УДК 005, 004.62
doi:10.31799/1684-8853-2019-3-45-54

Научные статьи
Articles

О моделях организации хранения и использования научных данных: основные принципы, процессы и механизмы

Ю. И. Шокин[a]***,*** *доктор физ.-мат. наук, академик РАН, orcid.org/0000-0002-5178-8294, dir@ict.nsc.ru*
А. В. Юрченко[a]***,*** *канд. физ.-мат. наук, первый заместитель директора, orcid.org/0000-0001-6435-1975, yurchenko@ict.nsc.ru*
[a]*Институт вычислительных технологий СО РАН, Академика Лаврентьева пр., 6, Новосибирск, 630090, РФ*

Введение: *проблемы организации хранения и использования научных данных усложняются с ростом их количества и разнообразия, при этом научные данные обладают рядом особенностей, не позволяющих полностью переносить на них подходы и инструменты, используемые в коммерческих и государственных структурах, работающих с данными. Обеспечение исследователей специализированными средствами для оперирования с данными — актуальная задача организации научных исследований.* ***Цель:*** *выявление и описание основных принципов работы с научными данными, процессов и этапов этой работы, механизмов реализации принципов и решений задач организации хранения и использования научных данных.* ***Результаты:*** *рассмотрены и описаны принципы, на которых может основываться хранение и использование научных данных, в том числе FAIR Data Principles. Основная цель организации работы с научными данными и центральный фокус принципов — эффективное использование и переиспользование научных данных. Представлена иерархия механизмов, которые могут применяться при работе с научными данными для решения научных и организационных задач. Перечислены основные процессы / этапы жизненных циклов научных данных и процессов исследований, основанных на них. Рассмотрен ряд принятых моделей таких жизненных циклов. Предлагается вместо попытки построить универсальную модель использовать или создавать модели на основе представленного списка этапов под конкретные случаи или классы задач исследований, основанных на данных.* ***Практическая значимость:*** *сформированная в работе иерархия классов понятий для области «организация хранения и использования научных данных» будет использована как ядро соответствующей онтологии и при разработке нормативных документов, рекомендаций и информационных систем поддержки научных исследований, основанных на данных.*

Ключевые слова *— научные данные, основанные на данных научные исследования, FAIR Data Principles, управление данными, жизненный цикл научных данных, научная информационная система.*

Для цитирования: Шокин Ю. И., Юрченко А. В. О моделях организации хранения и использования научных данных: основные принципы, процессы и механизмы. *Информационно-управляющие системы*, 2019, № 3, с. 45–54. doi:10.31799/1684-8853-2019-3-45-54
For citation: Shokin Yu. I., Yurchenko A. V. Models of organizing research data storage and usage: basic principles, processes and implementation mechanisms. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2019, no. 3, pp. 45–54 (In Russian). doi:10.31799/1684-8853-2019-3-45-54

Введение

Начнем с того, что под *научными данными* будем понимать любые данные в цифровом виде, генерируемые в ходе научных исследований и (или) используемые в них. Настоящая работа является продолжением исследований авторов в области организации и обеспечения технологиями и средствами научных исследований, основанных на использовании данных и цифровизации науки в целом.

Научные данные стали одним из важнейших источников получения новых знаний. Объемы этих данных постоянно растут. Возникают и растут также потребности в организации их хранения и использования. Эта проблема активно обсуждается на разных уровнях международного научного сообщества уже более 10 лет [1], но каких-то единых решений не найдено и, вероятно, не будет найдено. В России это направление разрабатывается не так активно, отчасти ввиду того, что российские исследователи долгое время пользовались устаревшей материально-технической базой, состоявшей в основном из аналогового оборудования. Но примерно с 2006 года материальную базу российских исследований начали активно обновлять, и в настоящее время она состоит уже преимущественно из цифровых приборов, т. е. приборов, выдающих результаты в виде цифровых данных. Соответственно, активизировалась и деятельность по организации работы с этими данными. Эту тему поднимает обзор [2], в котором ставится проблема управления научными данными и участия в ней научно-технических библиотек, проводится ее анализ и рассмотрены организационные подходы к ее решению, используемые в международном сообществе.

Активное развитие междисциплинарных исследований, методов работы с большими данными, примеры решения при этом совершенно новых либо представлявшихся ранее практически нерешаемыми задач с помощью анализа больших

данных (например, компьютерной диагностики рака [3]) или интеграции данных, происходящих из разных областей наук (например, географических, метеорологических, экологических и медицинских в задачах анализа заболеваемости [4]) заставляют искать пути и создавать инструменты для интеграции разнородных научных данных из различных источников. Актуальность проблемы повышают и не до конца раскрытые и использованные перспективы извлечения именно из разнородных научных данных новых неявных закономерностей либо выявления артефактов, которые могут дать новые представления об окружающем нас мире и нашем месте в нем. Задачи интеграции, которые выводят проблему организации работы с научными данными на новый уровень, решить одномоментно не представляется возможным. Необходимо применить итерационный подход к обеспечению исследователей инструментами для такой интеграции данных.

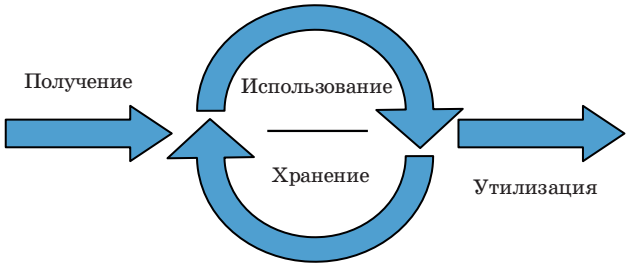
В работе [5] авторами предложена концепция построения информационной системы для поддержки исследований, основанных на данных, в работе [6] идеи развивались в направлении разработки архитектуры и проработки конкретных программных решений для такой системы.

Здесь будет уточнен и расширен понятийный аппарат и исследованы модели организации хранения и использования научных данных, в том числе на основе анализа опыта зарубежных коллег.

Особенности научных данных и задача моделирования их жизненного цикла

Рассматривая в качестве миссии своих исследований и разработок обеспечение исследователей средствами для работы с научными данными, мы двигаемся к достижению возникающих в рамках этой миссии целей с двух сторон: с одной стороны, практически реализуя и предоставляя исследователям конкретные инструменты [6, 7], с другой стороны, обобщая имеющийся опыт и накопленные знания в предметной области для того, чтобы создавать теоретический базис этой деятельности [5, 8, 9]. Цель настоящей работы состоит в построении концептов верхнего уровня деятельности по организации работы с научными данными.

На работу с научными данными распространяются основные подходы, используемые при работе с любыми другими данными, и, так или иначе, она вращается вокруг циклов «получение — хранение — использование — утилизация» (рис. 1). Однако научные данные и работа с ними обладают рядом особенностей, совокупность которых и заставляет формулировать подходы и строить системы для организации их хранения и использования, несколько отличные от принятых в других отраслях.



■ ***Рис. 1.*** Максимально упрощенная концептуальная схема работы с данными

■ ***Fig. 1.*** Simple conceptual scheme of common data lifecycle

Основные особенности данных.

1. *Множественность источников данных.* Каждая современная исследовательская единица в любой стране: единичный ученый, группа, лаборатория, институт и т. д. — генерируют научные данные.

2. *Неоднородность данных и их форматов.* Эта особенность является, в том числе, следствием предыдущей. Разнообразие источников данных, наличие множества областей знаний, при проведении исследований в которых генерируются и используются научные данные, индивидуальные особенности исследователей и используемого ими оборудования порождают многообразие видов научных данных и их форматов.

3. *Разное качество данных* с не всегда измеримыми и формализуемыми критериями и признаками — также одно из следствий множественности источников. Причиной может являться как неполное следование принятым методикам и стандартам измерений, так и их отсутствие, особенно в новых областях, где ведется активный научный поиск, и часто не фиксированы даже сами измеряемые характеристики. Конечно, немаловажен и человеческий фактор.

4. *Большие объемы данных* и при этом сложность или даже невозможность оценить перспективы их использования в дальнейшем и, тем более, получения с их помощью новых значимых научных результатов, особенно с учетом проблемы, обозначенной в предыдущем пункте.

Особенности работы с научными данными и происхождение этих особенностей.

1. *Необходимость обмениваться и делиться данными* — один из ключевых факторов, связанный с особенностями деятельности научного сообщества, его открытостью, необходимостью верификации результатов исследований как важнейшего процесса в рамках научного метода.

2. *Разнообразие и постоянное развитие методов и средств для анализа данных.* Методы ана-

лиза данных постоянно эволюционируют, исследователи обладают различными компетенциями, методами и средствами анализа данных, что означает, что из одних и тех же данных могут быть извлечены разные знания.

3. *Потребность в интеграции разнородных данных*. Нарастающая тенденция к объединению в единый объект исследования разнородных данных из разнородных источников, нетривиальные формы связанности данных (через предмет, объект, субъект исследований, через близость в пространстве / времени, по другим критериям), как следствие — необходимость одновременного анализа комплексов данных разных типов и происхождения.

4. *Необходимость использовать высокопроизводительные ресурсы*. Во многом — это следствие остальных перечисленных особенностей. Именно в науке с большими объемами и высокой интенсивностью генерации новых данных, со множеством неопределенностей, неясностей в отношении методов и средств решения задач (которые также выбираются в процессе научного поиска), в отношении источников данных (их нужно создавать, интегрировать, совершенствовать методики измерений / получения данных и т. д.) и, конечно, с потребностью совмещать сложные задачи анализа данных с не менее сложными и ресурсоемкими задачами компьютерного моделирования изучаемых процессов высокопроизводительные компьютерные ресурсы — важнейший инструмент научных исследований, основанных на данных.

В различных сочетаниях все перечисленные особенности могут быть характерны и для бизнес-данных, и для государственных данных, но полный комплекс свойственен именно научным данным.

Разработка модели организации хранения и использования (включая обмен и публикацию) научных данных, учитывающей перечисленные и другие особенности этих данных и работы с ними — цель настоящей работы и исследования в целом. Здесь определим ключевые сущности процесса работы с научными данными и виды их взаимодействия, рассмотрим различные модели жизненного цикла научных данных и основанных на них исследований, опишем основные процессы в рамках этих моделей, базовые принципы организации и управления научными данными, механизмы их реализации.

Базовые классы понятий

Основной сущностью в рамках модели является понятие «научные данные», определение которого сформулировано в начале статьи. Но прежде чем перечислять связанные с ними сущности и устанавливать их отношения, определим ряд фундаментальных, базовых классов сущностей / понятий.

В первую очередь нужно указать сопредельный класс / понятие, определяющее связь рассматриваемых данных с наукой, — это *научное исследование* (или просто *исследование*), под которым будем понимать процесс получения новых научных знаний с помощью научных методов. Упрощенная и детализированная схемы этого процесса в интересующем нас контексте приведены в работе [5]. Не вдаваясь глубоко в детализацию этого сложного класса понятий, установим три ключевых, связанных с научными исследованиями, понятия, которые будут являться основными точками соприкосновения (взаимодействия) этого класса с понятием «научные данные»:

— *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.

Отметим, что практически все современные виды научных исследований так или иначе оперируют с научными данными либо используя их, либо генерируя их, либо и то и другое. Также необходимо добавить, что и научные знания в их объективизированной форме, т. е. записанные в виде текстов или иным способом, являются одним из особых видов научных данных.

Дальнейшее построение иерархии классов понятий будем осуществлять в рамках двух базовых троек классов (рис. 2):

1) *объект — субъект — действие*, в рамках которой, в частности:

— субъект совершает действие (над объектом);

— объект подвергается воздействию, т. е. действию со стороны субъекта;

2) *принцип — механизм — этап* (процесс), в рамках которой, в частности:

— принцип реализуется через механизм;

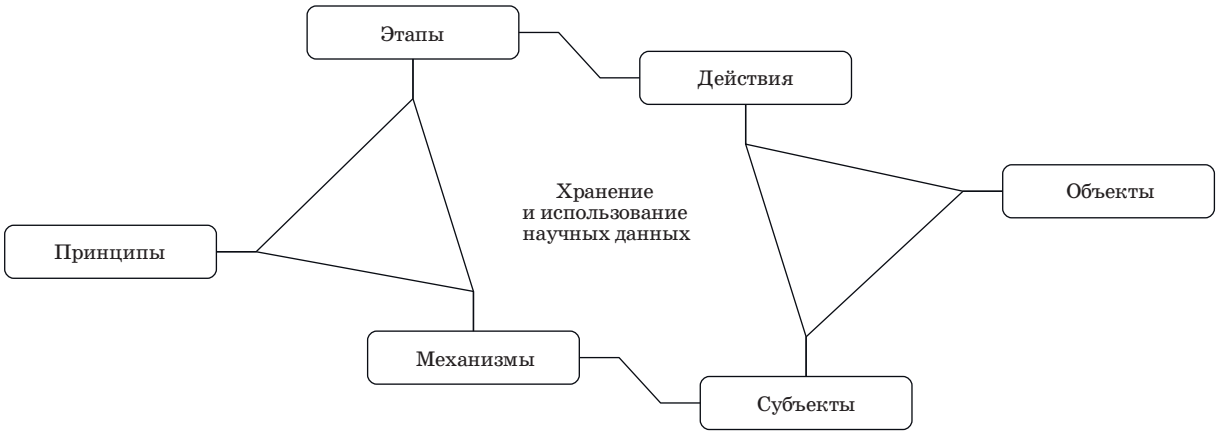
— механизм применяется на этапе.

Также установим важные связи между этими тройками:

— субъект применяет механизм;

— действие совершается на этапе.

Хотя эти тройки классов являются достаточно общими, мы будем рассматривать их конкретизацию в контексте организации хранения и исполь-



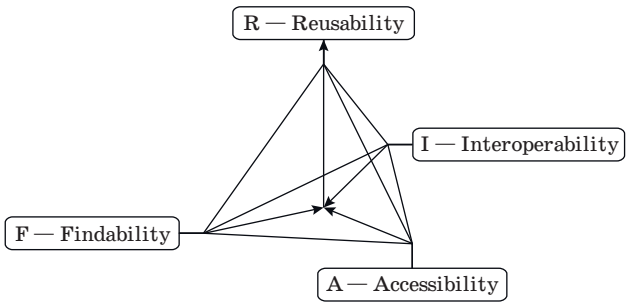
■ ***Рис. 2.*** Базовые концепты-классы организации хранения и использования научных данных
■ ***Fig. 2.*** Basic concepts-classes of research data storage and usage organizing

зования научных данных и связанных с этим процессов. В настоящей работе остановимся на тройке «принцип — механизм — этап».

Принципы

Начать детализацию базовых классов предлагается с принципов, так как их выбор во многом определяет структуру других классов.

Ключевые принципы организации хранения научных данных сформулированы международным научным сообществом под аббревиатурой FAIR: данные должны быть обнаружимы (findable), доступны (accessible), совместимы с множеством использующих их информационных систем и другого программного обеспечения (interoperable), «переиспользуемы» (reusable) [10]. В настоящее время актуальные документы и состояние вопроса (state-of-art) можно найти на сайте GO FAIR Initiative [11]. Разберем принципы FAIR.



■ ***Рис. 3.*** Целеустремленная система принципов организации работы с научными данными FAIR Data Principles
■ ***Fig. 3.*** FAIR Data Principles as a goal-driven system of research data organizing principles

F — Findable. Для того чтобы использовать существующие данные, их нужно найти. Чтобы это было легко сделать, GO FAIR Initiative предлагает:

— ассоциировать с данными глобально уникальный и устойчивый идентификатор (такой как DOI [12]);

— 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, а метаданные должны быть доступны, даже если сами данные уже недоступны.

I — Interoperable. Данные объединяются (интегрируются), анализируются, хранятся и обрабатываются различными приложениями, сервисами, системами. Чтобы это было возможно, нужно:

— использовать достаточно формализованный, общедоступный и общеизвестный, широко используемый язык для описания и представления данных и метаданных;

— использовать словари, соответствующие принципам FAIR (отметим, подобные вложенности / рекурсии могут вызывать коллизии при формальном, в том числе машинном, анализе);

— 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 (это должно позволить строить во-

круг данных своеобразные семантические сети, по аналогии с SemanticWeb [13]).

R — Reusable. Основная цель FAIR — оптимизация переиспользования данных. Именно ей и принципу «переиспользуемости» подчиняются остальные принципы FAIR (поэтому на рис. 3 они изображены в виде пирамиды). В дополнение к ним, для максимизации эффективности использования научных данных, предлагается сконцентрироваться на качественном описании и атрибутировании 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.

Здесь необходимо отметить, что возникает очень тонкая грань между элементами иерархии принципов и механизмами (их реализации). Так, принципы F, A и I детализируют принцип R, раскрывая часть механизмов для его реализации. Этой же цели соответствуют ключевые установки для информационно-аналитической системы поддержки научных исследований, основанных на интенсивном использовании цифровых данных, сформулированные в работе [5]. Упомянем те из них, которые дополняют принципы FAIR или механизмы их реализации (описанные в детализации этих принципов).

Usability@Top — удобство использования превыше всего. Установка ориентирует на обеспечение исследователей как тех, кто планирует использование чужих данных, так и тех, кто делится своими данными, необходимыми для этого инструментами, отдавая приоритет простоте и удобству использования.

StoreEverything, IntegrateData, CombineResources — хранение всех видов данных, интеграция данных и комбинирование ресурсов для снабжения исследователей инструментами для построения различных сложных коллекций данных и их совместного анализа с использованием всех доступных вычислительных ресурсов.

UseEverythingKnown — использовать при работе с данными (например, при формировании коллекций и подборе аналитических инструментов для работы с ними) всю возможную (доступную) информацию. Кроме прочего, эта установка расширяет виды хранимой информации дополнительными метаданными, предназначенными для выбора инструментов (программ, сервисов) для работы (анализа и обработки) с различными типами данных.

FromPrivate2Public — с одной стороны, это форма принципа A (accessibility), включающая вопросы аутентификации и авторизации, обеспечения конфиденциальности хранимых данных, с другой стороны — это установка на реализацию принципа открытости, который естественно дополняет другие принципы, направленные на реализацию принципа переиспользуемости: чтобы научные данные можно было эффективно переиспользовать, они должны быть открытыми [14, 15].

EnlightFromAnywhere и OntologizeAll — собирать всю информацию из всех доступных источников для построения и постоянного расширения и уточнения семантической сети вокруг данных: как их отдельных экземпляров, так и коллекций данных и даже всей доступной совокупности научных данных. Эта установка указывает, что нужно не только создавать возможности для построения семантических сетей вокруг научных данных, но и строить их, чтобы можно было использовать их при формировании коллекций данных, подборе инструментов для работы с ними [5].

Этапы и процессы

Детализацию класса этапы организации хранения и использования научных данных начнем с описания (выбора) жизненного цикла научных данных. В настоящее время нет какой-то общепринятой модели жизненного цикла научных данных или исследовательских проектов, основанных на использовании данных. Свои модели сформировали различные научные организации и сообщества. Чаще всего — это библиотеки университетов и научных организаций, что накладывает свой отпечаток на модели: они ориентированы больше на привычные для библиотек задачи сбора, сохранения и предоставления доступа к данным, организацию их каталогизации и цитирования. Одной из таких моделей является модель управления жизненным циклом информации [16], где основной упор сделан именно на проработку документальных источников, подготовку и публикацию документов, сопровождающих данные. Также это модели Библиотеки Наньянского технологического университета NTU Library (https://blogs.ntu.edu.sg/lib-datamanagement/data-lifecycle/), американской Национальной сети библиотек в области медицины NNLM (https://nnlm.gov/data/data-life-cycles), Библиотеки Университета Карнеги — Меллона CMU Library (https://library.cmu.edu/datapub/dms/data/101), Библиотеки Университета штата Вирджиния UVA Library (https://data.library.virginia.edu/data-management/lifecycle/). Большинство из них оперируют пятью базовыми стадиями-этапами жизненного цикла научных данных в различных вариациях: планирование — получение данных — обработка / ана-

лиз — сохранение — переиспользование, — по-разному раскрывая некоторые из них и формируя в них циклы. Схожая с названными выше модель предложена DDI Alliance (http://www.ddialliance.org/training/why-use-ddi). Американское географическое общество USGS представило нециклическую шестиэтапную модель, заканчивающую жизненный цикл работы с научными данными публикацией (см. [17], https://www.usgs.gov/products/data-and-tools/data-management/), имеющую, однако, отдельные уровни / слои для задач описания данных, управления их качеством и обеспечения информационной безопасности. В модели, представленной и используемой Data Observation Network for Earth's (DataONE, https://www.dataone.org/data-life-cycle), в отдельные этапы выделены задачи проверки данных (assure), их описания (describe) и интеграции (integrate). Выделение задачи проверки качества данных — важный шаг в направлении создания хранилищ и управления научными данными из разных источников, в том числе недоверенных, а задача интеграции данных необходима для обеспечения исследователей новыми возможностями, возникающими при анализе сложных комплексов разнородных данных из различных источников [5]. Более сложная и многослойная модель предложена Digital Curation Centre DCC (http://www.dcc.ac.uk/resources/curation-lifecycle-model), которая концентрируется вокруг задачи управления (курирования) данными, в этом смысле являющаяся наиболее полной из рассматриваемых, однако использованию данных в этой модели выделен только один блок (доступ / использование / переиспользование) из более чем 10; в их модели появляется важный этап — dispose, т. е. удаление / избавление от данных, который тесно связан с задачами оценки качества и отбора данных.

Число моделей жизненного цикла научных данных давно превысило 50. Комитет по спутниковым системам для наблюдения за Землей (Committee on Earth Observation Satellites — CEOS) делал несколько версий обзоров различных моделей, с версией 1.2 можно ознакомиться в отчете [18]. Обзоры с классификацией и критическим анализом ряда моделей жизненных циклов исследований и управления научными данными можно найти в работах [19 и 1].

Большинство упомянутых моделей жизненного цикла научных данных и основанных на них исследований так или иначе состоят из схожих этапов / процессов, по-разному выстраивая их в цепочки, организуя общий и внутренние циклы, а также параллельные процессы. Отметим, что существующее многообразие моделей заставляет осознать, что какой-либо общей модели нет и вряд ли целесообразно пытаться ее строить, что особенности исследовательских процессов таковы, что в каждом отдельном случае, в каждом проекте или научной задаче возникают собственные последовательности действий с данными. Однако все они так или иначе содержат и укладываются в четыре основные стадии: получение — сохранение и использование — избавление (удаление), — что соответствует максимально обобщенному жизненному циклу любого продукта.

Для дальнейшей детализации класса перечислим по возможности все этапы и процессы в ходе работы с научными данными.

1. Получение — процесс «возникновения» данных на стороне их источника. «Возникновение» здесь в кавычках, так как данные уже могли существовать в источнике (например, в системе хранения) и в рамках этого процесса могли быть предъявлены их соискателю, но в то же время они могли быть сгенерированы с помощью сенсоров / детекторов / компьютерного моделирования и др.

2. Сбор — процесс сбора или агрегации данных из разных источников для последующей работы с ними.

3. Размещение, соответственно, в системах хранения данных.

4. Проверка качества — процесс, в ходе которого выполняется своеобразное исследование данных на предмет соответствия требованиям или ожиданиям. Во время проверки качества могут быть проведены:

— оценка качества — т. е. присвоение некоторой исчислимой характеристики (в том числе вектора характеристик) качеству данных;

— ранжирование — т. е. определение «места» среди аналогов на основе полученной оценки.

5. Описание — снабжение данных метаданными, в том числе произвольным текстовым описанием.

6. Типизация — отнесение данных к какому-либо известному типу данных и привязка к соответствующим инструментам обработки и анализа.

7. Связывание — построение связей данных с другими данными, в том числе с документами, статьями, отчетами и пр., а также с иными объектами и субъектами, например, с получившими их исследователями или с оборудованием, на котором они были получены, методами обработки и анализа и др[1].

8. Организация — встраивание в существующий комплекс всех данных на основе построенных связей.

[1]Связывание с методами может происходить, например, путем отнесения к определенному классу данных, для которых уже установлены связи с методами обработки и анализа, т. е. в рамках типизации.

9. Сохранение — процесс, в рамках которого обеспечивается сохранность данных (включая их целостность).

10. Поиск — процесс получения точного указателя на конкретные данные или наборы данных, соответствующих критериям поиска.

11. Комбинирование — процесс построения коллекций (наборов, комбинаций, комплексов) данных по заданным критериям. Критериями могут быть как прямое указание на объекты данных, так и различные фильтры, а сами коллекции могут быть как неупорядоченными, так и упорядоченными или смешанными.

12. Выбор методов обработки — процесс, результатом которого является метод (комплекс методов) для обработки данных.

13. Обработка — процесс, в рамках которого данные преобразуются в другие данные, которые, например, удобнее использовать для последующего анализа.

14. Выбор методов анализа — аналогично п. 12 для анализа данных.

15. Анализ — процесс, результатом которого является некоторое новое знание об объекте или субъекте исследования, к которым относятся данные, или о самих данных.

16. Интеграция — комплексный процесс выбора или составления коллекции данных и методов (технологии) их совместной обработки / анализа, результаты которых невозможно получить без такой совместной обработки / анализа (т. е. при раздельном анализе элементов коллекции).

17. Обеспечение доступа — комплекс мер, позволяющий соискателю данных получить их непосредственно либо указать используемому для их обработки или анализа сервису, где и как их взять.

18. Изучение ограничений на передачу данных третьим лицам, ограничения могут быть связаны с действием юридических и этических законов и норм, а также с учетом прав первенства и коммерческой тайны.

19. Организация обмена — обеспечение доступа к данным избранному (указанному непосредственно или в форме критерия) кругу «третьих лиц» (например, из другой по отношению к владельцу данных научной организации / лаборатории).

20. Публикация — обеспечение доступа к данным произвольному кругу «третьих лиц».

21. Цитирование — процесс указания на использование данных при подготовке отчета, публикации или иной научной работы.

22. Утилизация — процесс уничтожения данных, например, по заданному критерию, такому, как оценка востребованности, качества или соответствующий рейтинг.

Здесь опущены этапы планирования, так как они относятся не непосредственно к работе с данными, а, в большей степени, к организации исследований и экспериментов в целом.

Комбинирование перечисленных этапов в циклы деятельности в соответствии с целями и задачами исследования — это один из механизмов организации хранения и использования научных данных, которым посвящен следующий раздел.

Механизмы

Очень важно, чтобы для каждого принципа существовали механизмы его реализации и каждому этапу / процессу также соответствовал хотя бы один механизм, реализуемый в рамках этого этапа / процесса.

Механизмы разнесем на три уровня: два инструментальных — аппаратный и программный (который далее может быть разделен на уровень платформы и уровень приложений), а также организационно-методический. Инструментальный уровень доступен пользователям (участникам процесса работы с научными данными) в виде информационных систем и сервисов, их детализация подробнее представлена в работах [5, 6].

Рассмотрим механизмы *организационно-методического уровня*. В первую очередь — это документы, регулирующие деятельность по организации хранения и использования научных данных: политики, регламенты, инструкции, рекомендации, лучшие практики, — которые образуют *регуляторный* класс.

Политики — комплексы принципов, описанные в форме установок для участников процесса работы с научными данными. Политики могут быть направлены на разные частные цели, поэтому их может быть много. Примерами могут быть «Политика открытия данных», описывающая, при каких условиях и когда научные данные могут или должны быть открыты (опубликованы), или «Политика сохранения данных», аналогично определяющая условия сохранения и удаления (утилизации) данных. Политики базируются на принципах и направлены на их соблюдение и применение. Характерные связи для политики — «субъект придерживается политики», «действие соответствует политике», «принцип охватывается политикой».

Регламенты — своды правил, определяющих условия взаимодействия при использовании ресурсов (например, информационных систем, объектов данных и классов объектов данных). Пример — «Регламент использования информационной системы». Характерные связи для регламента — это «субъект соблюдает регламент», «субъект совершает действие по регламенту», «регламент разрабатывается в соответствии с политикой».

Инструкции — содержат описания последовательностей действий в тех или иных ситуациях, предусмотренных регламентами, направленных на решение конкретных классов задач. Пример — «Инструкция по размещению данных в системе». Характерные связи для инструкций — «субъект действует по инструкции».

Рекомендации — наборы советов о действиях, которые можно предпринять при возникновении тех или иных проблем, для решения конкретных задач. Пример — «Рекомендации по аннотированию данных и снабжению их метаданными». Характерная связь — «субъект выполняет рекомендацию».

Лучшие практики — формы инструкций, которые по итогам использования приводили к лучшим результатам в решении конкретной задачи, когда решение этой задачи не является однозначным. Пример — «Лучшая практика при комплексном анализе данных конкретного типа». Характерная связь — «субъект следует лучшей практике».

Другой подкласс, который включаем в организационно-методический уровень механизмов — *методический*, он состоит из технологий, методик, методов и алгоритмов. Экземпляры этого класса могут быть описаны в виде инструкций или лучших практик, реализованы в виде компьютерных программ, доступны в информационных системах или в форме информационно-вычислительных сервисов.

Структурно-организационный подкласс организационно-методического уровня класса механизмов включает схемы организации деятельности и процессов хранения и использования научных данных. Это, собственно, модели жизненных циклов данных, организационные структуры, схемы взаимодействия субъектов и объектов при работе с научными данными.

Заключение

В работе описаны особенности научных данных и работы с ними, которые заставляют строить специальные модели, описывающие и определяющие способы и средства для эффективного использования таких данных, т. е. для получения на их основе максимального количества новых научных знаний.

Рассмотрены и описаны принципы, на которых может основываться организация хранения и использования научных данных. За основу приняты принципы FAIR [10], дополненные установками [5]. Принципы могут комбинироваться в различные политики. Придерживаясь тех или иных политик, исследователи, работающие с данными, особенно те, кто их генерирует, способствуют достижению целей политик, центральная из которых — эффективное использование и переиспользование научных данных. Представлена иерархия механизмов, которые могут применяться при работе с научными данными для решения научных и организационных задач. Перечислены основные процессы / этапы жизненных циклов научных данных и процессов исследований, основанных на них. Рассмотрен ряд принятых моделей таких жизненных циклов. Предлагается вместо попытки построить универсальную модель использовать или создавать модели на основе представленного списка этапов под конкретные случаи или классы задач исследований, основанных на данных.

Таким образом, построено ядро иерархии понятий для области знаний «организация хранения и использования научных данных», детализированное в части классов принципы, механизмы и этапы. Дальнейшие исследования будут направлены на детализацию классов понятий субъекты, объекты, действия и построение системы связей между элементами всей иерархии понятий.

Литература

1. **Cox A. M., Tam W.** A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, 2018, vol. 70, iss. 2, pp. 142–157. doi:10.1108/AJIM-11-2017-0251
2. **Земсков А. И.** Data Curation — хранение научных данных и обслуживание ими — новое направление деятельности библиотек. *Научные и технические библиотеки*, 2013, № 2, с. 85–101. http://www.gpntb.ru/ntb/ntb/2013/2/ntb_2_10_2013.pdf (дата обращения: 30.04.2019).
3. **Мелдо А. А., Уткин Л. В., Моисеенко В. М.** Алгоритмы диагностики XXI века. Искусственный интеллект в распознавании рака легкого. *Практическая онкология*, 2018, т. 19, № 3, с. 292–298. doi:10.31917/1903292
4. **Лушнов М. С., Лушнов А. М., Липовицкая И. Н., Головина Е. Г., Ступишина О. М.** Медицинская статистика и идентификация факторов риска для здоровья человека в пространстве биосферы. *Биосфера*, 2010, т. 2, № 1. https://cyberleninka.ru/article/n/meditsinskaya-statistika-i-identifikatsiya-faktorov-riska-dlya-zdorovya-cheloveka-v-prostranstve-biosfery (дата обращения: 30.04.2019).
5. **Юрченко А. В.** К концепции информационно-аналитической системы поддержки научных исследований, основанных на интенсивном использовании цифровых данных. *Вычислительные технологии*, 2017, т. 22, № 4, с. 105–120.
6. **Городничев М. А., Комиссаров А. В., Можина А. В., Прочкин П. В., Рудыч П. Д., Юрченко А. В.** Модели и проектные решения системы хранения и обработки исследовательских данных Ecclesia. *Вестник*

НГУ. Серия: Информационные технологии, 2018, т. 16, № 3, с. 87–104. doi:10.25205/1818-7900-2018-16-3-87-104

7. **Юрченко А. В.** О сервисном подходе к формированию и оценке востребованности киберинфраструктуры науки. *Информационные технологии*, 2018, т. 24, № 4, с. 219–232. doi:10.17587/it.24.219-232
8. **Шокин Ю. И., Федотов А. М., Барахнин В. Б.** *Проблемы поиска информации*. М., Наука, 2010. 197 с.
9. **Федотов А. М., Леонова Ю. В.** Требования к прототипу системы управления информационными ресурсами в распределенных информационных системах поддержки научных исследований. *Вычислительные технологии*, 2018, т. 23, № 5, с. 82–109. doi:10.25743/ICT.2018.23.5.008
10. **Wilkinson M. D.,** et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 2016, vol. 3, article number: 160018. doi:10.1038/sdata.2016.18
11. *GO FAIR Initiative*. https://www.go-fair.org/fair-principles/ (дата обращения: 30.04.2019).
12. **Paskin N.** *Digital Object Identifier (DOI®) System*. Encyclopedia of Library and Information Sciences. Third Edition. 2010. Available at: http://www.doi.org/overview/080625DOI-ELIS-Paskin.pdf (accessed 30 April 2019).
13. *Towards the Semantic Web: Ontology-driven Knowledge Management*/ Ed. John Davies, Dieter Fensel and Frank van Harmelen. John Wiley & Sons, 2003. 312 p.
14. **Yozwiak N. L., Schaffner S. F., Sabeti P. C.** Data sharing: Make outbreak research open access. *Nature*, 2015, vol. 518, iss. 7540, pp. 477–479. doi:10.1038/518477a
15. **Cutcher-Gershenfeld J.,** et al. Five ways consortia can catalyse open science. *Nature*, 2017, vol. 543, iss. 7647, pp. 615–617. doi:10.1038/543615a
16. **Allan R.** *Virtual research environments. From portals to science gateways*. Chandos Publishing, Oxford, UK, 2009. 284 p.
17. **Faundeen J. L., Burley T. E., Carlino J. A., Govoni D. L., Henkel H. S., Holl S. L., Hutchison V. B., Martín E., Montgomery E. T., Ladino C. C., Tessler S., Zolly L. S.** 2013. The United States Geological Survey Science Data Lifecycle Model. *U.S. Geological Survey Open-File Report*, 2013, 1265, 4 p. doi:10.3133/ofr20131265
18. *Data Life Cycle Models and Concepts. CEOS Version 1.2*. 2012. Available at: http://wgiss.ceos.org/dsig/whitepapers/Data%20Lifecycle%20Models%20and%20Concepts%20v12.docx (accessed 30 April 2019).
19. **Wissik T., Durco M.** Research data workflows: from research data lifecycle models to institutional solutions. *CLARIN 2015 Selected Papers, Linköping Electronic Conference Proceedings*, Annual Conference 2015, Wroclaw, Poland, October 14–16, 2015, Linköping University Electronic Press, Linköpings universitet, 2015, pp. 94–107. Available at: http://www.ep.liu.se/ecp/123/008/ecp15123008.pdf (accessed 30 April 2019).

UDC 005, 004.62
doi:10.31799/1684-8853-2019-3-45-54

Models of organizing research data storage and usage: basic principles, processes and implementation mechanisms

Yu. I. Shokin[a], Dr. Sc., Phys.-Math., RAS Academician, orcid.org/0000-0002-5178-8294, dir@ict.nsc.ru
A. V. Yurchenko[a], PhD, Phys.-Math., Vice-director, orcid.org/0000-0001-6435-1975, yurchenko@ict.nsc.ru
[a]Institute of Computational Technologies of SB RAS, 6, Academician M. A. Lavrentiev Ave., 630090, Novosibirsk, Russian Federation

Introduction: Storage and usage of research data become more sophisticated as their quantity and diversity grow. Research data have a number of features which do not allow you to copy the approaches and tools used in commercial or governmental data-processing facilities. Providing researchers with specialized tools for working with data is an urgent task in research management. **Purpose:** Identifying and describing the basic principles for working with research data, the processes and stages of this work, the mechanisms for implementing the principles and solving the problems of organizing the storage and usage of research data. **Results:** We review and discuss the principles on which the storage and usage of research data can be based, including the FAIR Data Principles. The main goal of organizing the work with research data and the central focus of its principles is the effective use and reuse of this data. We present a hierarchy of mechanisms which can be applied when working with research data for solving scientific and organizational problems. The main processes and lifecycle stages of scientific data and research processes based on them are listed in the article. A number of well-known models of such lifecycles are considered. It is proposed, instead of trying to build a universal model, to use or create models based on the presented list of stages for specific cases or classes of data-driven research. **Practical relevance:** The hierarchy of concept classes developed in the work for the field "Organizing the storage and usage of scientific data" will be used as an ontology core, and for the development of regulatory documents, recommendations and information systems supporting data-driven research.

Keywords — research data, data-driven research, FAIR Data Principles, data management, research data lifecycle, scientific information system.

For citation: Shokin Yu. I., Yurchenko A. V. Models of organizing research data storage and usage: basic principles, processes and implementation mechanisms. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2019, no. 3, pp. 45–54 (In Russian). doi:10.31799/1684-8853-2019-3-45-54

References

1. Cox A. M., Tam W. A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, 2018, vol. 70, iss. 2, pp. 142–157. doi:10.1108/AJIM-11-2017-0251
2. Zemskov A. I. Data Curation — storage and maintenance of scientific data — a new direction of libraries activities. *Nauchnye i tehnicheskie biblioteki*, 2013, no. 2, pp. 85–101 (In Russian). Available at: http://www.gpntb.ru/ntb/ntb/2013/2/ntb_2_10_2013.pdf (accessed 30 April 2019).
3. Meldo A. A., Utkin L. V., Moiseyenko V. M. XXI century diagnostic algorithms. Artifitial intelligance in lung cancer detection. *Practical oncology*, 2018, vol. 19, no. 3, pp. 292–298 (In Russian). doi:10.31917/1903292
4. Lushnov M. S., Lushnov A. M., Lipovitskaya I. N., Golovina E. G., Stupishina O. M. Medical statistics and identification of risk factors for human health in the biosphere. *Biosfera*, vol. 2, no. 1, pp. 157–165 (In Russian). Available at: https://cyberleninka.ru/article/n/meditsinskaya-statistika-i-identifikatsiya-faktorov-riska-dlya-zdorovya-cheloveka-v-prostranstve-biosfery (accessed 30 April 2019).
5. Yurchenko A. V. On the concept of information-analytical system for supporting data intensive science. *Computational technologies*, 2017, vol. 22, no. 4, pp. 105–120 (In Russian).
6. Gorodnichev M. A., Komissarov A. V., Mozhina A. V., Prochkin P. V., Rudych P. D., Yurchenko A. V. Information models and project solutions for the ecclesia research data storing and processing system. *Vestnik NSU. Series: Information Technologies*, 2018, vol. 16, no. 3, p. 87–104 (In Russian). doi:10.25205/1818-7900-2018-16-3-87-104
7. Yurchenko A. V. On the approach considering scientific IT-service as a base unit for cyberinfrastructure of science. *Information Technologie*, 2018, vol. 24, no. 4, pp. 219–232 (In Russian). doi:10.17587/it.24.219-232
8. Shokin Yu. I., Fedotov A. M., Barakhnin V. B. *Problemy poiska informatsii* [Problems of information retrieval]. Novosibirsk, Nauka Publ., 2010. 198 p. (In Russian).
9. Fedotov A. M., Leonova Y. V. Requirements for the prototype of the information resources management system in distributed information systems for the support of scientific research. *Computational technologies*, 2018, vol. 23, no. 5, pp. 82–109 (In Russian). doi:10.25743/ICT.2018.23.5.008
10. Wilkinson M. D., et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 2016, vol. 3, article number 160018. doi:10.1038/sdata.2016.18
11. *GO FAIR Initiative*. Available at: https://www.go-fair.org/fair-principles/. (accessed 30 April 2019).
12. Paskin N. *Digital Object Identifier (DOI®) System*. Encyclopedia of Library and Information Sciences. Third Edition. 2010. Available at: http://www.doi.org/overview/080625DOI-ELIS-Paskin.pdf (accessed 30 April 2019).
13. *Towards the Semantic Web: Ontology-driven Knowledge Management*. Ed. John Davies, Dieter Fensel and Frank van Harmelen. John Wiley & Sons, 2003. 312 p.
14. Yozwiak N. L., Schaffner S. F., Sabeti P. C. Data sharing: Make outbreak research open access. *Nature*, 2015, vol. 518, iss. 7540, pp. 477–479. doi:10.1038/518477a
15. Cutcher-Gershenfeld J., et al. Five ways consortia can catalyse open science. *Nature*, 2017, vol. 543, iss. 7647, pp. 615–617. doi:10.1038/543615a
16. Allan R. *Virtual research environments. From portals to science gateways*. Chandos Publishing, Oxford, UK, 2009. 284 p.
17. Faundeen J. L., Burley T. E., Carlino J. A., Govoni D. L., Henkel H. S., Holl S. L., Hutchison V. B., Martín E., Montgomery E. T., Ladino C. C., Tessler S., Zolly L. S. 2013. The United States Geological Survey Science Data Lifecycle Model. *U.S. Geological Survey Open-File Report*, 2013, 1265, 4 p. doi: 10.3133/ofr20131265
18. *Data Life Cycle Models and Concepts. CEOS Version 1.2*. 2012. Available at: http://wgiss.ceos.org/dsig/whitepapers/Data%20Lifecycle%20Models%20and%20Concepts%20v12.docx (accessed 30 April 2019).
19. Wissik T., Durco M. Research data workflows: from research data lifecycle models to institutional solutions. *CLARIN 2015 Selected Papers, Linköping Electronic Conference Proceedings*, Annual Conference 2015, Wroclaw, Poland, October 14–16, 2015, Linköping University Electronic Press, Linköpings universitet, 2015, pp. 94–107. Available at: http://www.ep.liu.se/ecp/123/008/ecp15123008.pdf (accessed 30 April 2019).