов решения задач и технологий организации хранения и использования научных данных, управления ими и обработки содержащейся в них информации**
- **Системный анализ, проектирование и реализация инфраструктурных и информационно-технологических решений, используемых при организации хранения и использования научных данных**
- **Разработка проблемно-ориентированной системы для управления аппаратно-программным инструментарием хранения и использования научных данных**

Постановка задачи системного анализа и управления процессами работы с научными данными, организации их хранения и использования

Постановка задачи: общая схема

Постановка задачи: базовые понятия – определения

- **Научное исследование** – это процесс изучения, эксперимента, концептуализации и проверки теории, связанной с получением научных знаний
- **Научные исследования, основанные на данных** (Data Driven Science) – это меж(мульти)дисциплинарная область, в которой используются научные методы, процессы, алгоритмы и системы для извлечения знаний и открытий из структурированных и неструктурированных данных
- **Научные данные** – это любые данные в цифровом виде, генерируемые в ходе научных исследований и / или используемые в них
- **Информационные технологии**, как область науки (Informational Science) – это область, связанная с анализом, сбором, классификацией, управлением (организацией), хранением, поиском, перемещением, распространением и защитой информации
- **Хранение и использование данных** – это ключевые процессы работы с данными, включающие размещение данных в системах хранения, их упорядочивание, поиск и агрегацию, обработку и анализ, обмен и опубликование
- **Организация** – это целевой процесс, деятельность по созданию или усовершенствованию взаимосвязей между частями и элементами с целью внесения упорядоченности в процессы и повышения их эффективности

Постановка задачи: формулировка основных проблем и целей

Проблема	Цель
Низкая эффективность использования научных данных	**Повысить эффективность использования научных данных**
Нежелание или невозможность (в т.ч. неумение) делиться данными и методами их анализа	Повысить мотивацию и создать условия для публикации данных и методов
Сложность поиска и подбора данных и методов их анализа для исследования	Облегчить поиск и комбинирование данных, подбор методов их анализа
Низкое качество данных и возникновение «цифрового мусора»	Повысить мотивацию к генерации качественных данных, разработать инструменты и критерии для фильтрации и утилизации данных
Недостаток ресурсов для размещения данных, для их обработки и анализа	Организовать создание и обеспечение доступа к ресурсам для хранения данных и их анализа
…	…

Постановка задачи: критерии оценки достижения целей

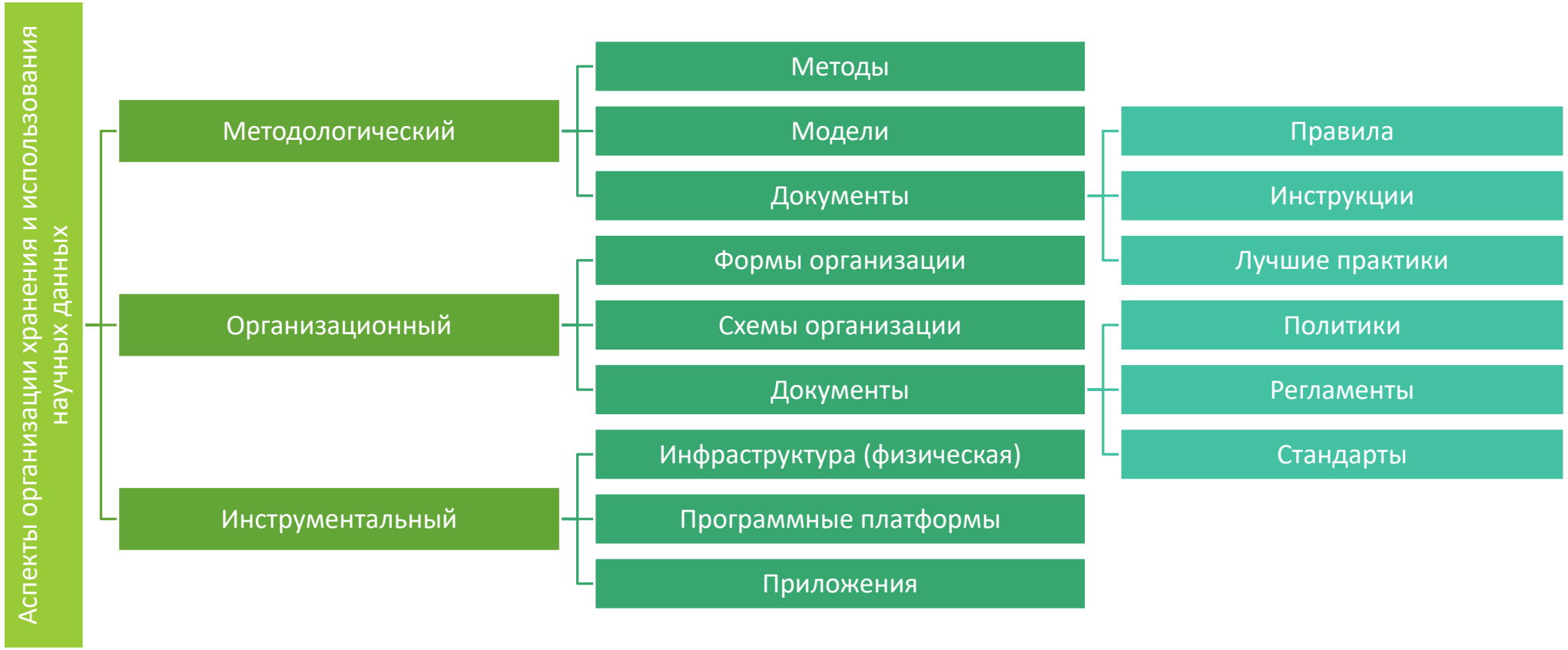
- Цель – повысить эффективность использования научных данных:
  - Количество (абсолютное, относительное) данных, использованных более чем в одном исследовании и более чем одной исследовательской единицей. Подтверждается публикацией результатов со ссылкой на использование данных в изданиях с определенным рейтингом
- Цель – повысить мотивацию и создать условия для публикации данных и методов:
  - Количество опубликованных данных и методов (алгоритмов)
- Цель – облегчить поиск и комбинирование данных, подбор методов их анализа:
  - Наличие средств для автоматизации поиска и создания наборов данных, формирования пулов методов для анализа данных под конкретное исследование
- Цель – повысить мотивацию к генерации качественных данных, разработать инструменты и критерии для фильтрации и утилизации данных:
  - Количество (доля) данных низкого качества (по установленным критериям) не превышает заданного порога
  - Разработаны и применяются при приеме, поиске и принятии решений об утилизации данных инструменты, реализующие критерии оценки качества данных
- Цель – организовать создание и обеспечение доступа к ресурсам для хранения данных и их анализа:
  - Количество ресурсов (в заданных единицах, например ПБ и Пфлопс) доступных для размещения данных и их анализа
  - Количество отказов в размещении данных достаточного качества не превышает заданного порога
  - Количество отказов и время ожидания (вычислительных) ресурсов для обработки и анализа данных не превышает заданного порога
- …

Постановка задачи: «облако» методов достижения целей

Необходимо рассмотреть и организовать как **систему** весь комплекс инструментов и средств, объектов и субъектов, возникающих между ними отношений при работе с научными данными.

Этот комплекс является трудноформализуемым, однако даже частичная его систематизация может дать большой синергетический эффект в решении задачи получения новых знаний о природе, человеке и их взаимоотношениях.

Постановка задачи: декомпозиция проблемы на методологический, организационный и инструментальный аспекты



Постановка задачи: 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

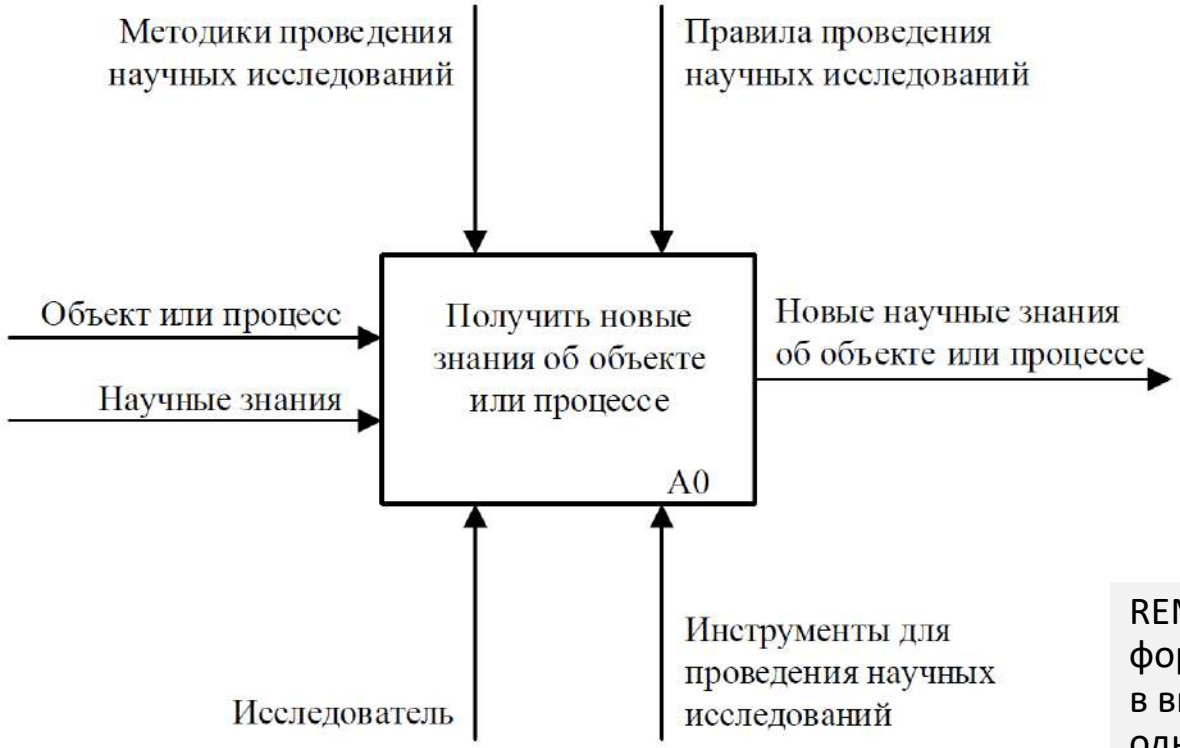
Модели описания процессов работы с научными данными, организации их хранения и использования

Модели: Иерархия классов-понятий – сопредельные понятия

Область «научные исследования» включает набор понятий, являющихся точками соприкосновения или взаимодействия с исследуемой областью, это

- **объект исследований** – явление, процесс, объект или комплекс объектов, которые изучаются в ходе исследований;
- **субъект исследований** – лицо, группа лиц или другая, более сложная организационная структура, осуществляющая исследования, отдельных представителей класса обобщенно будем также называть исследовательская единица или исследователь;
- **метод исследования** – определенная последовательность действий, приемов, операций, применяемая при научном исследовании.

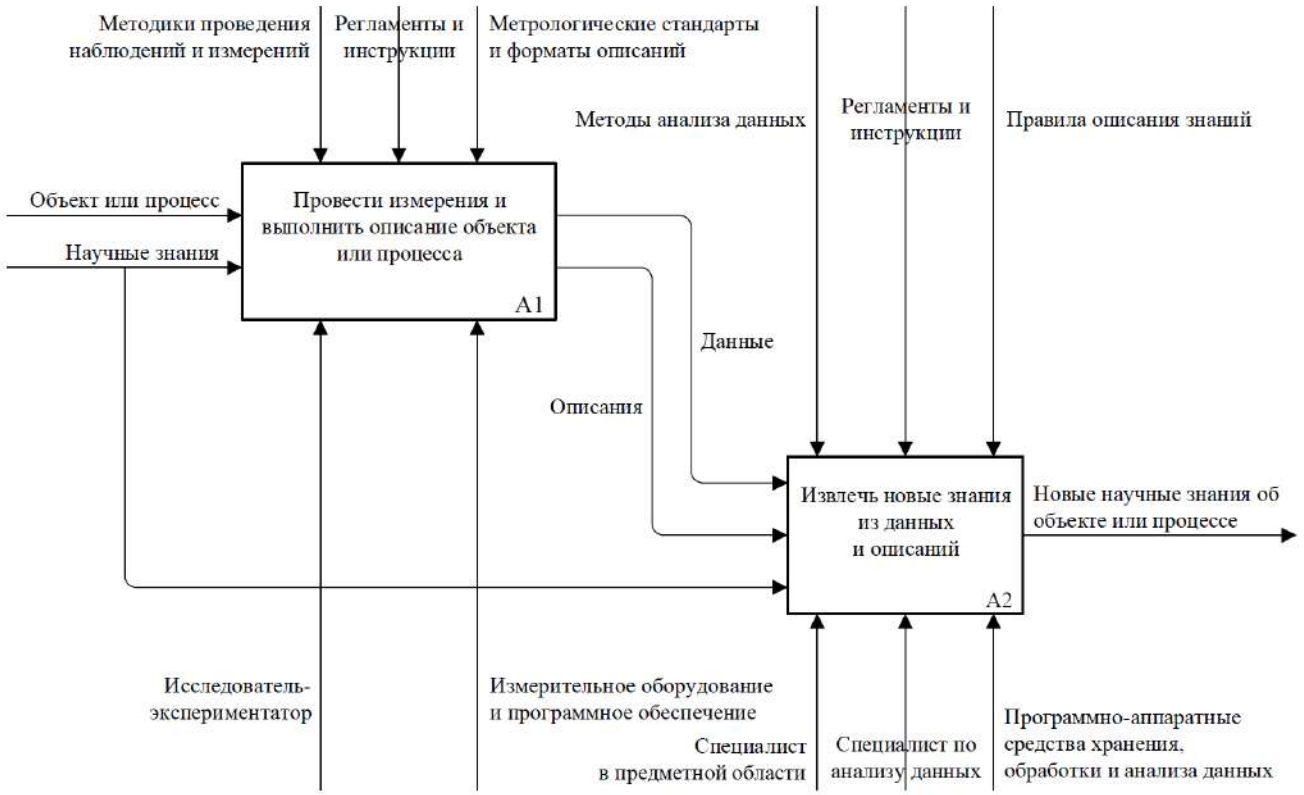
Модели: IDEF0 описание исследовательского процесса – верхний уровень



- 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

Модели: IDEF0 описание исследовательского процесса, основанного на анализе экспериментальных данных



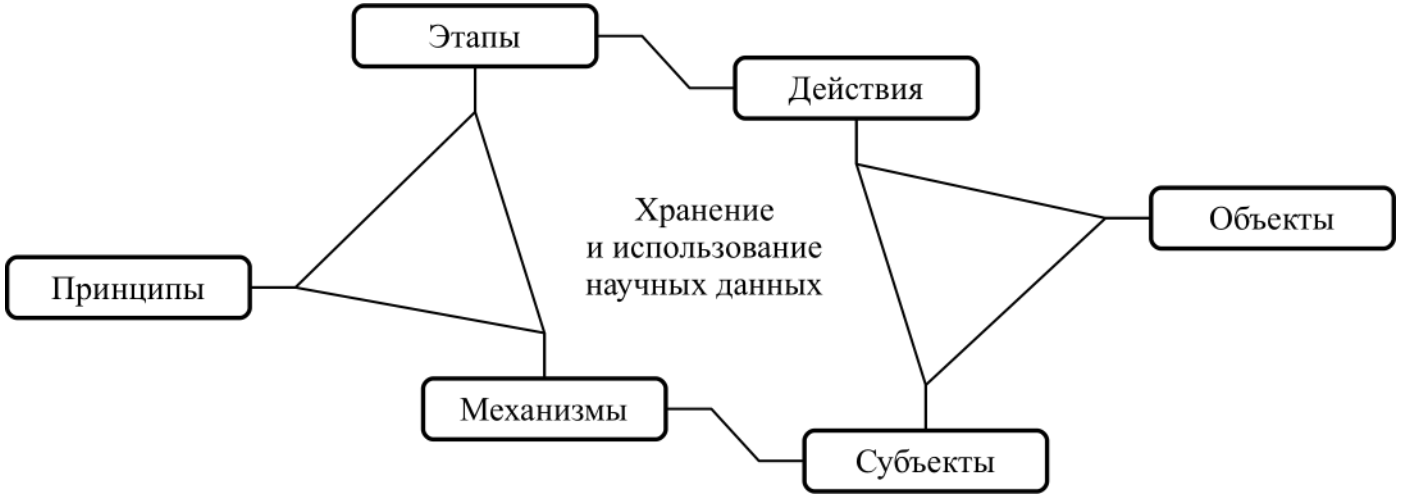
- Научные данные получаются двумя основными путями
  - при проведении измерений и описания объекта исследований
  - при компьютерном моделировании объекта
- Выделив процессы измерения / описания и моделирования мы определяем «зону ответственности» Data Science
- Соответственно разделяются
  - акторы исследовательских процессов
  - методологическая база
  - инструментальная база

Модели: IDEF0 описание исследовательского процесса, основанного на компьютерном моделировании и анализе его результатов

- На этапе компьютерного моделирования и вычислительного эксперимента мы работаем не с самим объектом, а с его компьютерной моделью
- Однако результатом все равно являются данные и описания
- Важно, что независимо от источника научных данных общая схема (модель) основанных на них исследований не меняется

Модели: Иерархия 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

Модели: Иерархия классов-понятий – детализация базовых понятий: принципы

- Принципы организации хранения и использования научных данных:
  - FAIR Data Principles
  - аналог FAIR для методов и алгоритмов
  - Usability@Top – удобство использования превыше всего
  - StoreEverything, IntegrateData, CombineResources – хранение всех видов данных, интеграция данных и комбинирование ресурсов
  - UseEverythingKnown – использовать при работе с данными всю возможную (доступную) информацию
  - FromPrivate2Public – обеспечивать требуемую приватность, мотивируя / стимулируя раскрытие (публикацию) данных
  - EnlightFromAnywhere и OntologizeAll – собирать всю информацию из всех доступных источников для построения и постоянного расширения и уточнения семантической сети вокруг данных

Принципы должны способствовать повышению уровня переиспользования данных, в том числе – постановке и решению с их использованием новых научных задач

Модели: Иерархия классов-понятий – детализация базовых понятий: этапы

- Этапы организации хранения и использования научных данных:
  - Получение
  - Сбор
  - Размещение
  - Проверка качества
  - Описание
  - Типизация
  - Связывание
  - Организация
  - Сохранение
  - Поиск
  - Комбинирование
  - Выбор методов обработки
  - Обработка
  - Интеграция
  - Обеспечение доступа
  - Изучение ограничений на передачу / публикацию
  - Организация обмена
  - Публикация
  - Цитирование
  - Утилизация

Модели: Иерархия классов-понятий – детализация базовых понятий: механизмы

- Механизмы организации хранения и использования научных данных:
  - Методологические
    - Методы и методики
    - Алгоритмы (действий)
    - Технологии
    - Документы: рекомендации, инструкции, лучшие практики
  - Регуляторные
    - Политики
    - Регламенты
    - Правила
  - Организационные
    - Модели процессов
    - Схемы взаимодействия
    - Организационные структуры (институты)
  - Инструментальные
    - Инфраструктура
    - Аппаратные ресурсы
    - Программные платформы
    - Приложения
    - Сервисы

Важно!

- **Каждый принцип должен иметь механизмы реализации**
- **Каждому этапу должен соответствовать хотя бы один механизм**

Модели: Иерархия классов-понятий – детализация базовых понятий: объекты и субъекты

- Объекты в контексте хранения и использования научных данных:
  - Объект исследований
  - Источник данных
  - Объект данных
  - Объект метаданных
  - Коллекция данных
  - Объект типа связь
  - Алгоритм
  - Объект инфраструктуры
  - Аппаратный ресурс
  - Виртуальный ресурс
  - Программная платформа
  - Инсталляция программной платформы
  - Программа для ЭВМ
  - Сервис
- Субъекты в контексте хранения и использования научных данных:
  - Исследователь
    - Исследователь-предметник
    - Исследователь-экспериментатор
    - Исследователь-вычислитель
    - Исследователь данных (Data Scientist или дата-аналитик)
  - Производитель (создатель, автор) объекта данных
  - Владелец объекта данных
  - Пользователь объекта данных
  - Администратор данных
  - Библиотекарь (Data Steward)
  - Владелец ресурса
  - Пользователь ресурса
  - Администратор ресурса
  - Оператор ресурса
  - Регулятор

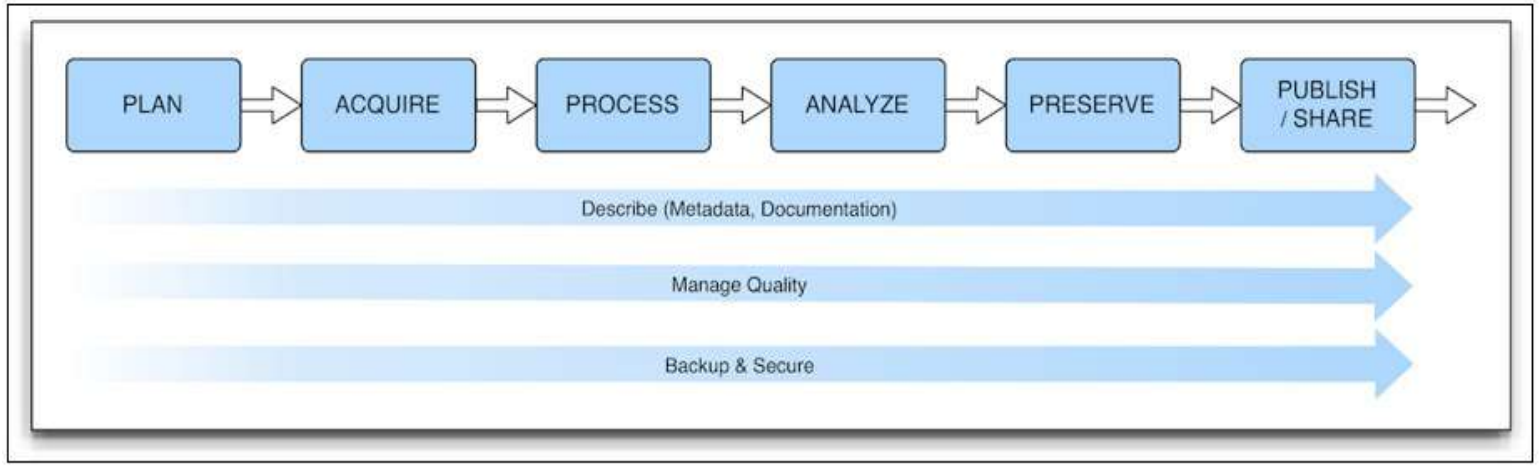
Под **объектом (научных) данных** будем понимать выделенную (фиксированную) конечную последовательность (цифровых) данных, снабженную (связанную с) объектом метаданных, описывающим ее, размещенную на компьютерных ресурсах, имеющую самостоятельную научную ценность. Объект данных – базовая единица хранения данных.

Модели: Иерархия классов-понятий – детализация базовых понятий: действия

- Действия при организации хранения и использования научных данных:
  - Получение данных
  - Сбор данных
  - Формирование и загрузка объекта данных
  - Описание объекта данных
  - Построение связей
  - Организация
  - Верификация данных
  - Оценка объекта данных
  - Ранжирование объектов данных
  - Хранение данных
  - Выгрузка объекта данных
  - Поиск
  - Формирование коллекции данных (интеграция данных)
  - Формирование процесса обработки / анализа (комбинирование ресурсов)
  - Обработка данных
  - Анализ данных
  - Обмен данными
  - Публикация объекта данных
  - Уничтожение объекта данных
- Дополнительные действия, связанные с пользованием данными
  - Техническая поддержка пользователей
  - Сопровождение действий пользователей
  - Обучение пользователей

Модели: Жизненный цикл (научных) данных

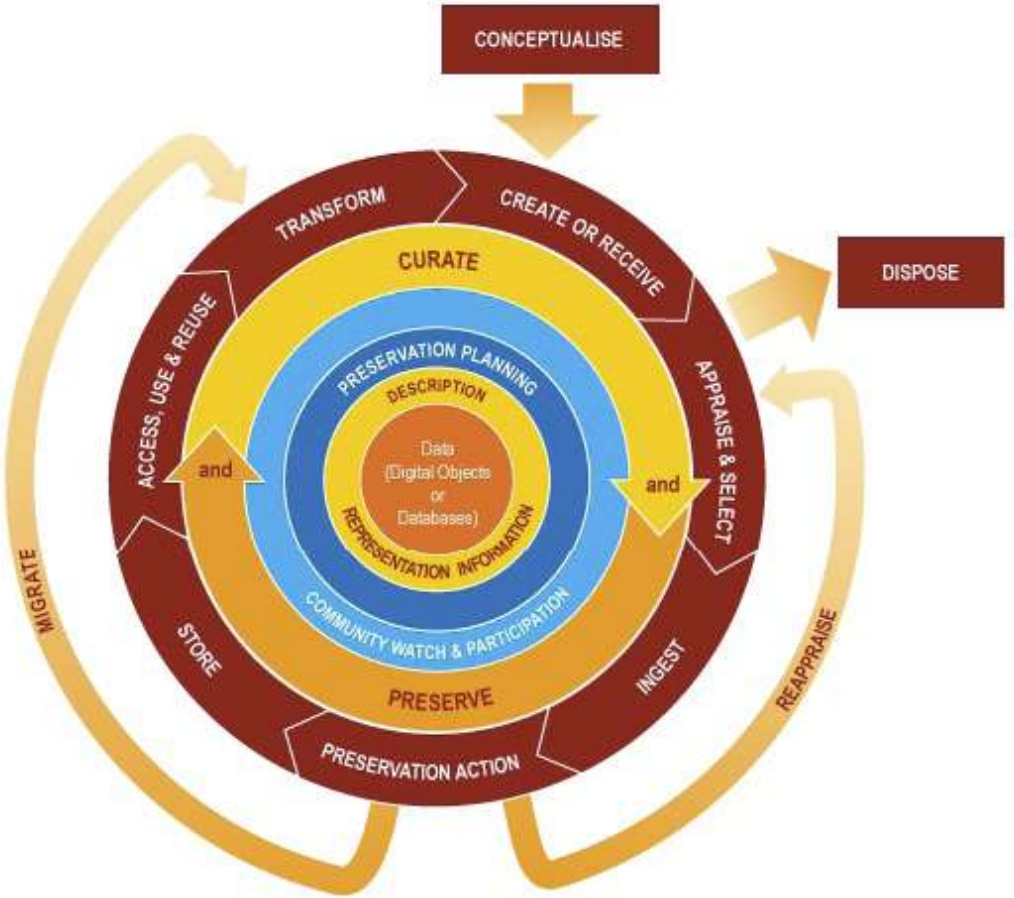
- В мире разработано более 50 моделей жизненного цикла данных, отличающихся набором этапов, линейных или цикличных, последовательных или с параллельными процессами, ориентированных на ученых или библиотекарей, либо учитывающих потребности нескольких сторон, и т.д.



Простая линейная модель жизненного цикла данных с параллельными процессами, принятая USGS (U.S. Geological Survey – Геологическая служба США). В модели выделяется основная линия с детализацией этапов и дополнительные. Не предусмотрено уничтожение данных.

Модели: Жизненный цикл (научных) данных

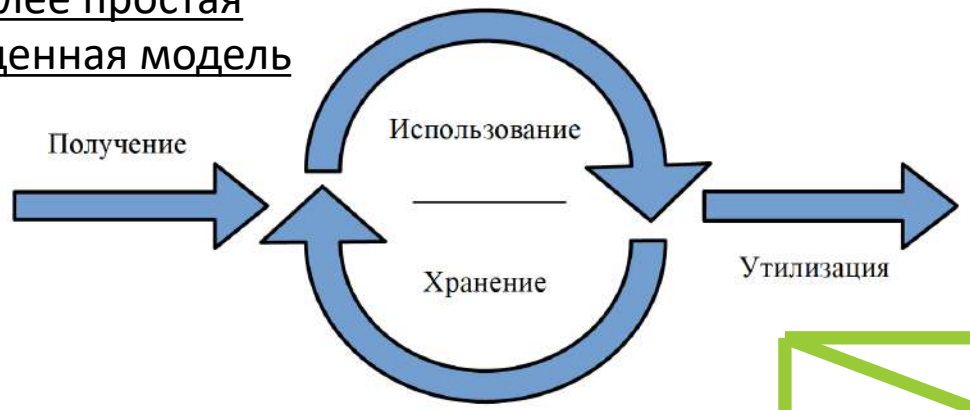
- Большинство моделей жизненного цикла данных в науке разрабатывается «библиотекарями», поэтому концентрируется на близких и понятных библиотекарю задачах каталогизации, сохранения и предоставления данных.



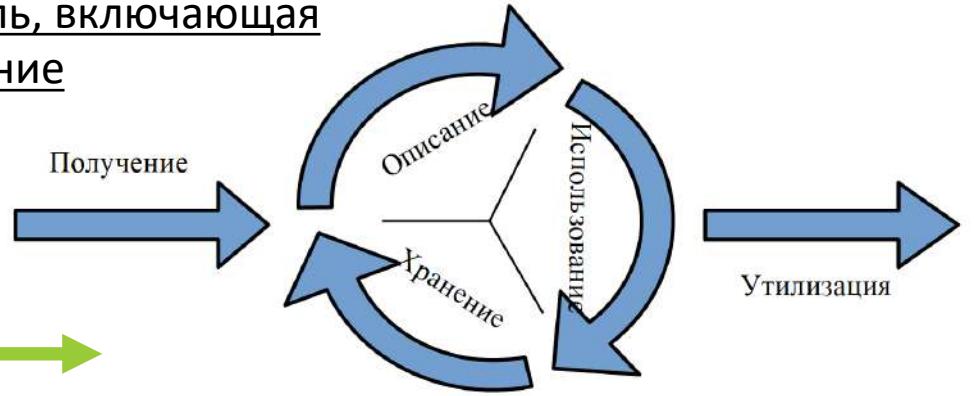
Циклическая модель жизненного цикла данных с параллельными процессами и ветвлениями британского Центра цифрового сопровождения (Digital Curation Centre). Модель ориентирована на «библиотечные задачи», уделяя крайне мало внимания вопросам использования (Access, Use & Reuse)

Модели: Жизненный цикл (научных) данных

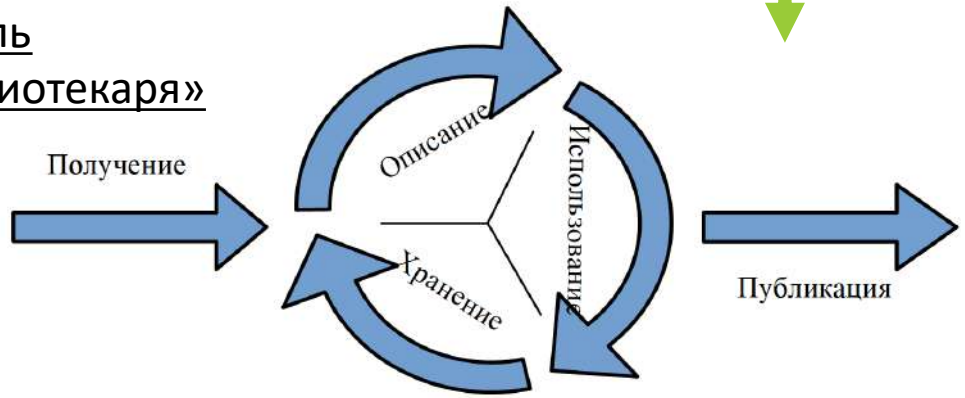
Наиболее простая обобщенная модель



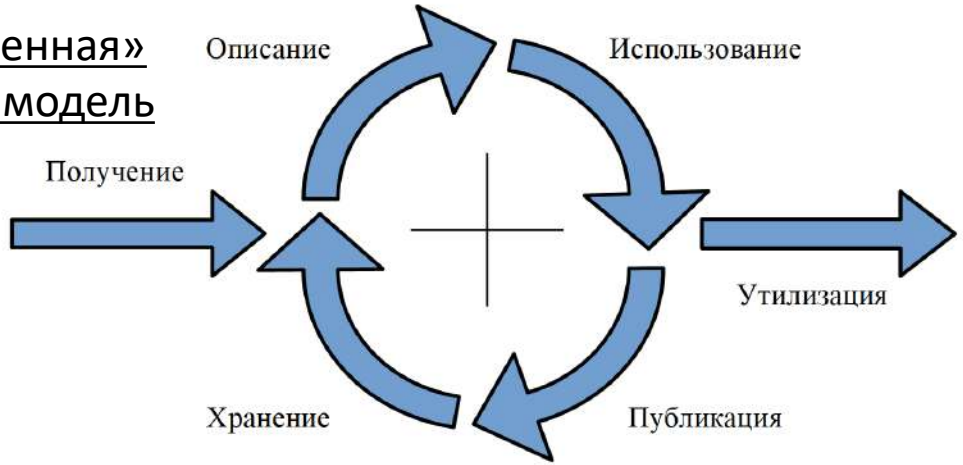
Модель, включающая описание



Модель «библиотекаря»



«Уточненная» общая модель



Модели: Жизненный цикл (научных) данных

- Множественность и существенная разнородность моделей жизненного цикла данных затрудняет выбор при организации деятельности исследователей, работающих с ними
- Преобладание в области управления научными данными библиотечных подходов оставляет в стороне вопросы организации использования данных, особенно – их анализа и обработки
- Решение проблемы моделирования процессов работы с данными и организации их жизненного цикла лежит в плоскости адаптивного конструирования их под задачи конкретных исследовательских групп на основе анализа потребностей и целей, с использованием расширенной иерархии понятий предметной области – это работа специалиста нового типа – Data Steward

Методы решения задач и технологии организации хранения и использования научных данных

Методы: О реализации FAIR Data Principles

- F – Findable:
  - ассоциировать с данными глобально уникальный и устойчивый идентификатор (такой, как DOI);
  - снабжать данные богатыми метаданными / описанием;
  - устанавливать четкую связь между метаданными и данными, включая в метаданные идентификатор данных;
  - размещать метаданные на индексируемых ресурсах с возможностью поиска.
- A – Accessible:
  - и данные, и метаданные должны быть достижимы и доступны для получения по стандартизованным (открытым, свободным и универсальным, поддерживающим аутентификацию и авторизацию) протоколам
  - метаданные должны быть доступны даже если сами данные уже недоступны.

Методы: О реализации FAIR Data Principles

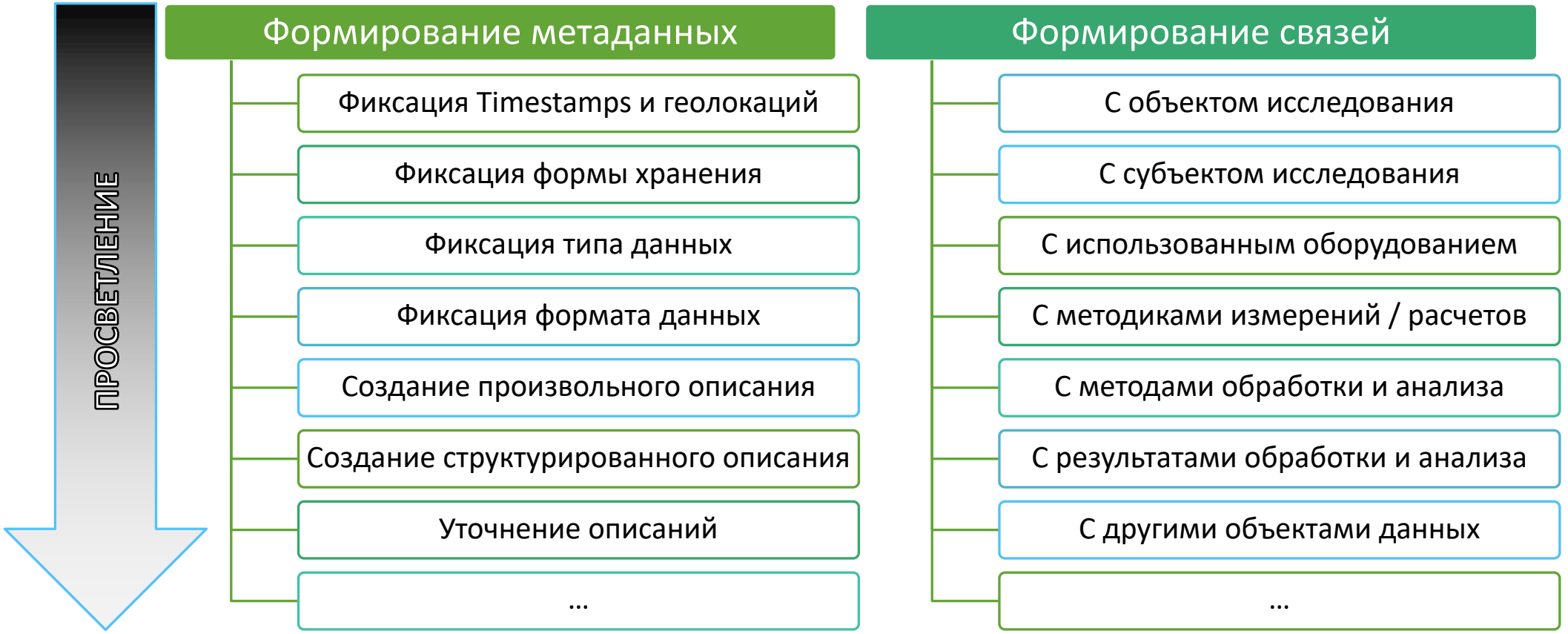
- I – Interoperable:
  - использовать достаточно формализованный, общедоступный и общеизвестный, широко используемый язык для описания и представления данных и метаданных;
  - использовать словари, соответствующие принципам FAIR;
  - включать в данные и метаданные «квалифицированные» ссылки на другие данные и метаданные, т.е. такие ссылки, которые будут формализовать вид связи, а не просто указывать на ее наличие.
- R – Reusable: качественное описание и атрибутирование данных, в частности
  - выпускать данные в сопровождении понятной и доступной лицензии на использование;
  - подробно описывать (или указывать) происхождение данных;
  - соответствовать стандартам и рекомендациям сообщества в соответствующей области знаний.

Методы: Концепция ***просветления***

- Неполнота информации – обычная ситуация
- Объект данных возникает как «черный ящик»
- Чем сложнее объект данных, тем сложнее его описать
- Типичное состояние объекта данных – «серый ящик»
- Идеальное состояние «белого ящика» практически недостижимо

Просветление – процесс обрастания объекта данных дополнительной информацией, способствующей лучшему его пониманию и облегчающей его анализ

Методы: Пути ***просветления***



Методы: Источники и методы *просветления*

- Активные:
  - Автор объекта данных
  - Пользователи объекта данных
  - «Управляющие» данными (библиотекари данных)
  - …

- Ручное и автоматизированное внесение информации
- Экспертная верификация информации

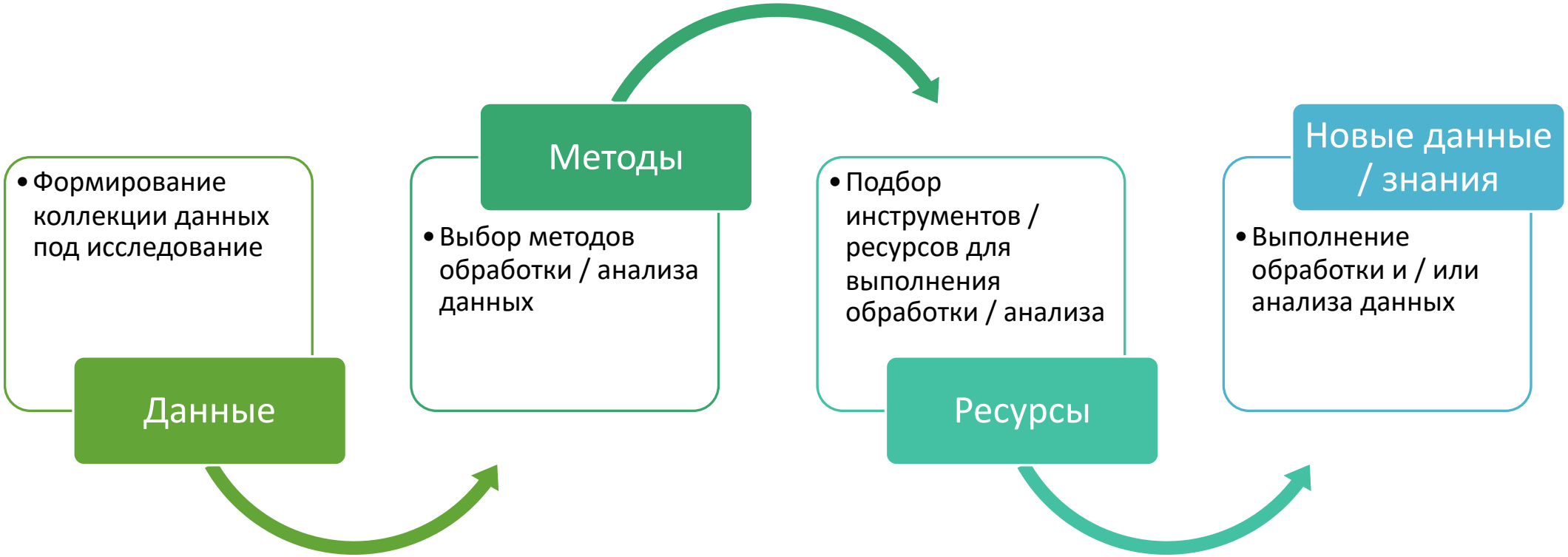
- Пассивные:
  - Информация (метаданные) от приборов
  - Действия пользователя с объектом данных
  - Содержимое системы
  - Содержимое других научных баз данных и знаний
  - Прочее содержимое Интернет
  - …

- Отслеживание и анализ действий
- Анализ содержимого
- Автоматическая верификация
- Автоматический поиск

Методы: Результат ***просветления***

- Просветление – непрерывный процесс, одним из его результатов должно быть сопровождение объекта данных *описанием*. Описание может храниться как отдельный объект, тогда исходный объект и его описание должны содержать взаимные ссылки друг на друга, либо описание может храниться вместе с объектом. Это реализует принцип Accessible.
- Важно, чтобы *описание* было опубликовано и доступно «всем нуждающимся». Информацию о нем можно опубликовать в открытых источниках. Описание объекта данных, тогда, может быть использовано различными системами-агрегаторами для формирования собственных баз данных с информацией об объектах данных, например, специализированных по областям наук, которые уже, в свою очередь, могут обеспечивать реализацию принципа Findable.

Методы: Использование данных

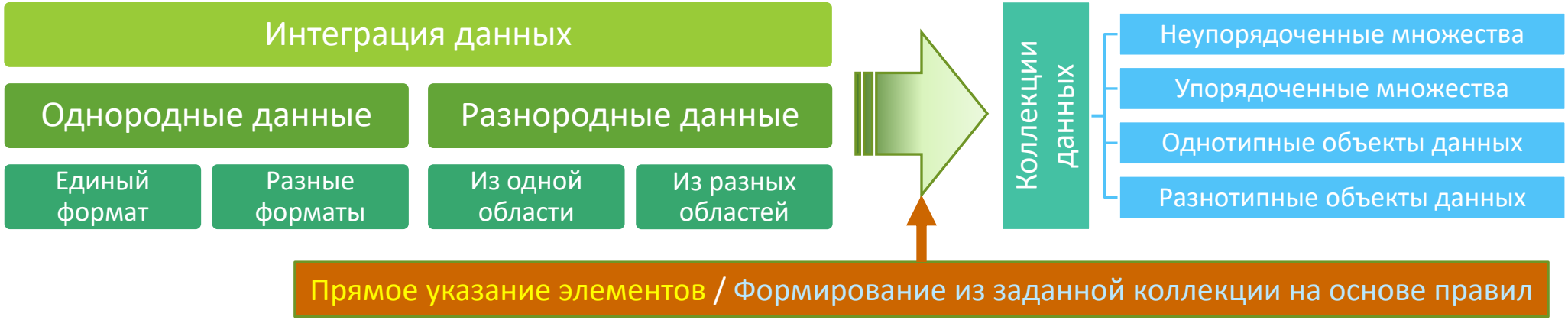


Методы: Интеграция данных

- **Данных не бывает много. Обычно данных не хватает**
- Расширение наборов данных, используемых в исследовании, может дать дополнительные возможности для
  - повышения точности и надежности, в том числе – валидности исследований
  - расширения класса исследовательских задач, которые можно решить с их помощью
- Объединение данных из разных источников в коллекции повышает их переиспользуемость, позволяет решать новые задачи

Методы: Интеграция данных

- **Интеграция данных – процесс формирования коллекций данных с заданными свойствами**



- Эффективность формирования коллекции кроме квалификации специалиста напрямую зависит от того, насколько просветленные объекты данных в реестре
- Особый интерес представляет работа с коллекциями, включающими данные из различных областей – это позволяет решать задачи анализа данных и построения прогнозов на их основе на новом уровне

Методы: Интеграция данных – примеры

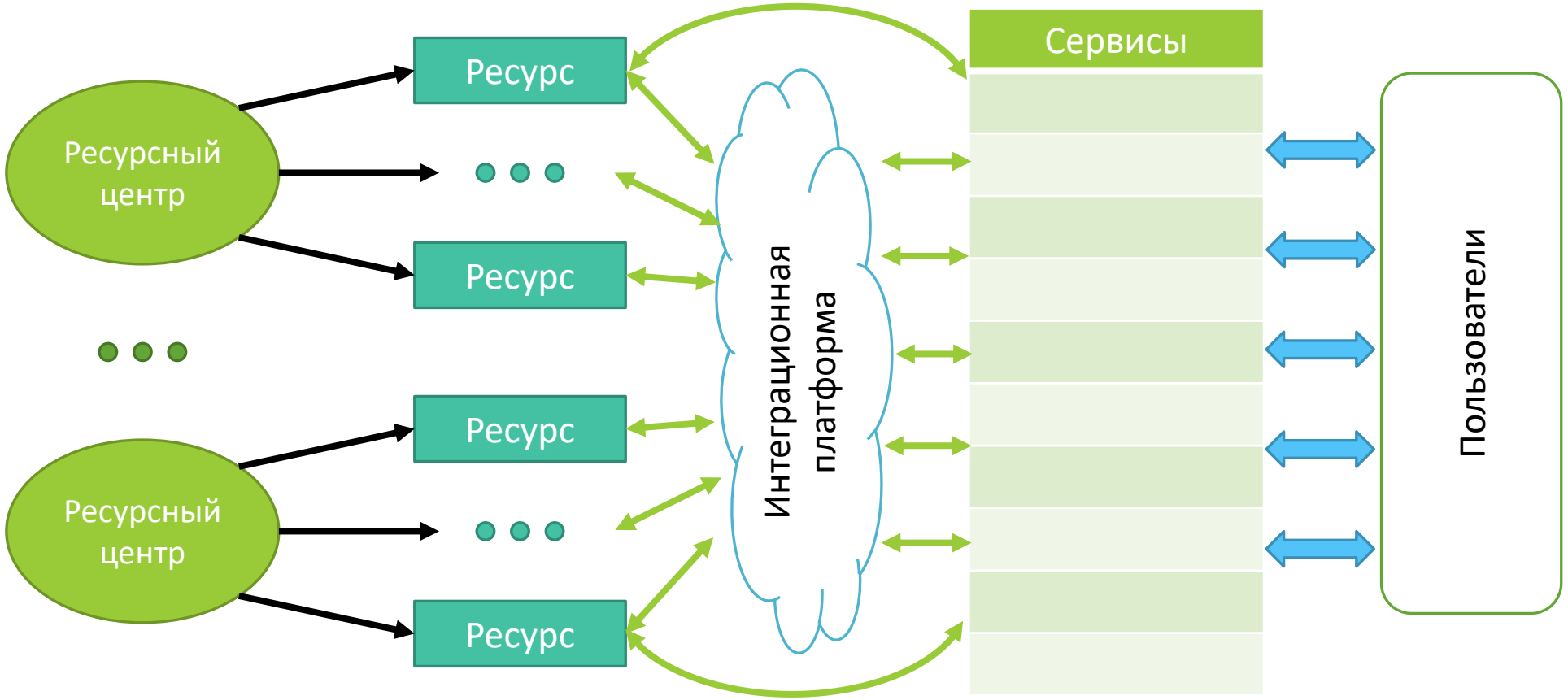
- Извлечение географической информации из негеографических документов
  - Произвольные документы, содержащие упоминания географических объектов + Пространственные географические данные и карты, в том числе с ретроспективой -> Комплекс связей между документами и географическими объектами
- Изучение территориальных рисков здоровья населения
  - Актуальные и ретроспективные пространственные данные дистанционного зондирования Земли из космоса, в том числе по аэрозолям + Пространственные данные наземных наблюдений + Пространственные данные медицинской статистики -> Корреляции между состоянием здоровья и антропогенным воздействием -> Экологические и медицинские рекомендации
- Оценка рисков лесных пожаров
  - Актуальные и ретроспективные пространственные данные дистанционного зондирования Земли из космоса о термоаномалиях + Актуальные и ретроспективные пространственные данные дистанционного зондирования Земли из космоса об аэрозолях + Актуальные и ретроспективные метеоданные + Данные метеопрогноза + Данные с наземных станций и наблюдений -> Прогнозы и оценка рисков возникновения и распространения лесных пожаров

Инфраструктурные и информационно-технологические решения организации хранения и использования научных данных

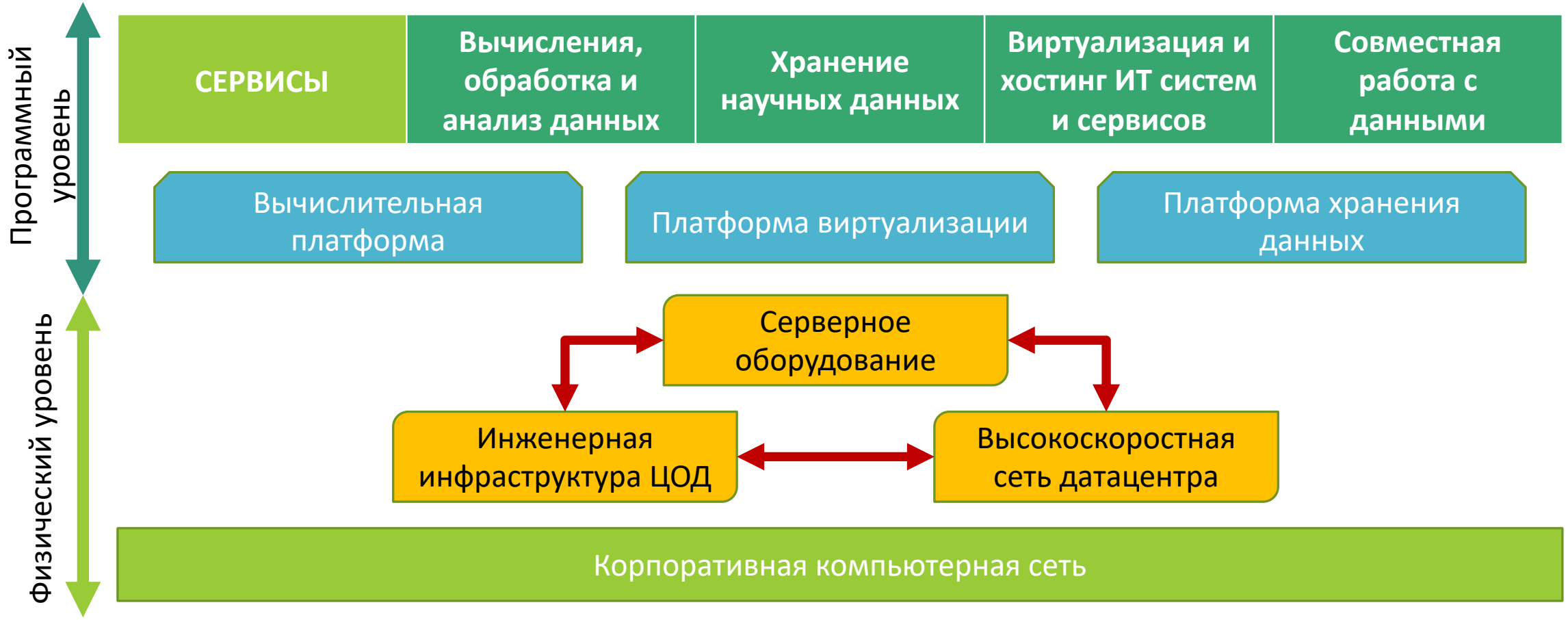
Решения: Сервисы – как форма доступа к ресурсам

- Сервис – это форма обеспечения потребителя без передачи материального продукта или средств реализации (производства)
- Научным ИТ-сервисом называем реализованную в программном и аппаратном виде технологию работы с научными данными, предоставляемую, соответственно, без передачи самих программных и аппаратных средств
- Особенностью научных ИТ-сервисов является
  - Опора на значительные ИТ-ресурсы: суперкомпьютеры, системы хранения и обработки больших данных, мегасайенс-установки в качестве источников цифровых данных, …
  - Постоянная разработка, совершенствование, адаптация и применение новейших методов и алгоритмов обработки и анализа данных всеми участниками исследовательских процессов
  - Отсутствие ориентации на извлечение прибыли
- Высокая неравномерность потребностей в ИТ-ресурсах для отдельной исследовательской группы в совокупности с большим числом таких групп и часто неподъемной для них стоимостью научных ИТ-ресурсов заставили искать пути организации совместного использования этих ресурсов, реализацию которых целесообразно осуществлять в форме соответствующих услуг – сервисов.

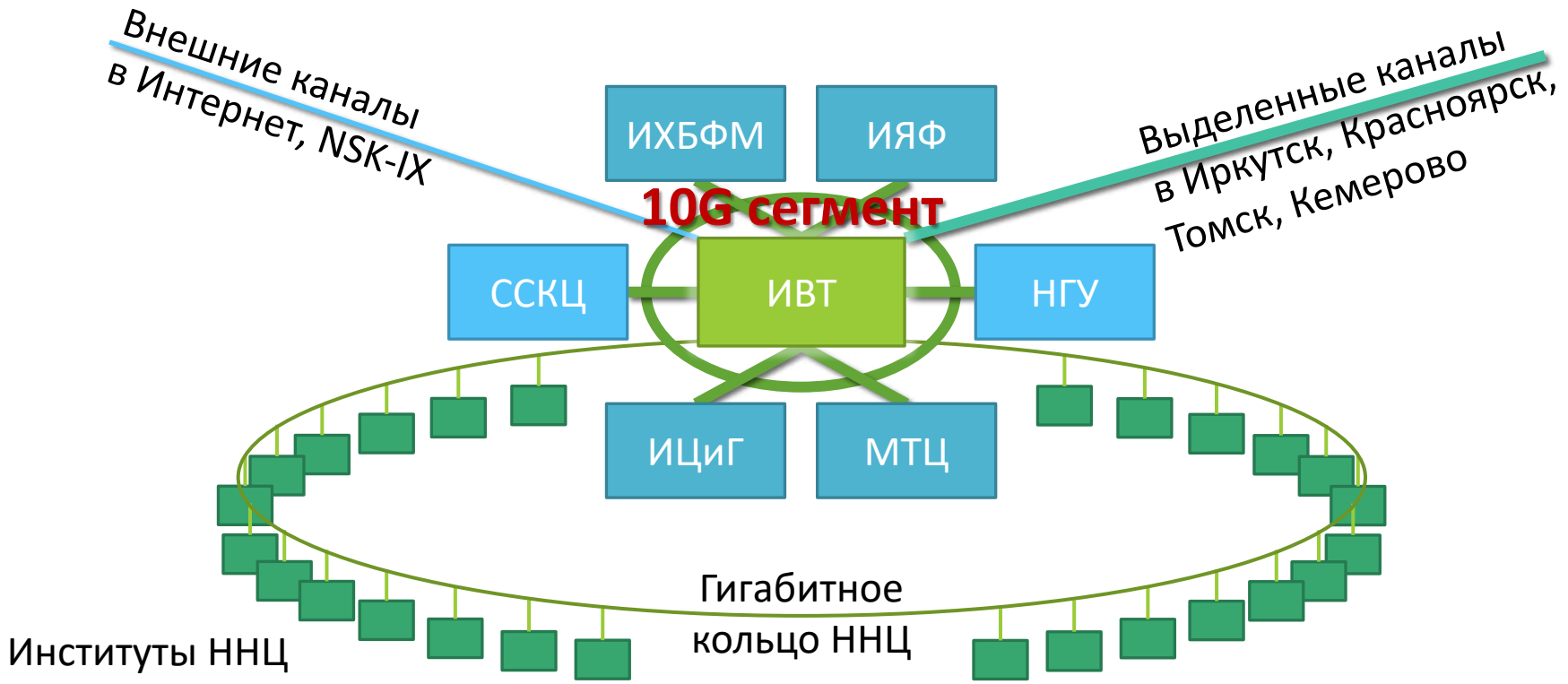
Решения: Ресурсные центры – как основные единицы инструментального обеспечения



Решения: концепция организации Центра научных ИТ-сервисов

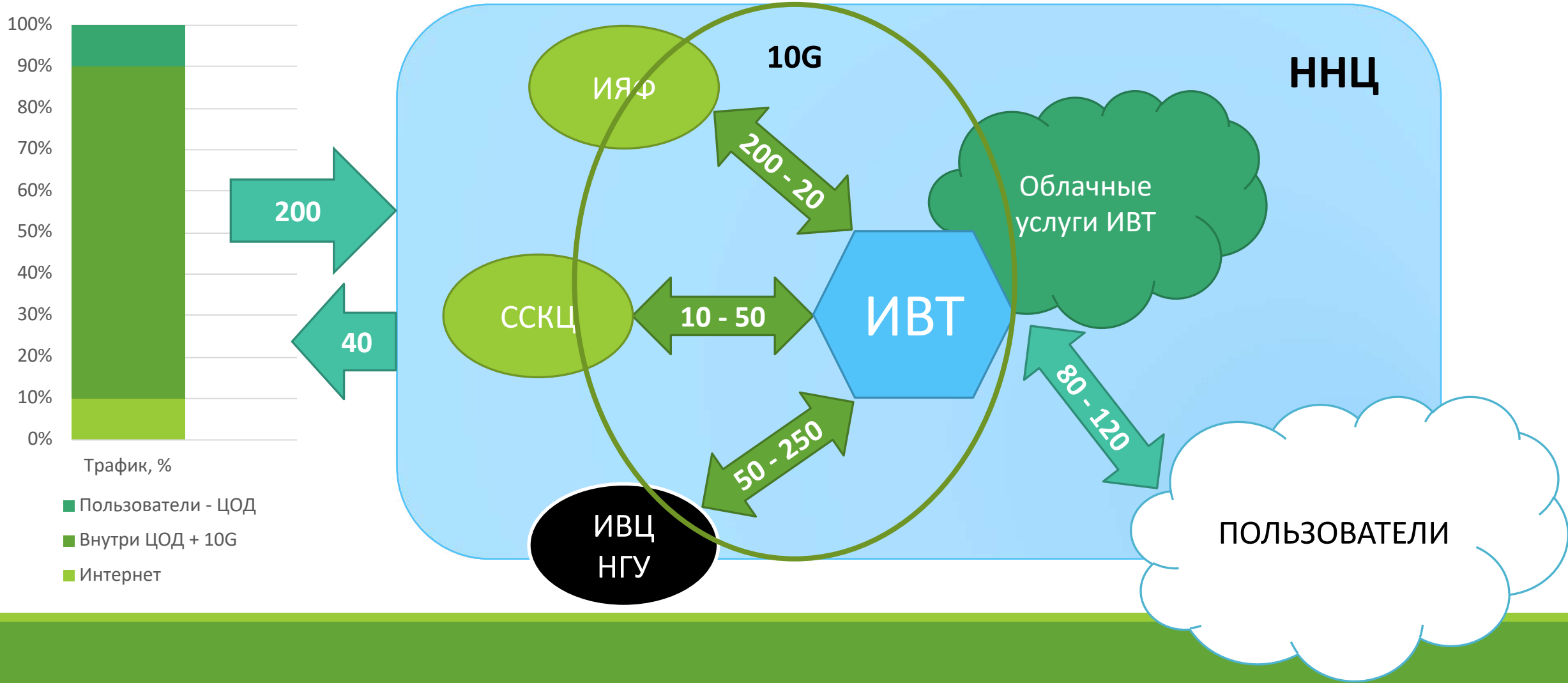


Структура корпоративной компьютерной сети Новосибирского научного центра



Структура сетевого трафика ННЦ

объем месячного трафика в ТБ (прием - отдача)



Центр научных ИТ-сервисов ФИЦ ИВТ: аппаратные ресурсы и инженерная инфраструктура

- Два больших (по 15-20 стоек) и два маленьких (4 и 7 стоек) серверных зала с системами воздушного охлаждения и газового пожаротушения.
- Электропитание с двух независимых вводов. Аккумуляторные системы бесперебойного питания для телекоммуникационного и серверного оборудования.
- Вычислительные кластеры для обработки и анализа данных, в том числе гибридный **GPU-TPU-кластер** для задач машинного и глубокого обучения, нейросетевого и высокопроизводительного моделирования (производительность FP64 – до 96 Терафлопс, тензорная – до 1,44 Петафлопс).
- Кластеры виртуализации для размещения информационных систем, научных каталогов и вычислительных сервисов.
- Системы отказоустойчивого хранения данных, суммарной доступной пользователям емкостью более **3 Петабайт**.

Решения: Ресурсы Центра научных ИТ-сервисов ФИЦ ИВТ

- Корпоративная компьютерная сеть Новосибирского научного центра
- Комплекс небольших датацентров
- Сеть уровня датацентра
- Распределенный комплекс для долговременного хранения и организации доступа к исследовательским данным больших объемов
- Распределенный вычислительный комплекс для обработки больших объемов данных, в том числе в режиме реального времени
- Кластеры виртуализации
- Тестовый кластер для разработки и отработки решений по хранению и обработке данных, тестированию информационных систем и ИТ-сервисов

Решения: Сервисы Центра научных ИТ-сервисов ФИЦ ИВТ

- Сервис хранения исследовательских данных: от Терабайтов до Петабайтов
- Виртуализация в отказоустойчивых кластерных конфигурациях
- Хостинг научных информационных систем и ресурсов (в том числе веб-порталов и веб-сайтов организаций, баз данных и знаний) на готовых платформах
- Сервис организации совместной работы с научными данными и документами
- Выделение вычислительных ресурсов
- Комплекс научно-организационных сервисов
  - Сервис ИТ-подержки организации и проведения конференций
  - Формирование каталогов публикаций сотрудников
  - Электронные системы поддержки издания журналов
- Сервисы обработки данных ДЗЗ
- Биологические и биомедицинские ИТ-сервисы

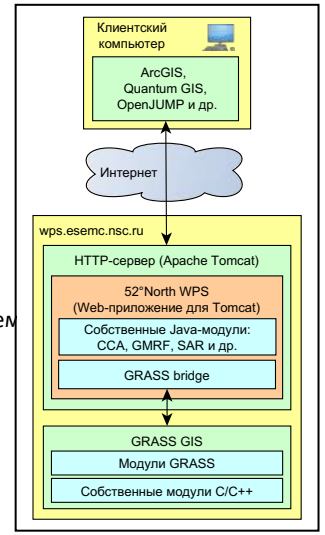
Примеры ИТ-сервисов и их использования

Сервисы обработки данных ДЗЗ

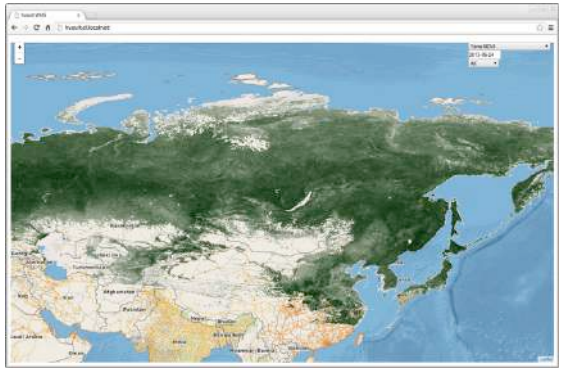
Система WPS-сервисов обработки изображений

http://wps.esemc.nsc.ru:8080/wps

- Включает алгоритмы обработки спутниковых изображений, разработанных в ФИЦ ИВТ (выбор информативных признаков, сегментация на основе обучаемой и необучаемой классификации данных и др.)
- Интегрирована с GRASS GIS
- Позволяет решать практические задачи на стороне клиента, в том числе с использованием ArcGIS
- Разработана на основе программных продуктов с открытым исходным кодом
- Ведутся работы по созданию веб-ориентированного клиентского приложения на базе открытой библиотеки OpenLayers

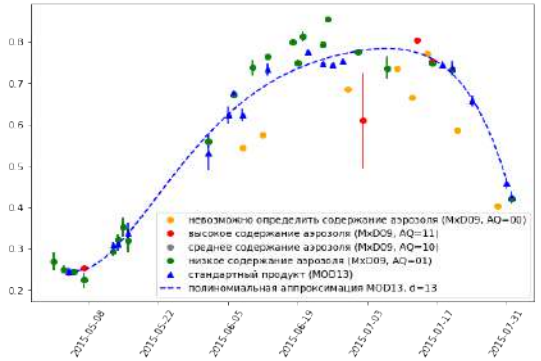


Результаты использования WMS сервиса для построения картосхем на основе данных дистанционного зондирования с функциями агрегации

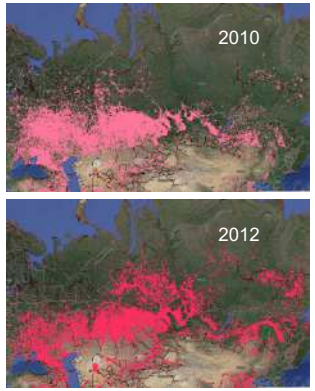


Расчет средних значений вегетационного индекса NDVI

Временной ход индекса NDVI, для отдельного поля пшеницы



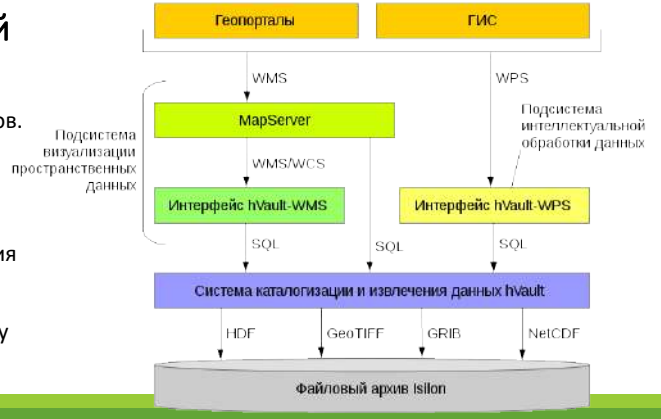
Отслеживание развития отдельных пожаров (пожар вблизи поселка Тура, Красноярского края, июль 2013)



Исследование особенностей отдельных пожарных сезонов, близость природных пожаров к объектам транспортной инфраструктуры

Активная система сбора, хранения и оперативной обработки данных ДЗЗ

- Основана на пополняемом архиве данных ДЗЗ, получаемых с различных сенсоров. История архива составляет более 10 лет.
- Архив хранится на отказоустойчивой высоконадежной СХД, емкость 500 Тбайт.
- Содержит инструменты поддержки каталогизации, быстрого поиска и извлечения как исходных данных ДЗЗ, так и тематических продуктов на их основе.
- Предоставляет внешним ГИС-системам и геопорталам интерфейсы как к каталогу данных ДЗЗ и тематических продуктов, так и к подсистема обработки данных.

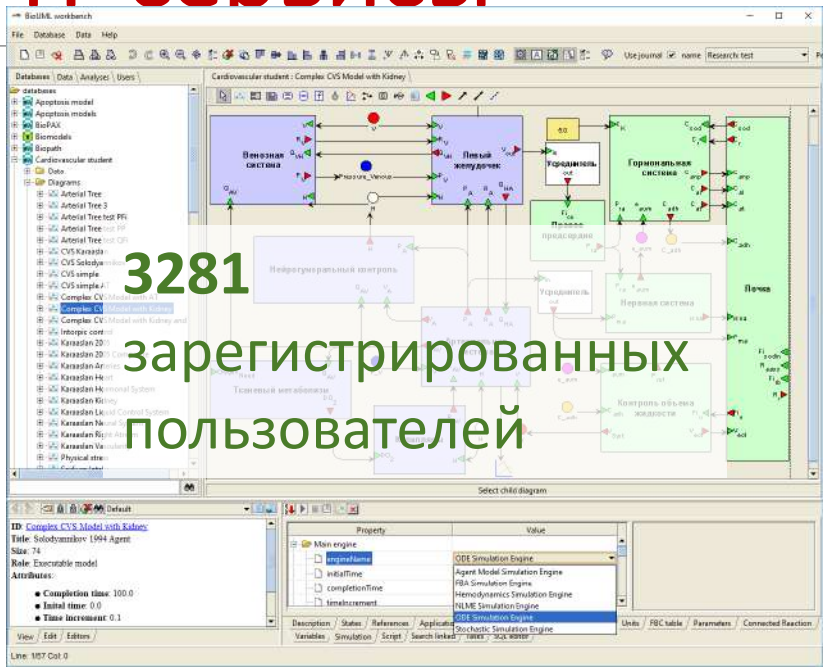


Примеры ИТ-сервисов и их использования

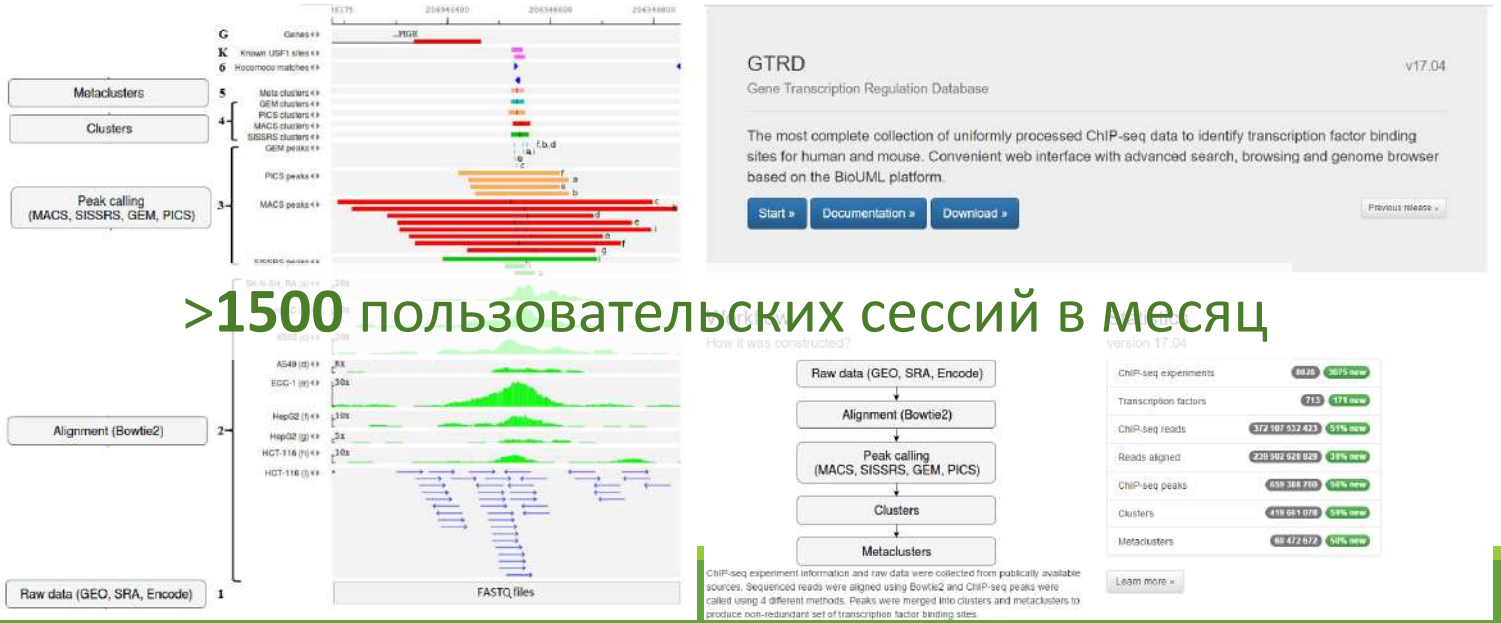
Биологические и медицинские ИТ-сервисы

BioUML – платформа для модульного моделирования сложных биологических систем и анализа биомедицинских данных

- Построены модульные модели апоптоза, сердечно-сосудистой системы и лечения гипертонии, множественной миеломы и ее лечения
- Анализ данных: 100+ готовых сценариев; 1000+ методов анализа; интеграция с R/Bioconductor и Galaxy.



3281 зарегистрированных пользователей



>1500 пользовательских сессий в месяц

GTRD: база данных по сайтам связывания транскрипционных факторов, идентифицированных при помощи ChIP-seq экспериментов

Центр научных ИТ-сервисов ФИЦ ИВТ: программные решения и сервисы

- Выделение пространства для размещения исследовательских данных любого вида
  - Подключение в виде «внешнего диска» по файловым или блочным протоколам
  - Облачное хранилище с возможностью обмена и публикации, совместной работы с документами, общения через чаты и аудио-, видеосвязь
- Выделение вычислительных ресурсов в форме доступа к системе управления заданиями
  - Вычисления, анализ данных, машинное и глубокое обучение на узлах с GPU: доступны 3 узла с характеристиками 2xCPU Intel Xeon Gold 6136, 384 MB RAM, 4xGPU-TPU NVidia Volta V100 32GB with NVLink
  - Вычисления на узлах предыдущих поколений: Opteron O2435 и Tesla C2070
- Размещение информационных систем, в том числе веб-сайтов и веб-порталов
  - Выделение виртуальных серверов различной конфигурации
  - Подготовка и предоставление готовых платформ (LAMP, LAMP+Drupal и др.) для размещения ИС

http://sits.ict.sc

Заключение

Заключение

- Описана проблемная область организации хранения и использования научных данных
- Произведена формализация задачи системного анализа и управления процессами работы с научными данными, осуществлена ее декомпозиция на методологический, организационный и инструментальный аспекты
- Построены модели исследовательских процессов, основанных на данных, иерархия классов-понятий для онтологии проблемной области, проведен сравнительный анализ ряда моделей жизненного цикла научных данных и управления ими
- Сформулирована концепция просветления объектов научных данных и предложены способы ее реализации
- Описана система методов работы с научными данными, позволяющая повысить эффективность их использования
- Представлена идеология сервисного обслуживания исследователей, использующих научные данные, для обеспечения их необходимыми ИТ-ресурсами, создан в ФИЦ ИВТ соответствующий ресурсный центр

Спасибо

ЮРЧЕНКО АНДРЕЙ ВАСИЛЬЕВИЧ, ФИЦ ИВТ, YURCHENKO@ICT.NSC.RU