ПРОБЛЕМЫ ДОСТУПА К ДАННЫМ В ИССЛЕДОВАНИЯХ С ИНТЕНСИВНЫМ ИСПОЛЬЗОВАНИЕМ ДАННЫХ В РОССИИ*

Л. А. Калиниченко[1], А. А. Вольнова[2], Е. П. Гордов[3], Н. Н. Киселева[4], Д. А. Ковалева[5], О. Ю. Малков[6], И. Г. Окладников[7], Н. Л. Подколодный[8], А. С. Позаненко[9], Н. В. Пономарева[10], С. А. Ступников[11], А. З. Фазлиев [12]

Аннотация: Целью данного обзора является анализ глобальных тенденций создания массивных коллекций данных в мире и обеспечения возможности совместного использования таких коллекций при решении задач исследования и принятия решений в различных областях с интенсивным использованием данных (ОИИД) в России. Конкретный набор ОИИД, отобранный для обзора, включает астрономию, материаловедение, науки о Земле, геномику и протеомику, нейронауку. По каждой из рассмотренных ОИИД представлены крупные стратегические инициативы США и ЕС, примеры крупных коллекций данных в мире до 2025 г., известные проекты информационных и телекоммуникационных инфраструктур и центров данных. Включенный в обзор набор массивных коллекций данных, планируемых к получению в мире, предлагается использовать в качестве ориентира при планировании и развитии исследовательских инфраструктур для накопления и анализа данных, совместимых с зарубежными открытыми инфраструктурами в науке. В частности, рассматриваемые в обзоре коллекции данных, цели их создания и научные исследования, планируемые к осуществлению с их помощью, позволяют перейти к постановке и решению задач создания компонентов перспективных информационных и телекоммуникационных инфраструктур, таких как, например, средства концептуализации ОИИД, необходимые метамодели, средства обеспечения возможности повторного использования коллекций данных, воспроизводимости программ и потоков работ и др.

Ключевые слова: 4-я парадигма; области с интенсивным использованием данных; исследовательские инфраструктуры; коллекции данных; большие данные

DOI: 10.14357/19922264160101

1 Введение

Исследования и принятие решений в различных областях деятельности людей реализуются на основе анализа данных, накопленных в соответствующих областях, объем и разнообразие которых в наши дни растут экспоненциально.

В соответствии с 4-й парадигмой научных исследований [1], проведение исследований, движи-

*Подготовка настоящего обзора была частично поддержана различными грантами, полученными группами из вовлеченных в эту работу исследовательских организаций: для ИПИ ФИЦ ИУ РАН грантами 13-07-00579, 14-07-00548 и 16-07-01028; для ИМКЭС СО РАН грантами РФФИ 13-05-12034 и 14-05-00502; для ИМЕТ РАН грантами РФФИ 14-07-00819 и 15-07-00980; для ИНАСАН РАН грантом РФФИ 15-02-04053 и грантом Президиума РАН по программе П-41; для ИЦиГ СО РАН грантом РНФ 14-24-00123; для ИКИ РАН грантом РФФИ 15-02-10203-К; для НЦН грантами РФФИ 15-04-08744 и 15-04-05066; для ИОА СО РАН грантом РФФИ 13-07-00411.

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук; факультет вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова, leonidandk@gmail.com

[2]Институт космических исследований Российской академии наук, alinusss@gmail.com

[3]Международный исследовательский центр климатоэкологических исследований Института мониторинга климатических и экологических систем Сибирского отделения Российской академии наук, gordov@scert.ru

[4]Институт металлургии и материаловедения им. А. А. Байкова Российской академии наук, kis@imet.ac.ru

[5]Институт астрономии Российской академии наук, dana@inasan.ru

[6]Институт астрономии Российской академии наук, malkov@inasan.ru

[7]Международный исследовательский центр климатоэкологических исследований Института мониторинга климатических и экологических систем Сибирского отделения Российской академии наук, igor.okladnikov@gmail.com

[8]Центр коллективного пользования «Биоинформатика» Федерального исследовательского центра «Институт цитологии и генетики Сибирского отделения Российской академии наук», pnl@bionet.nsc.ru

[9]Институт космических исследований Российской академии наук, apozanen@iki.rssi.ru

[10]Научный центр неврологии, ponomare@yandex.ru

[11]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, sstupnikov@ipiran.ru

[12]Центр интегрированных информационных систем Института оптики атмосферы Сибирского отделения Российской академии наук, faz@iao.ru

мых данными, становится неотъемлемой частью различных областей науки, экономики, бизнеса (далее — областей с интенсивным использованием данных — ОИИД, или data intensive domains — DID). Без обеспечения все новыми данными, являющимися результатом наблюдений, измерений в природе и обществе, развитие исследований в различных ОИИД становится немыслимым.

По мере развития ОИИД извлечение данных из природы становится все более сложным и изощренным из-за необходимости проникновения во все более масштабные микро-, мезо- и макроявления. Организуются глобальные проекты и миссии (в том числе космические) по сбору и накоплению данных при помощи специализированных новейших высокотехнологичных инструментов, размещаемых не только на Земле, но и в космосе. Получение данных становится все более сложным, дорогостоящим делом, требующим развития специальных технологий и серьезных капиталовложений. В результате удается получать сырые данные, подлежащие дальнейшей обработке и анализу. Сам процесс сбора данных при изучении определенного вида явлений в конкретной ОИИД может занимать многие годы.

Наряду со сбором данных беспрецедентно быстро развиваются методы и средства накопления, обработки, анализа и управления накапливаемыми данными в разнообразных ОИИД, происходит быстрое расширение спектра задач, требующих решения на основе полученных данных, накопление опыта решения подобных задач и обеспечение возможности его междисциплинарного использования.

Главный побудительный мотив настоящей инициативной работы[1] заключается в необходимости положить начало систематическому анализу развития массивных коллекций данных в различных ОИИД в мире, создания и развития инфраструктур для накопления и использования больших коллекций данных, систематизации опыта решения задач в ОИИД и пр. Некоторыми прагматическими целями такого анализа являются выявление технических, правовых и финансовых проблем на пути обеспечения доступа ученых России в различных ОИИД к уже накопленным и ожидаемым коллекциям данных в мире[2], определение потребности создания специальных инфраструктур технических и программных средств в России для поддержки такой возможности, а также способности России заметным образом участвовать во вкладе в мировую сокровищницу данных, в создание соответствующих инфраструктур, методов и средств решения задач анализа данных.

Предварительный анализ показывает, что западный мир весьма озабочен проблемами, порожденными все возрастающим «наводнением» ОИИД большими данными, проблемами их анализа (включая анализ публикаций в виде текстов на естественном языке как части когнитивного процесса), накопления и совместного (в том числе междисциплинарного) использования данных и опыта решения задач, планированием специальных инфраструктур, позволяющих справиться с подобным наводнением по мере ввода в действие новых инструментов получения данных. Для этого организуются крупные совместные проекты, рабочие группы, обсуждаются возможные решения, планируются новые инфраструктуры и уже тестируются их фрагменты, ориентированные на получение конкретных коллекций данных после 2020 г., создаются методы и средства поддержки таких инфраструктур, отрабатываются характерные примеры будущих задач, проводятся конференции, специализированные симпозиумы рабочих групп и пр.

Вместе с тем в ряде ОИИД в России обстановка такова, что если своевременно не позаботиться о том, чтобы иметь эффективный доступ к данным (наиболее важные из которых, увы, собираются и накапливаются за пределами России), то исследования по ряду направлений во многих ОИИД можно просто прекратить.

Все это служит мотивацией для проведения анализа глобальных тенденций создания массивных коллекций данных в мире и обеспечения возможности совместного использования таких коллекций при решении задач исследования и принятия решений в различных ОИИД в России. Анализ проведен на выборке коллекций, отражающей современные тенденции научных исследований.

Отбор глобальных коллекций данных и примеров их использования в данной работе ограничен временны́ми рамками (учитываются крупные проекты накопления и использования данных в различных ОИИД, выполняемые до 2025 г.), а также конкретным набором предварительно определен-

[1]Данная статья является расширенным русскоязычным вариантом работы *Kalinichenko L., Fazliev A., Gordov E., Kiselyova N., Kovaleva D., Malkov O., Okladnikov I., Podkolodny N., Ponomareva N., Pozanenko A., Stupnikov S., Volnova A.* New data access challenges for data intensive research in Russia // 17th Conference (International) on Data Analytics and Management in Data Intensive Domains Proceedings, 2015. P. 215–237.

[2]Следует заметить, что на этом пути из-за высокой технологической сложности и стоимости средств извлечения данных во многих ОИИД, высокой стоимости самого процесса получения конкретных коллекций данных в течение ближайших 10 лет ни о каком «импортозамещении» не может быть и речи.

ных ОИИД, включающим астрономию, материаловедение, науки о Земле, геномику и протеомику, нейронауку. Этот список дополнен информатикой, в силу того что все инфраструктурные проекты в значительной мере посвящены проблемам информационно-коммуникационных технологий (ИКТ).

По каждой из рассмотренных ОИИД авторы стремились представить в нижеследующих разделах работы следующую информацию:

- крупные стратегические инициативы США и ЕС по направлению;
- примеры крупных коллекций данных в мире до 2025 г. по направлению;
- известные проекты ИКТ-инфраструктур и центров данных;
- сравнимые проекты в России, при наличии таковых.

2 Астрономические данные

2.1 Большой обзорный телескоп

Обзорный широкоугольный (поле зрения — около 10 кв. град.) 8-метровый телескоп (Large Synoptic Survey Telescope, LSST) строится в Чили на высоте 2700 м и начнет функционировать в начале 2020-х гг. Он предназначен для регистрации объектов всей доступной полусферы неба и прежде всего для обнаружения темной материи и темной энергии, поиска околоземных астероидов, изучения природы транзиентных явлений и картографирования Млечного Пути.

Поток данных, ожидаемый от проекта LSST начиная с 2020 г., будет поступать с беспрецедентной скоростью. Телескоп будет собирать данные о более чем 40 млрд объектов, а также проводить исследования переменных источников и транзиентных событий. Предполагается, что доступная для наблюдений полусфера будет полностью покрываться наблюдениями LSST в 6 фотометрических фильтрах (ugrizy) не реже, чем раз в неделю.

Объем наблюдений за ночь достигнет 15 ТБ (терабайт), что за 10 лет приведет к суммарному объему около 60 ПБ (петабайт) сжатых сырых данных, 15 ПБ баз данных, 0,5 ЭБ (эксабайт) в коллекции изображений. Эти данные соответствуют каталогу, содержащему 20 млрд галактик и 17 млрд звезд, 7 трлн детектируемых источников и около 30 трлн измерений.

Две группы данных планируется сделать публичными: автоматическая система будет посылать оповещения о транзиентных событиях, а также публично доступными будут данные верхнего уровня (каталоги). Для институтов, сотрудничающих в рамках миссии, гарантирован доступ к вычислительным ресурсам для эффективного поиска, запуска программ над базой данных в 15 ПБ и обработки изображений в базе данных в 100 ПБ [2]. Институты России не участвуют в этом проекте.

2.2 Массив квадратного километра

Массив квадратного километра (Square Kilometer Array, SKA) — наиболее амбициозный проект в радиоастрономии. Радиотелескоп, расположенный и Южной Африке, содержит тысячи отдельных антенн, занимающих площадь около квадратного километра. Он работает в широком диапазоне частот (от 50 МГц до 14 ГГц), имеет чувствительность, в пятьдесят раз превышающую возможности современных радиотелескопов, и способен производить обзор неба в 10 тыс. раз быстрее.

Количество собранной информации ставит сложную задачу хранения и потребует обработки данных в реальном времени. По оценкам, SKA может создавать эксабайт сырых данных ежедневно, который после обработки в режиме реального времени можно будет сжимать до 10 ПБ [3]. Требования к мощности компьютеров для обработки данных превышают характеристики имеющихся самых быстрых компьютеров в 2015 г., а передача данных в Интернете требует нового вида высокоскоростных сетей. Ученые России не участвуют в этом проекте.

2.3 Космическая обсерватория Гайя

Гайя (Gaia) является космической обсерваторией Европейского космического агентства (European Space Agency, ESA), созданной для астрометрии и выведенной на орбиту в 2013 г. Цель — создание трехмерного каталога 1 млрд астрономических объектов, главным образом звезд, позволяющего понять образование и эволюцию нашей Галактики. Этот каталог станет основой для нового взгляда на Галактику и ключом для решения фундаментальных астрономических проблем. Дополнительно ожидается обнаружение от тысяч до десятков тысяч планет, подобных Юпитеру, за пределами Солнечной системы (экзопланет), около полумиллиона квазаров и десятков тысяч астероидов и комет в Солнечной системе. За пять лет миссии общий объем данных составит 20 ТБ. Окончательная версия каталога будет доступна в 2020 г. Доступ к полученным данным ограничен. Российские ученые участвуют в поддержке проекта наземными наблюдениями (*Gaia Follow-Up Network*).

2.4 Детекторы гравитационных волн (gravitational wave astronomy)

Проекты LIGO (Laser Interferometer Gravitational-Wave Observatory) или Advanced LIGO и Virgo ориентированы на экспериментальное подтверждение поступления гравитационных волн от их наиболее мощных источников — взрывов коллапсирующих сверхновых и слияния нейтронных звезд в тесных двойных системах. Исходные данные проекта LIGO составляют 1 ПБ, и к ним предоставлен публичный доступ. В рамках проектов в реальном времени работает система оповещения об обнаружении и возможных областях локализации транзиентных источников гравитационных волн. Российские ученые участвуют в коллаборации LIGO. Система оповещения для поддержки проекта наблюдениями областей локализации в оптическом диапазоне доступна всем ученым после подписания соглашения с коллаборацией LIGO.

2.5 Публичные коллекции данных

Sloan Digital Sky Survey (SDSS) — один из основных продолжающихся проектов астрономических наблюдений, более 15 лет его поддержки посвящены созданию карты Вселенной. Каждую ночь широкоугольный телескоп производит более 200 ГБ многоцветных фотометрических обзоров и спектроскопических данных (для сравнения, LSST за ночь будет производить 15 ТБ). В настоящее время фотометрическими наблюдениями в пяти фильтрах покрыто около 37% всего неба. На их основе создаются каталоги, включающие звезды, звездные скопления, галактики, экзопланеты и др.

NASA/IPAC Extragalactic Database (NED) — это база данных внегалактических объектов, обеспечивающая систематический анализ интегрируемой информации из сотен обзоров неба и десятков тысяч публикаций. Диапазон наблюдаемых спектров — от гамма- до радиочастотного излучения. По мере опубликования наблюдения кросс-идентифицируются с предшествующими данными и интегрируются в базу данных для упрощения запросов и извлечения требуемых данных. Приблизительный объем базы данных NED около 20 ТБ.

Mikulski Archive for Space Telescope (MAST). Основой MAST является архив научных данных, полученных от чрезвычайно успешно до сих пор функционирующего космического телескопа им. Э. Хаббла. В него включены также данные таких космических проектов, как Kepler, IUE (International Ultraviolet Explorer), GALEX (Galaxy Evolution Explorer) и др. Объем данных составляет чуть более 100 ТБ, и они публично доступны [4].

The ESO (European South Observatory) Science Archive представляет коллекцию данных Европейской южной обсерватории. Месячный поток данных составляет $7 \sim 8$ ТБ, а полный объем превышает 100 ТБ за последние несколько лет. Данные после научной обработки в большей части становятся доступными через год. Публичная часть архива доступна для зарегистрированных пользователей из международного сообщества. России пока не удалось стать членом ESO.

2.6 Примеры астрономических миссий и коллекций данных в России

В России наиболее близким аналогом архива ESO является архив общих наблюдений Специальной астрофизической обсерватории РАН, содержащий в 2010 г. данные объемом 250 ГБ [5].

Данные международной космической обсерватории INTEGRAL составляют несколько десятков терабайт и являются публично доступными по прошествии одного года, в течение которого исключительные права на данные принадлежат заявителям ежегодных открытых программ.

В России запланировано несколько космических проектов: наряду с уже работающим с 2011 г. орбитальным радиотелескопом «Радиоастрон» [6] это Спектр-Рентген-Гамма (http://hea.iki.rssi.ru/ru/index.php?page=srg), WSO-UV [7], а также «Миллиметрон» [8].

3 Данные в исследованиях мозга

Нейронаука — это совокупность анатомии, физиологии, генетики, биохимии, патологии нервной системы, психологии. Она является передним краем изучения мозга и мышления. Изучение мозга важно для понимания того, как мы воспринимаем и взаимодействуем с внешним миром.

Количество данных, генерируемых в типовой лаборатории, проводящей исследования в нейронауке, растет с поражающей быстротой. Интеграция полученных данных в единую картину является сложной задачей. Для ее решения необходима нейроинформатика, возникающая при сотрудничестве исследователей в нейронауке с информатиками, для того чтобы как новые, так и ранее известные данные стали доступнее сообществу исследователей для ускорения нашего понимания работы мозга [9].

3.1 Исследование мозга в рамках стратегической инициативы развития инновационных нейротехнологий (BRAIN)

Инициатива Белого дома BRAIN, объявленная в апреле 2013 г., — это десятилетняя программа, нацеленная на создание динамического понимания функций мозга и демонстрацию того, как отдельные клетки и сложные нейросети взаимодействуют в здоровом или больном организме.

Главные цели анализа сетей взаимодействующих нейронов:

- идентификация и описание компонентов нейронов, определяющих их (клеток) синаптические связи друг с другом, на основе изучения динамики активности во время функционирования нейросетей в живом организме;
- понимание алгоритмов управления обработкой информации внутри нейросетей и между взаимодействующими нейросетями в мозге в целом.

Ожидается, что в результате данного исследования появится концептуальная база понимания биологической основы ментальных процессов вследствие развития новых теоретических инструментов, а также инструментов обработки данных. Теоретические и статистические исследования, а также моделирование способствуют пониманию комплексных, нелинейных функций мозга.

Программе BRAIN необходима инфраструктура для обобщения и обмена релевантными наборами данных, а также методами анализа данных. Значительным препятствием, которое затрудняет понимание работы мозга, является раздробленность исследований мозга и получаемых в результате этих исследований данных. Основной целью является согласование международных усилий по интеграции этих данных в единую картину мозга как отдельной многоуровневой системы. Объем данных о мозге на клеточном уровне имеет порядок эксабайтов. Планируется построить комплексную систему исследовательских платформ, основанных на ИКТ, которая позволила бы нейробиологам, медикам-исследователям и разработчикам новых технологий ускорить темпы их исследований.

3.2 Проект Европейского Союза по исследованию человеческого мозга

Human Brain Project (HBP) — это главный десятилетний проект Европейского Союза с бюджетом в 1 млрд Евро, нацеленный на ускорение процесса понимания работы человеческого мозга. Данный проект включает исследования по диагностике и определению расстройств мозга, а также по разработке новых технологий, основанных на принципах работы мозга [10].

Human Brain Project состоит из 13 подпроектов, охватывающих стратегические данные нейробиологии, когнитивную архитектуру, теорию, этику, менеджмент, а также развитие новых платформ, основанных на информатике.

Основной целью HBP является создание реалистичной симуляции человеческого мозга. Для этого потребуется молекулярная и клеточная информация, позволяющая моделировать и понять биологические процессы в норме и патологии. Это позволит использовать данную информацию для разработки и применения новых типов компьютеров и робототехники, т. е. для применения полученных результатов для разработки новых технологий (создания нейроморфных устройств).

Планируется построить управляемые данными модели, которые отображают то, что удалось узнать о мозге экспериментальным путем, его глубинную механику, а также познать основные принципы, на которых основан мыслительный процесс. Модели мозга будут создаваться при помощи правил обучения, максимально приближенных к реальным закономерностям, которые использует мозг. Ожидается, что подобные модели будут обучаться с помощью тех же механизмов, которые используются человеческим мозгом, и что они будут проявлять подобное интеллектуальное поведение.

Проект HBP развивает 6 новых платформ, основанных на информатике:

(1) нейроинформатика (поисковые атласы и анализ данных мозга);

(2) симуляция мозга (построение и симуляция многоуровневых моделей нервных сетей и церебральных функций);

(3) медицинская информатика (анализ клинических данных для лучшего понимания болезней мозга);

(4) нейроморфные вычисления (применение функций, подобных функциям мозга, в аппаратном обеспечении);

(5) нейроробототехника (тестирование моделей мозга и их симуляция в виртуальной среде);

(6) высокопроизводительные вычисления (обеспечивающие необходимую вычислительную способность).

3.3 Проект коннектома человека

Структурные (анатомические) связи мозга (его коннектом) могут быть отображены на нескольких

уровнях: макро- (в сантиметровом и миллиметровом масштабе), мезо- (в миллиметровом и микронном масштабе) и микромасштабе (в микронном и нанометровом разрешении). Текущие разработки по человеческому коннектому (отображению всех нейронных связей в нервной системе) проводятся только в макромасштабе [11]. Данные по Human Connectome Project (HCP) (объемом в десятки терабайт) уже доступны для анализа.

3.4 Нейробиологические базы данных

Атласы мозга (Аллена) — это проекты по совмещению геномики и нейроанатомии при помощи создания карт экспрессии генов для мозга мыши и человека [12]. Данные этих проектов будут способствовать развитию различных областей нейронаук, они помогут выяснить роль определенного гена в том или ином заболевании мозга. Разные типы клеток центральной нервной системы возникают в связи с изменением экспрессии генов. Карта экспрессии генов в мозге позволяет исследовать отношения между формой и функцией. Атлас мозга дает исследователю вид областей с отличием экспрессии генов в мозге, которые позволяют исследовать пути формирования нейронных связей. Изучение этих путей, также и с помощью методов нейровизуализации, позволит установить отношения между экспрессией генов, типами клеток и функцией различных путей мозга в организации поведения и фенотипами. Атлас позволит показать, какие гены и области мозга связаны с неврологическими и психическими расстройствами.

3.5 Данные в нейронауке в России

В ряде российских исследовательских центров (Научный центр неврологии, Институт высшей нервной деятельности и нейрофизиологии РАН, Институт мозга человека им. Н. П. Бехтеревой и др.) накоплены большие коллекции данных по анатомии, гистологии, генетике и биохимии нервной системы, компьютерной томографии, структурной и функциональной магнитно-резонансной томографии мозга, электроэнцефалографии и вызванным потенциалам при нормальном развитии и старении, а также при неврологических и психических заболеваниях. В настоящее время эти данные доступны в центрах, где коллекции были получены, и в сотрудничающих с ними организациях. Разработка открытых коллекций данных будет способствовать повышению эффективности исследований в области нейронаук.

4 Данные в геномике и протеомике

Для современной молекулярной генетики характерно появление качественно новых возможностей, связанных с использованием в исследованиях высокопроизводительных экспериментальных технологий, которые привели к беспрецедентному объему накопленных данных и знаний [13]. Эти данные используются для сравнительного анализа геномов, поиска генетических вариаций и биомаркеров, которые применяются в биотехнологии, сельском хозяйстве, фармакологии, клинических исследованиях, персонализированной медицине и т. д.

Прогнозные оценки указывают на то, что общий объем геномных данных по всем проектам ежегодно будет увеличиваться в 3 раза и достигнет к 2018 г. объема 3300 ПБ.

В настоящее время существует около 7400 высокопроизводительных геномных секвенаторов, которые работают в 1027 центрах по всему миру. В России находится только 14 геномных секвенаторов в 6 научных центрах. Поэтому бóльшая часть геномных данных генерируется за рубежом: в США, Европе, Китае, Южной Корее и др.

Мультимодальность, многоуровневость и широкомасштабность биологических систем порождают огромный объем неоднородных и распределенных данных, для которых характерна изменчивость и несогласованность, необходимость контроля точности этих данных.

Как и в других предметных областях, отсутствие технологий поиска, распределения, хранения, поддержки целостности, передачи, интеграции и визуализации больших данных существенно затрудняют анализ и систематизацию больших данных [14–16].

4.1 Коллекции геномных данных

Целью проекта «*1000 геномов*» является создание наиболее детализированного каталога генетического разнообразия человеческого генома, основанного на результатах секвенирования геномов более чем 2600 человек из 26 популяций по всему миру.

Проект «*1001 геном*» ориентирован на поиск генетических вариаций в геномах различных штаммов растения Arabidopsis thaliana, которое используется как модель для детального изучения молекулярно-генетических механизмов у растений. Эта информация открывает новые возможности в генетике, определяя аллели, ответственные за фенотипическое разнообразие целого генома одного вида как на разных уровнях, включая биохимический,

метаболический, физиологический, морфологический, так и на уровне целого растения. Результаты исследования проекта «*1001 геном*» важны для развития таких наук, как селекция растений, биотехнология и медицина.

Проект «*Геном 10K*» содержит коллекцию более чем 16 000 последовательностей геномов позвоночных, включая ныне живущих и недавно вымерших млекопитающих, птиц, рептилий, амфибий, рыб и многих других видов, находящихся под угрозой исчезновения или вымирающих [17].

Целью проекта «*Человеческий микробиом*» является описание метагенома микробных сообществ, найденных во многих частях человеческого тела, а также поиск соотношений между изменениями в микробиоме и здоровьем человека.

Проект «*Атлас генома рака*» содержит исследования геномов пациентов, страдающих от более 33 видов рака. В настоящий момент накоплена информация о более чем 7000 вариантах рака [18]. Эта информация важна для поиска генетических маркеров рака и использования их для диагностики.

4.2 Атлас протеомы человека

Интерактивный атлас протеом человека, созданный в Стокгольме, в Royal Institute of Technology, ориентирован на фундаментальные исследования в области биологии человека и применение в трансляционной медицине. В настоящее время атлас содержит 13 млн аннотированных изображений человеческих тканей.

В рамках этого проекта могут изучаться различного типа протеомы человека, например протеома домашнего хозяйства, включающая белки, экспрессирующиеся во всех типах тканей, тканеспецифические протеомы, включающие белки, которые показывают повышенную экспрессию только в одной или нескольких типах тканей, или протеомы, связанные с определенными функциями, такими как лекарственные протеомы, включающие все белки-мишени лекарств, раковая протеома, включающая белки, участвующие в патогенезе рака, а также секретома — все белки, которые секретируются, и т. д.

4.3 ELIXIR — Европейская медико-биологическая инфраструктура биологической информации

ELIXIR — это проект Европейской молекулярной биологической обсерватории (European Molecular Biology Laboratory, EMBL), реализуется как панъевропейская исследовательская инфраструктура. Целью ELIXIR является предоставление средств, необходимых для всех исследователей в области медицины и биологии, начиная с полевых биологов и заканчивая химико-информатиками, позволяя им получить полную информацию из быстрорастущего хранилища информации о живых системах. Эти данные являются основой, на которой базируется наше понимание жизни.

Задачей ELIXIR является управление сбором, контролем качества и архивированием больших объемов биологических данных, полученных вследствие биологических и медицинских экспериментов. Некоторые из этих наборов данных ранее были слишком специализированными и доступными лишь для ученых той страны, в которой они были получены.

4.4 Интеграция BILS-ProteomeXchange на основе ресурсов EUDAT

Этот пилотный проект нацелен на интеграцию хранилищ сырых данных масс-спектроскопии, данных протеомики, собираемых в BILS (Швеция) и ProteomeXchange (через базу данных PRIDE (Proteonics Identifications), EMBL-EBI, U.K.), используя Европейскую инфраструктуру EUDAT. Проект служит примером объединения национальных хранилищ данных и международных репозиториев посредством ELIXIR.

4.5 Проект BD2K

Проект *От больших данных к знаниям* (Big Data to Knowledge, BD2K) позволяет использовать биомедицинские большие данные для укрепления человеческого здоровья посредством создания, индексирования и распространения методов, инструментов и обучающих материалов. Проект BD2K (начатый в 2012 г.) имеет четыре главные цели, которые в совокупности расширят использование биомедицинских больших данных:

- упростить широкое использование биомедицинских цифровых ресурсов, сделав их более доступными, распространенными и цитируемыми;
- проводить исследования и разрабатывать методы, программное обеспечение и инструменты, необходимые для анализа биомедицинских больших данных;
- усилить обучение развитию и использованию методов и инструментов, необходимых для науки биомедицинских больших данных;

– поддержать экосистему данных, ускоряющую открытия.

В проекте участвуют 185 институтов, 11 BD2K центров мастерства (centers of excellence).

5 Данные в материаловедении

Современные материалы во многом определяют развитие человеческой цивилизации. Они широко применяются в промышленности, включая те отрасли, которые непосредственно связаны с национальной безопасностью, разработкой источников чистой энергии и обеспечением высокого уровня жизни людей. Особенностью данных в неорганической химии и материаловедении является то, что они представляют собой результат обработки и систематизации больших (сотни петабайт) объемов исходных экспериментальных данных. В связи с этим создание инфраструктуры для хранения и поиска данных — одна из важнейших проблем разработки информационных систем для материаловедения.

5.1 Инициатива генома материалов

Согласно Инициативе генома материалов (Materials Genome Initiative, MGI), объявленной Белым домом в 2011 г., ускоренное создание новых материалов, обладающих заданными свойствами, критично для достижения высокого уровня конкурентоспособности промышленности США [19]. Цель MGI — обеспечение разработки и внедрения новых материалов за счет координации исследований и предоставления доступа к расчетным моделям и инструментарию для оценки свойств и поведения материалов, а также использования прорывных методов моделирования и анализа данных. Главной целью MGI является создание механизмов, способствующих обмену данными и знаниями о материалах не только между исследователями, но и между академической наукой и промышленностью.

Инициатива генома материалов будет способствовать поддержке лидирующей роли США во многих секторах современного материаловедения и промышленности: от энергетики до электроники, от обороны до здравоохранения, а также поддержке недавних прорывов в теории, моделировании свойств материалов и data mining для существенного прогресса в материаловедении, что приведет к снижению затрат на разработку, исследование и получение новых материалов. Основой MGI является *Инфраструктура инноваций в материаловедении* (Materials Innovation Infrastructure), которая обеспечит интеграцию методов и средств современного моделирования, включающего данные, а также экспериментальный и теоретический инструментарий.

5.2 Средства организации данных о материалах

В июне 2014 г. консорциум Национальных сервисов данных (National Data Service, NDS) объявил о первом показательном проекте разработки средств для организации данных, выбрав для этого область материаловедения (Materials Data Facility, MDF) [20]. Этот проект является реакцией на инициативу Белого дома MGI по ускорению разработки современных материалов. MDF обеспечит материаловедов масштабируемым репозиторием для хранения экспериментальных и расчетных данных, в том числе и до их публикации, снабженных ссылками на соответствующие библиографические источники. MDF станет рычагом для создания национальной инфраструктуры коллективного использования информации, включая разработанные в мире базы данных по свойствам материалов и информационные системы для расчета и моделирования, а также будет способствовать организации обмена данными о материалах, в том числе и еще не опубликованными. Доступность данных и средств расчета обеспечивается современной информационной и телекоммуникационной инфраструктурой, которая позволяет предоставить данные исследователям материалов для многоцелевого использования, дополнительного анализа и проверки.

5.3 Программа VAMAS

Versailles Project on Advanced Materials and Standards (VAMAS) [21] — это программа международного сотрудничества, призванная продвигать исследования и разработки, которые обеспечивают подготовку новых стандартов для современных материалов. Предполагается, что эта программа приведет к согласованию стандартов по всему миру. Предварительные исследования при разработке стандартов особенно необходимы в случае современных материалов, поскольку традиционные тесты не всегда подходят для них. VAMAS создан для преодоления барьеров в обмене новыми технологиями, необходимыми для исследований на базе международных стандартов.

5.4 Коллекции данных в материаловедении

Коллекция данных Национального института стандартов и технологии (National Institute of Stan-

dards and Technology, NIST) США содержит информацию о широком наборе веществ и материалов: неорганических и органических веществах, включая пластмассы, углеродные нанотрубки, высокопрочные сплавы, искусственные кости и т. д., для которых в институте развиваются стендовые испытания и определяются эталонные тесты.

Коллекция данных Национального института материаловедения (Япония) содержит информацию о веществах и материалах разной природы: неорганических веществах, композитах, промышленных сплавах и т. д.

Немецкая сеть научно-технической информации STN (Scientific and Technical Network) предоставляет доступ к опубликованным экспериментальным данным о структуре и свойствах материалов, патентам и иной информации.

Коллекция данных Springer Materials (Германия) обеспечивает доступ к данным о 3000 физических и химических свойств более 250 000 материалов и веществ.

5.5 Проекты информационных систем в области материаловедения в России

Развитие информационных систем по материаловедению в России является инициативой разработчиков. Наиболее известные информационные системы разработаны в Объединенном институте высоких температур РАН [22] и Институте металлургии и материаловедения РАН [23]. Базы данных в этих системах объединены с подсистемами расчета термодинамических свойств веществ [22] и системами data mining, позволяющими конструировать еще не полученные неорганические соединения [22].

6 Коллекции данных в науках о Земле

Объектами исследования наук о Земле являются планета Земля и ее атмосфера. Комплексные исследования процессов, происходящих в литосфере, атмосфере, гидросфере, биосфере и криосфере, направлены на понимание функционирования Земли как системы. Особенностью наук о Земле является сложная иерархия предметных областей, включающих в себя как фундаментальные, так и прикладные науки. Эта иерархия накладывает жесткие ограничения как на данные отдельных предметных областей наук о Земле, так и на структуры интегрированных данных наук о Земле, предназначенных для использования в таких приложениях, как метеорология, климатология, океанология, экология.

Основные массивы данных в науках о Земле получаются в результате локальных и дистанционных наблюдений, а также численного моделирования изучаемых процессов. При этом объемы соответствующих архивов, например для данных дистанционного зондирования, достигают десятков петабайт, а данных климатических вычислительных экспериментов — единиц петабайт (CMIP5 — Coupled Model Intercomparison Project Phase 5, ERA CLIM). В области климатологии основные усилия направлены на выяснение причин и последствий происходящих сейчас и возможных в будущем глобальных климатических изменений. Прикладной целью этих исследований является создание «службы климата» как аналога службы погоды. Инструментами получения данных здесь являются сети метеостанций, сети плавающих в океанах буев, сети наземных измерительных комплексов, осуществляющие наблюдения за локальными климатическими и экологическими характеристиками, системы спутников, осуществляющих наблюдения за атмосферой и поверхностью Земли, и климатические модели.

Основным источником больших массивов данных являются спутники. Петабайтные коллекции данных формируются, поддерживаются и обслуживаются в США профильными национальными ведомствами (NOAA — National Oceanic and Atmospheric Administration, NASA — National Aeronautics and Space Administration, DoE — Department of Energy), а в Европе — либо наднациональными тематическими структурами (ECMWF — European Center for Medium range Weather Forecasting, ESA), либо консорциумами ведущих по теме университетов и исследовательских центров. Эти структуры активно участвуют в реализации указанных программ.

6.1 Примеры крупных проектов получения и накопления данных в науках о Земле

В области наук о Земле, точнее об окружающей среде, наибольший прогресс в деле обеспечения всего комплекса работ, связанных с большими объемами данных, достигнут в области дистанционного зондирования Земли. Примеры соответствующих программ рассматриваются далее.

В ЕС наиболее амбициозную программу *Copernicus* возглавляют Европейская комиссия и ESA. Европейское космическое агентство координирует доставку данных с 30 спутников, а комиссия отвечает за проект, устанавливает требования и управляет сервисами.

Европейское космическое агентство создает семейство спутников (Sentinels) для оперативных нужд этой программы. Спутники будут проводить уникальный набор наблюдений, таких как всепогодные круглосуточные радарные изображения, получать оптические изображения высокого разрешения для наземных сервисов, данные для сервисов, относящихся к океану и приземному слою, данные по мониторингу состава атмосферы с геостационарных и полярных орбит, а также данные с радарного высотомера для измерения высоты морской поверхности в океанографии.

Программа *Copernicus* [24] обеспечит сервисы для предсказания качества воздуха, предупреждения наводнений, раннего обнаружения засух и опустынивания, оценки качества морской воды и анализа урожая зерновых, мониторинга лесов, контроля изменений землепользования, предсказания катастрофических погодных условий, фиксации разливов нефти, слежения за отклонением кораблей от курса и т. д.

Программа *Copernicus* (получая 5 млрд Евро за период 2014–2020 гг.) предусматривает доставку высококачественных данных (до 8 ТБ в день) в рамках политики, основанной на полном и открытом доступе к данным. Предоставляя данные высокого разрешения о приземном слое, океане и атмосфере, Copernicus получит возможность управлять развитием исследований и сотрудничества в новых приложениях наук о Земле.

Развиваемая в США *Earth Observing System* (EOS) — это координированный набор спутников для долговременных глобальных наблюдений приземного слоя Земли, биосферы, земной поверхности, атмосферы и океанов, позволяющих улучшить понимание Земли как сложной интегрированной системы. Информационная инфраструктура EOS содержит 12 национальных центров в США, которые хранят и обеспечивают непрерывный доступ к широкому разнообразию геофизической информации о Земле и космосе: полярных и приземных процессах; верхней атмосфере, глобальной биосфере, атмосферной динамике и геофизике; физической океанографии, радиационному бюджету, тропосферной химии, облакам и аэрозолям; глобальном распределении снега и льда; криосфере; биохимической динамике; воздействии человека на окружающую среду; гидрологическом цикле; климате и погоде; геофизике земной тверди, геологии и геофизике морей, солнечно-земной физике, палеоклиматологии; спутниковом дистанционном зондировании. Данные и средства работы с ними объединены в *Earth Observing System Data and Information System* (EOSDIS) [25].

В настоящее время этот опыт активно используется в мире для создания национальных сегментов такой глобальной информационной системы и для глобальной инфраструктуры международного проекта по наблюдению Земли из космоса *GEOSS* (*Global Earth Observation System of Systems*) — международного проекта, рассчитанного на несколько десятилетий. Объем финансирования — десятки миллиардов долларов.

Проект *Data Observation Network for Earth* (DataONE, https://www.dataone.org) является основой для создания науки об окружающей среде в форме распределенной базы и устойчивой киберинфраструктуры для открытого, постоянного, устойчивого и безопасного доступа к качественным описаниям и легкодоступным данным наблюдений о Земле. Проект не предназначен для хранения данных. Он является основой для соединения многочисленных репозиториев в федеральных сетях для поиска, извлечения и обеспечения репликаций на репозиториях данных внутри сетей.

В проекте будет создано легкое и просто инсталлируемое программное обеспечение и развита совместимость программного обеспечения, уже развернутого в репозиториях по всему миру. Новыми чертами проекта будут:

- семантический поиск результатов измерений;
- отслеживание всех этапов жизненного цикла данных;
- сервисы обработки данных, дающие исследователям возможность простыми способами обращаться к большим данным.

Проект *Satellite Observations for Climate Modeling* (SOCM) посвящен интеграции спутниковой информации и моделированию процессов. Новое поколение инфраструктуры будет поддерживать сравнение спутниковых наблюдений с климатическими моделями. Публикация данных дистанционного зондирования вместе с результатами моделирования климата будет способствовать их сравнению и пониманию. Кроме того, лица, принимающие ключевые решения о будущем климата, состоянии регионального уровня туризма, водных ресурсах и управлении питанием: штаты, федеральное правительство и иностранные структуры, — будут использовать эту более полную информацию.

Следующим шагом является преобразование климатической аналитики в сервисы [26]. Например, сервис CAaaS (Continuous Analitics as a Service) сочетает вычисления высокой производительности и аналитику данных с масштабируемым управлением данными, виртуализацией облачных вычислений, представлением адаптивной аналитики

и API (Application Programming Interface), связанными с предметными областями для улучшения доступа к большим коллекциям климатических данных.

В рамках международного сотрудничества *Earth System Grid Federation* (ESGF) созданы порталы, интегрирующие коллекции научных данных, распределенные по всему миру. В рамках этого сотрудничества развивается виртуальная среда *Earth System Grid* (ESG) для содействия анализу глобальных климатических изменений и обеспечивается доступ к предсказанным климатическим данным. В частности, исторические климатические данные и результаты моделирования по климатическим сценариям, выполненные при подготовке недавнего доклада Межправительственной группы экспертов по изменению климата, распространялись через ESGF. В настоящее время около 2 ПБ данных архивированы в узлах ESGF, распределенных по всему свету.

В Европе скоординированный подход к созданию глобальной инфраструктуры данных был выработан в ходе выполнения проекта 7-й Рамочной программы ЕС «Global Research Data Infrastructures: The Big Data Challenges». Реализующие эту программу проекты в Европе можно разделить на три группы: проекты наднациональных структур (ESA, ECMWF), внутригосударственные проекты развития ведомств, о которых имеется очень скупая информация, и межгосударственные научные проекты в рамках программ ЕС, информация о которых доступна в Сети (http://cordis.europa.eu). В частности, список инфраструктурных проектов 7-й Рамочной программы включает более чем 350 проектов. Не менее десятой части этих проектов связано с науками о Земле.

Следует добавить, что основой многих прикладных направлений наук о Земле являются результаты фундаментальных наук. В частности, существенную роль играют количественные данные, полученные в таких фундаментальных науках, как спектроскопия, химия атмосферы и др. Учитывая, что число молекул, рассматриваемых при решении задач, например, прогноза качества воздуха в регионе, достигает почти тысячи, а с учетом их изотопов — более двух тысяч, объем спектральных данных и затраты на анализ их качества, с учетом постоянного потока данных в новых спектральных интервалах, делают такие задачи чрезвычайно трудозатратными. Одним из выполненных в Европе проектов, относящихся к фундаментальным наукам, стал проект VAMDC — Virtual Atomic and Molecular Data Center [27, 28]. Этот проект ориентирован на исследовательские группы и институты, играющие центральную роль в производстве атомных и молекулярных данных, которые критичны для использования в широкой области применений.

6.2 Сравнимые проекты в России

В области создания информационных ресурсов для наук о Земле инфраструктурных научных проектов, сравнимых по масштабу с названными в подразд. 6.1, в России не было.

Крупным ведомственным проектом является ЕСИМО (Единая государственная система информации об обстановке в Мировом океане).

Более мелкие проекты связаны с пространственными данными субъектов РФ. Проекты в области наук о Земле, финансируемые РФФИ и РНФ, не являются частью каких-либо долгосрочных государственных программ.

7 Инфраструктуры данных и проекты для доступа к данным и анализа перспективных источников информации

7.1 Проекты исследовательских инфраструктур в Европейском Союзе

Исследовательские инфраструктуры, создаваемые в ЕС, представляют собой средства, ресурсы или сервисы уникальной природы, которые были идентифицированы в различных областях сообществами исследователей Европы для поддержки соответствующей деятельности на высоком уровне. Подобное определение *исследовательской инфраструктуры*, включая ассоциированные с ней людские ресурсы, охватывает крупное оборудование или наборы инструментов вместе с содержащими знания ресурсами, такими как коллекции данных, архивы или банки данных.

Европейский стратегический форум исследовательских инфраструктур (European Strategy Forum on Research Infrastructures, ESFRI) является стратегическим механизмом, образованным в 2002 г. странами — членами ЕС и Еврокомиссией, чтобы способствовать научной интеграции Европы и усилению ее международного влияния. Члены ESFRI назначаются министрами науки стран — членов или ассоциированных членов ЕС, а также включают представителей Еврокомиссии. Они работают совместно для определения объединенного видения и общей стратегии, включающих в качестве инструментов планирования и реализации новых

панъевропейских исследовательских инфраструктур регулярно обновляемые дорожные карты, отчеты и критерии. Подобный стратегический подход нацелен на обеспечение Европы наиболее современными исследовательскими инфраструктурами, отвечающими нуждам быстро развивающихся областей науки, продвижение основанных на знаниях технологий и расширение их применений.

Ряд примеров исследовательских инфраструктур, деятельность которых приводит к образованию новых коллекций данных и знаний, к их совместному использованию, рассматривается ниже.

ЦЕРН (http://home.web.cern.ch) — наибольшая в мире лаборатория ядерной физики частиц; именно ЦЕРН стал родоначальником идеи исследовательской инфраструктуры.

GEANT (http://www.geant.net/pages/home.aspx) — проект высокоскоростной сети, является примером инфраструктуры, способствующей совместному использованию данных и знаний учеными.

EMMA — Европейский архив мышиных мутантов (European Mouse Mutant Archive, http://www.emmanet.org) — типичный пример распределенной инфраструктуры с узлами в шести странах, представленной для пользователей в виде единственного центра.

SIOS (Svalbard Integrated Arctic Earth Observation System, http://www.sios-svalbard.org/servlet/Satellite?c=Page&pagename=sios/Hovedsidemal&cid=1234130481072) — интегрированная система наблюдений Арктики на Шпицбергене, предназначена для изучения геофизических, химических и биологических процессов, охватывая всю арктическую систему, начиная от верхних уровней атмосферы до процессов в морских глубинах и земной коре.

BBMRI-LPC (Biobanking and Biomolecular Resources Research Infrastructure — Large Prospective Cohorts, http://www.bbmri-lpc.org/about) — исследовательская инфраструктура для получения биобанков биомолекулярных ресурсов — одна из наибольших сетей поддержки биобанков в Европе; целью проекта является изучение подобных коллекций и связи накопленных данных со здоровьем людей.

EMbaRC (European Consortium of Microbial Resource Centres, http://www.embarc.eu) — Европейский консорциум центров микробиомных ресурсов, служит для координации обеспечения информационными микробиомными ресурсами исследователей в Европе и в мире.

SYNTHESYS (Synthesis of Systematic Resources, http://www.synthesys.info) — проект создания интегрированной европейской инфраструктуры для поддержки коллекций естественной истории.

7.2 Панъевропейская инфраструктура данных EUDAT

Европейская комиссия поддерживает развитие панъевропейской междисциплинарной инфраструктуры данных в рамках программы Horizon 2020, следуя нескольким ведущим принципам.

Федерализация. Предполагается, что основные действия над данными реализуются в федерациях данных. Они являются сетями репозиториев и центров данных, которые предоставляют структуры для обработки данных и действуют на основе соглашений о легальных или этических правилах, интерфейсах и спецификациях протоколов, а также стека общих сервисов манипулирования данными. Такие центры могут являться членами многих федераций. Координированный подход предполагает, что каждый центр создает описание своих возможностей, а каждая федерация может использовать одни и те же описания для извлечения необходимой информации. Такой подход способствует открытому представлению исследовательских данных и помогает изменять существующую культуру исследований для поддержки совместного использования данных.

Открытое совместное использование данных. Поскольку научные дисциплины интернациональны по своей природе, то критичным является следование международным подходам к снижению барьеров при обмене данными или при их повторном использовании. На этом пути основными препятствиями являются неоднородность данных и языков запросов, способность к пониманию и обнаружению данных, перемещение данных сквозь семантические границы между многозначными контекстами, а также проблемы рассогласования данных (относительно качества, неполноты, абстракции данных).

Европейская инфраструктура данных EUDAT является начальным шагом в этих направлениях. EUDAT (http://www.eudat.eu) объединяет 25 европейских партнеров, включающих центры данных, провайдеры технологий, сообщества исследователей и фондовые агентства из 15 стран. EUDAT предлагает общие сервисы данных в рамках географически распределенной сети, связывающей центры данных и специализированные репозитории, а также решения для поиска, совместного использования, хранения, репликации, стадийности первичных и вторичных данных исследований и выполнения их анализа. Такая сеть образует Совместную инфраструктуру данных (Collaborative Data Infrastructure), обозначаемую далее СИД, которая развивается как сервис-ориентированная, междисциплинарная и устойчивая инфраструктура.

EUDAT2020 — новый трехлетний большой проект развития СИД, начатый в 2015 г., целями которого являются: поддержка политики Европейской комиссии открытого доступа к данным исследований, достижение интероперабельности существующих в Европе инфраструктур научных исследований (ИНИ) для доступа ученых к сетевым, вычислительным ресурсам и ресурсам данных в различных ИНИ, включая гриды и облачные инфраструктуры. Так, например, будут достигнуты возможности подключения данных в СИД к высокопроизводительным ресурсам, организуемым в рамках PRACE (Partnership for Advanced Computing in Europe), для их анализа или в качестве входных данных моделей и репликации полученных результатов в систему хранения EUDAT; подключения данных в СИД к гридам и облачным ресурсам, поддерживаемым EGI (European Grid Infrastructure); а также федерализации данных при их подключении к ряду европейских инициатив (таких как Nebula, GEANT, TERENA, OpenAIR и др.).

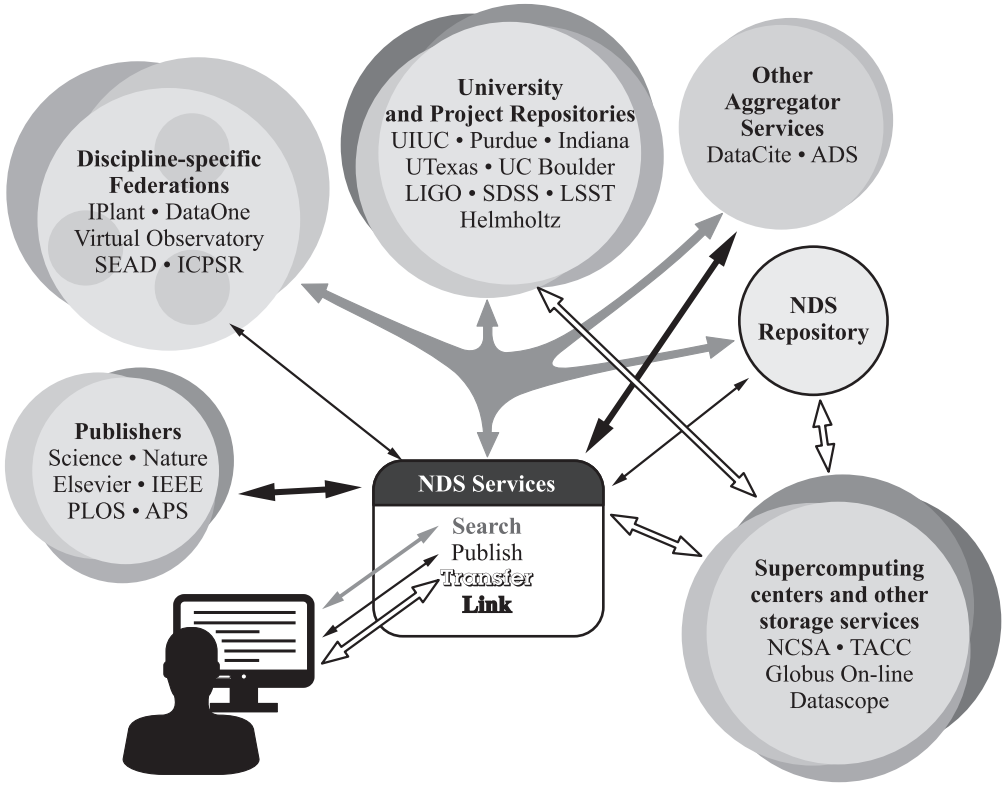
При организации EUDAT2020 достигнута договоренность о партнерстве с NDS по образованию совместных пилотных проектов (междисциплинарных и межконтинентальных). В СИД будет поддерживаться функция долгосрочного архивирования данных, репликации, каталогизации, цитируемости данных наряду с обеспечением обнаружения, доступа, повторного использования коллекций и отдельных объектов данных. Функции анализа данных будут поддерживаться ресурсами EGI и PRACE, а также средствами, образуемыми на основе виртуализации вычислительного оборудования центров данных и кластерных платформ.

Специальная программа в рамках EUDAT2020 ориентирована на создание средств оценки качества данных и сертификации репозиториев данных в СИД. EUDAT2020 развивает мультидисциплинарный подход, охватывая сообщества исследователей в гуманитарных областях и в социальных сетях (CLARIN — Common Language Resources and Technology Infractructure, DARIAH, CESSDA), в науках о Земле и атмосфере (EPOS — European Plate Observing System, ICOS, EMSO, VERCE, IAGOS, DRIHM), науке о климате (ENES — European Network for Earth System), биоразнообразии (LifeWatch, LTER, iMarine), науке о жизни (VPH, ELIXIR, BBMRI, ECRIN, INCF, DiXA) и физике (EISCAT, EURO-VO, ISIS, WLCG, PaNdata). Значительное внимание в проекте будет уделено динамическим данным и научным потокам работ, созданию сервисов управления динамическими данными, оставаясь в рамках СИД. Эти исследования будут опираться на сценарии динамического использования данных из ENES и EPOS и обобщения их для анализа будущих динамических данных при решении реальных научных задач. Одним из планируемых результатов будет создание модели и языка представления жизненного цикла данных, сервисных инфраструктур и происхождения данных. Одновременно будут происходить исследования инфраструктурных операций более эффективных, надежных, устойчивых и близких к потребностям научных сообществ. Примерами планируемых задач являются следующие: оценка объектно-ориентированной среды хранения для машин баз данных центров данных при создании масштабируемой и интероперабельной СИД на основе облачных решений (одной из целей этого анализа является определение возможности реализации B2SHARE без использования iRODS); расширение возможностей уровня долговременного хранения путем применения распределенной графовой базы данных для поддержки отношений между объектами данных вместо собственной базы метаданных, используемой в настоящее время в B2SHARE-сервисе (по замыслу это должно сблизить подходы в СИД с применениями семантического веба, поддержкой происхождения данных и семантического аннотирования).

7.3 Инфраструктура проекта «Национальные сервисы данных»

США и ряд международных научных сообществ нуждаются в унификации структур и сервисов для хранения, совместного использования, публикации, размещения и верификации данных. Нужны стандартные средства доступа к данным, программному обеспечению, метаданным, инструментам и иным компонентам, характерным для многих дисциплин. Отсутствие таких стандартных средств создает трудности при проведении исследований и репродуцировании опубликованных научных результатов. США планируют открытую инфраструктуру для поддержки интегрированного набора сервисов национального масштаба для эффективного, удобного и безопасного хранения, совместного использования, публикации, обнаружения, верификации и атрибуции данных на уровне индивидуальных, групповых и кооперативных потребностей. Именно такую инфраструктуру и сервисы формирует проект NDS [29] (см. рисунок).

Международные партнеры, в особенности Research Data Alliance (Альянс исследовательских данных) — RDA, будут способствовать NDS в обеспечении транспарентного, глобального доступа к данным.



Среда NDS

7.4 Альянс исследовательских данных

Альянс исследовательских данных был образован для поддержки совместного использования данных сквозь барьеры в 2013 г. Ядро организаторов включало Европейскую комиссию, National Science Foundation, NIST, Министерство инноваций Австралии. В настоящее время число членов альянса превышает 2600 из 90 стран. В рамках альянса образовано большое число рабочих групп и групп по интересам. Дважды в год организуются пленарные совещания в различных местах мира. Например, в марте 2015 г. на совещании в Сан-Диего рассматривались крупномасштабные инфраструктурные проекты организации и анализа данных (включая EUDAT, DataOne, CLARIN, Supercomputing and Big Data, ELIXIR, NDS и др.). Пока еще RDA находится в состоянии обсуждения и уточнения целей альянса.

7.5 Проекты обеспечения доступа к ожидаемым данным (на примере астрономии)

Разнообразные проекты (миссии) в мире в различных предметных областях, рассматриваемые в настоящем обзоре, недавно начали получать данные или планируют начать получать их до либо после 2020 г. В разных странах исследователи в соответствующих областях X-информатики уже начали или подготавливают исследования инфраструктур, поддерживающих доступ к данным, их анализ и управление данными в подобных проектах (миссиях). В настоящем обзоре астрономия выбрана для того, чтобы показать примеры подобных исследований, относящихся к проекту LSST, миссии Gaia, а также к проекту ASTERICS,

7.5.1 Подготовка к доступу к данным в проекте LSST

В марте 2015 г. заключено партнерское соглашение между Institut National de Physique Nucléaire et de Physique des Particules (IN2P3), Корпорацией LSST, проектом LSST, а также NCSA (Национальным центром суперкомпьютерных приложений Иллинойского университета) о вкладе в доступ и обработку версий данных LSST во время формирования телескопом обзоров неба [30]. По этому соглашению IN2P3 должен реализовать операции обработки данных LSST посредством коммуникаций, средств обработки данных, а также персонала для образования годичных версий обзора в узле об-

работки, сателлитном по отношению к Архивному центру NCSA LSST. Цель проекта CNRS/LSST заключается в том, чтобы предоставить получаемые LSST данные ученым и более широкой аудитории в мире в виде двух видов данных: (а) извещения о транзиентах, посылаемые в течение 60 с после завершения формирования изображения; (б) годичные релизы данных, которые будут содержать наиболее полно обработанные данные обзора. Каталог годичного релиза будет состоять из более 100 таблиц, самыми важными из которых будут каталог объектов, суммирующий для каждого физического источника всю информацию, накопленную за время действия проекта, а также исходный каталог, обеспечивающий доступ к данным каждого конкретного наблюдения одного объекта за одну экспозицию. Данные будут представлены в виде, при котором алгоритмы их анализа смогут сосредоточиться на извлечении знаний из каталогов, накопленных в базе данных без необходимости доступа к первоначальным пикселам. Согласно проекту, предполагается применить массивно-параллельную реляционную технологию баз данных (основанную на принципах архитектуры shared nothing), которая при текущем уровне развития оценивается как более эффективная, чем Map-Reduce. Предварительные измерения производительности и масштабируемости были проведены в проекте LSST на различных кластерах: от 20 узловых 100-терабайтных кластеров до 300 узловых 30-терабайтных кластеров с таблицами, содержащими порядка 50 млрд строк.

7.5.2 Подготовка к доступу к данным миссии Gaia

Космический аппарат Gaia в среднем за день передает 40 ГБ данных. К концу миссии ожидается накопление данных, превышающих 1 ПБ. Для доступа к данным созданный для обработки и анализа данных Европейский научный консорциум (DPAC) образовал шесть центров обработки данных (DPC), разбросанных по Европе: Мадрид (DPCE), Тулуза (DPCC — Data Processing and Coordinating Center), Кембридж (DPCI), Турин (DPCT), Барселона (DPCB) и Женева (DPCG) [31].

Различаются два вида DPC: (а) основанные на «инфраструктурном пакете», соединенном с централизованной файловой системой (DPCE, DPCT, DPCB, DPCG); (б) использующие Hadoop (DPCC и DPCI). DPCC ориентирован на обработку спектров, в конце миссии планируется установка в нем 6000 ядер в кластере с 2-гигабитной сетью для связи между узлами и 10-гигабитной сетью между стойками (racks).

7.5.3 Подготовка к доступу к разнообразным данным в комплексном проекте исследовательской инфраструктуры ASTERICS

В 2015 г. стартовал координируемый ESFRI и финансируемый Horizon 2020 комплексный проект ASTERICS (http://www.asterics2020.eu). Это первый проект, в котором совместно рассматриваются проблемы астрономии, астрофизики и физики космических частиц. В проекте будут использованы различные координируемые ESFRI инфраструктуры (включая SKA, массив телескопов Черенкова CTA (https://portal.cta-observatory.org/Pages/Home.aspx), глубоководный нейтринный телескоп KM3NeT (http://www.km3net.org), гигантский, создаваемый ESO телескоп E-ELT (http://www.eso.org/public/unitedkingdom/teles-instr/e-elt) и другие проекты). Основные цели ASTERICS — поддержка и ускорение реализации телескопов, находящихся в ведении ESFRI, и обеспечение их интероперабельности в рамках интегрированного, многочастотного и многоцелевого посредника. Основные ожидаемые результаты четырехлетнего проекта включают: создание технологий для обеспечения надежного и гибкого манипулирования гигантскими потоками данных, генерируемыми названными выше инфраструктурами, охватываемыми ASTERICS, адаптацию и оптимизацию систем управления огромными базами данных для нужд создаваемой инфраструктуры, адаптацию средств виртуальной обсерватории IVOA для использования в результирующей инфраструктуре. Кроме того должны быть проведены исследования возможности анализа данных в создаваемой инфраструктуре, применяя средства статистического анализа и data mining над коллекциями данных петабайтного масштаба.

8 Заключение

Практически во всех ОИИД данные становятся стратегическим ресурсом, затрагивающим все сферы деятельности людей и определяющим конкурентоспособность, уровень развития науки, промышленности, здравоохранения, обороноспособности страны. Анализ состояния пяти представительных областей науки в обзоре показал следующее.

Новизна ситуации заключается в том, что повсеместно в мире развивается процесс образования петабайтных коллекций данных как результат применения новых высокотехнологичных наземных или космических миссий и инструментов в крупных программах (инициативах) исследований, посвященных изучению разнообразных явлений окру-

жающей среды в различных ОИИД. В некоторых областях массивные коллекции данных образуются как результат интеграции большого числа относительно небольших баз данных, создаваемых в различных исследовательских лабораториях мира. В США получение петабайтных коллекций данных часто является одним из естественных результатов стратегических инициатив, объявляемых на уровне Президента США, вовлекающих большое число государственных ведомств и ведущих исследовательских центров в выполнение соответствующих проектов. В Европейском Союзе подобные программы являются межгосударственными.

В России практически нет межведомственных крупных исследовательских программ, которые требовали бы создания новейших инструментов изучения природных явлений, а также крупных международных информационных инфраструктур для накопления и анализа данных (например, со странами БРИКС и ШОС). Большая часть исследовательских проектов организуется инициативно в рамках межличностных, академических и университетских связей. Потребности в научных данных в стране не формируются системно государственными органами науки, ими не регулируются процессы дублирования действий разных ведомств, научных институтов и университетов в области накопления, стандартизации и контроля качества данных.

В результате вклад России в мировые коллекции данных незначительный; более того, трудно прогнозировать изменение ситуации в ближайшие 10 лет ввиду неразвитости соответствующих технологий в стране и отсутствия возможности образования адекватных программ, требующих значительных ассигнований. Таким образом, одной из важнейших проблем сохранения уровня научных исследований в России является обеспечение возможности эффективного доступа исследовательских организаций России к данным, накапливаемым в мире. Доступ к центрам данных, размещенным на территории иностранных государств, требует решения ряда серьезных технических проблем, а также преодоления политических и финансовых ограничений (часто требующих заключения международных соглашений). Эффективный доступ означает возможность проведения анализа данных с темпом их предоставления для ученых в мире. При этом следует понимать, что совершенно недостаточно создания методов решения типовых классических проблем — статистических, машинного обучения, data mining и пр. Опыт показывает, что в конкретных ОИИД каждая конкретная задача анализа данных, особенно больших, требует проведения исследований и экспериментов для создания специального подхода к решению задачи, по возможности опираясь на типовые методы.

Анализ показал, что, в отличие от России, за рубежом идет активная подготовка к использованию новых источников данных (примеры подготовки даны в разд. 7), включая обсуждение и планирование проектов новых информационных инфраструктур (таких как, например, ASTERICS, NDS, EUDAT2020, RDA, DataONE, MDF, ELIXIR и др.), создание и отработка элементов таких инфраструктур (например, для анализа данных, которые начнут поступать в ближайшее время (миссия Gaia), или не ранее чем через пять лет (телескоп LSST), или по завершении проекта (ASTERICS)). Для этого в каждом крупном проекте образуются большие международные междисциплинарные сообщества специалистов, рабочие группы для спецификации новых функций, которые должны поддерживаться новыми инфраструктурами.

Приведенные в обзоре примеры коллекций данных, создаваемых в мире, инфраструктур формирования новых коллекций данных в процессе исследований предполагается использовать в качестве ориентира при планировании создания и развитии исследовательских инфраструктур для накопления и анализа данных в России, совместимых с зарубежными открытыми инфраструктурами данных в науке. В частности, рассматриваемые в обзоре коллекции данных, цели их создания и научные исследования, планируемые к осуществлению с их помощью, позволяют планировать создание в России компонентов перспективных ИКТ-инфраструктур, таких как, например, средства концептуализации ОИИД, необходимые метамодели, средства обеспечения возможности повторного использования коллекций данных, воспроизводимости программ и потоков работ и др.

Для достижения эффективного доступа исследовательских организаций России к данным, накапливаемым в мире, с целью их использования в исследовательских проектах России представляется целесообразной организация целевой междисциплинарной программы для реализации пилотного проекта распределенной инфраструктуры для накопления и анализа данных, совместимой с зарубежными открытыми инфраструктурами в науке. Предполагается, что программа должна включать решение следующих основных задач:

– анализ и выбор вариантов инфраструктур и платформ для поддержки решения задач анализа больших данных в различных ОИИД, а также для обеспечения доступа исследователей к разнообразным видам данных в мире и совместного междисциплинарного их исполь-

зования (наряду с техническими проблемами, в том числе коммуникационными, предполагается решение на международном уровне правовых и финансовых проблем, вызываемых установленными ограничениями доступа к конкретным коллекциям данных);

– организация рабочих групп и формирование сообществ в различных областях с интенсивным использованием данных, принятие мер для установления контактов с международными сообществами аналогичного назначения;

– создание высокопроизводительного междисциплинарного центра интенсивного использования данных (МЦИИД) для исследователей и практиков из разнообразных ОИИД, накопление междисциплинарного опыта создания подходов к решению конкретных задач анализа данных в конкретных ОИИД, реализация проектов с интенсивным использованием данных в МЦИИД, выработка предложений по тиражированию МЦИИД в стране, их интероперабельности и размещению в составе распределенной междисциплинарной инфраструктуры совместного использования данных.

Литература

1. The fourth paradigm: Data-intensive scientific discovery / Eds. T. Hey, S. Tansley, K. Tolle. — Redmond, WA, USA: Microsoft Research, 2009. 284 p. http:// goo.gl/edvr6W.
2. *Juric M., Tyson T.* LSST data management: Entering the era of petascale optical astronomy // High. Astron., 2015. Vol. 16. P. 675.
3. *Taylor A. R.* Data intensive radio astronomy en route to the SKA: The rise of big radio data // High. Astron., 2015. Vol. 16. P. 677.
4. *Fleming S. W., Abney F., Donaldson T., et al.* Beyond the prime directive: The MAST discovery portal and high level science products // American Astronomical Society Meeting (AAS 225), 2015. #336.59.
5. *Zhelenkova O., Vitkovsky V., Plyaskina T.* Electronic archive of observational data of astrophysical observatory // Russ. J. Digital Libraries, 2010. Vol. 13. Iss. 4. http:// www.elbib.ru/index.phtml?page=elbib/rus/journal/2010/part4/ZVP.
6. *Kardashev N. S., Khartov V. V., Abramov V. V., et al.* "RadioAstron" — a telescope with a size of 300 000 km: Main parameters and first observational results // Astron. Rep., 2013. Vol. 57. Iss. 3. P. 153–194.
7. *Shustov B. M., Gomez de Castro A. I., Sachkov M., et al.* WSO-UV progress and expectations // Astrophys. Space Sci., 2014. Vol. 354. Iss. 1. P. 155–161.
8. *Кардашёв Н. С., Новиков И. Д., Лукаш В. Н. и др.* Обзор научных задач для обсерватории Миллиметрон // УФН, 2014. Т. 184. № 12. С. 1319–1352.
9. Why neuroinformatics? International Neuroinformatics Coordinating Facility. http://www.incf.org/about/why-neuroinformatics.
10. Human Brain Project. https://www.humanbrainproject.eu.
11. Human Connectome Project. WU-Minn HCP 500 Subjects Data Release: Reference manual. 2014. 166 p. http://www.humanconnectome.org/documentation/S500/HCP_S500_Release_Reference_Manual.pdf.
12. *Hawrylycz M. J., Lein E. S., Guillozet-Bongaarts A. L., et al.* An anatomically comprehensive atlas of the adult human brain transcriptome // Nature, 2012. Vol. 489. P. 391–399.
13. *Gomez-Cabrero D., Abugessaisa I., Maier D., Teschendorff A., Merkenschlager M., Gisel A., Ballestar E., Bongcam-Rudloff E., Conesa A., Tegner J.* Data integration in the era of omics: Current and future challenges // BMC Syst. Biol., 2014. Vol. 8. Suppl. 2. P. I1.
14. *Greene C. S., Tan J., Ung M., Moore J. H., Cheng C.* Big data bioinformatics // J. Cell. Physiol., 2014. Vol. 229. Iss. 12. P. 1896–1900.
15. *Herland M., Khoshgoftaar T. M., Wald R.* A review of data mining using big data in health informatics // J. Big Data, 2014. Vol. 1. Iss. 2. 35 p.
16. *Kamesh D. B. K., Neelima V., Ramya Priya R.* A review of data mining using bigdata in health informatics // Int. J. Sci. Res. Publ., 2015. Vol. 5. Iss. 3. 35 p.
17. Genome 10K community of scientists. Genome 10K: A proposal to obtain whole-genome sequence for 10 000 vertebrate species // J. Heredity, 2009. Vol. 100. Iss. 6. P. 659–674.
18. *Davis-Dusenbery B., Onder Z., Locke D., Kural D.* Petabyte-scale cancer genomics in the cloud // TCGA Symposium Oral Presentations, 2015. P. 34.
19. Materials Genome Initiative for Global Competitiveness. 2011. http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf.
20. The Materials Data Facility. http://www.nationaldataservice.org/mdf.
21. Versailles Project on Advanced Materials and Standards (VAMAS). http://www.vamas.org.
22. *Belov G. V., Iorish V. S., Yungman V. S.* IVTANTHERMO for Windows — database on thermodynamic properties and related software // CALPHAD, 1999. Vol. 23. Iss. 2. P. 173–180.
23. *Киселева Н. Н., Дударев В. А., Земсков В. С.* Компьютерные информационные ресурсы неорганической химии и материаловедения // Усп. хим., 2010. Т. 79. Вып. 2. С. 162–188.
24. Copernicus. Observing the Earth. http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview3.
25. *Ramapriyan H. K., Behnke J., Sofinowski E., Lowe D., Esfandiari M. A.* Evolution of the Earth Observing System (EOS) data and Information System (EOSDIS) // Standard-based data and information systems for Earth

observation / Eds. L. Di, H. K. Ramapriyan. — Lecture notes in geoinformation and cartography ser. — Berlin–Heidelberg: Springer, 2010. P. 63–92.

26. *Schnase J. L., Duffy D. Q., McInerney M. A., et al.* Climate analytic as a service // Conference on Big Data from Space (BiDS'14) Proceedings. — Luxembourg: Publications Office of the European Union, 2014. P. 90–93.
27. *Dubernet M. L., Boudon V., Culhane J. L., et al.* Virtual atomic and molecular data centre // J. Quant. Spectrosc. Ra. Transfer, 2010. Vol. 111. Iss. 15. P. 2151–2159.
28. *Rixon G., Dubernet M.-L., Piskunov N., et al.* VAMDC — the Virtual Atomic and Molecular Data Centre — a new way to disseminate atomic and molecular data — VAMDC Level 1 Release // J. Phys. Conf. Ser., 2011. Vol. 1344. P. 107–115.
29. National Data Service (NDS). http://www.nationaldataservice.org.
30. *Gangler E.* Big data challenge posed by the Large Synoptic Survey Telescope // Conference on Big Data from Space (BiDS'14) Proceedings. — Luxembourg: Publications Office of the European Union, 2014. P. 194–197.
31. *Frezouls B., Brunet P.-M.* Big data technology in the service of the Gaia data processing // Conference on Big Data from Space (BiDS'14) Proceedings. — Luxembourg: Publications Office of the European Union, 2014. P. 198–201.

Поступила в редакцию 02.12.15

DATA ACCESS CHALLENGES FOR DATA INTENSIVE RESEARCH IN RUSSIA

L. A. Kalinichenko[1,2], A. A. Volnova[3], E. P. Gordov[4], N. N. Kiselyova[5], D. A. Kovaleva[6], O. Yu. Malkov[6], I. G. Okladnikov[4], N. L. Podkolodnyy[7], A. S. Pozanenko[3], N. V. Ponomareva[8], S. A. Stupnikov[1], and A. Z. Fazliev[9]

[1]Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

[2]Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

[3]Space Research Institute of the Russian Academy of Sciences, 84/32 Profsoyuznaya Str., Moscow 117997, Russian Federation

[4]Siberian Center for Environmental Research and Training, Institute of Monitoring of Climatic and Ecological Systems of the Siberian Branch of the Russian Academy of Sciences, 10/3 Akademicheski Av., Tomsk 634055, Russian Federation

[5]A. A. Baikov Institute of Metallurgy and Materials Science of the Russian Academy of Sciences, 49 Leninsky Av., GSP-1, Moscow 119991, Russian Federation

[6]Institute of Astronomy of the Russian Academy of Sciences, 48 Pyatnitskaya Str., Moscow 119017, Russian Federation

[7]Center for Bioinformatics, Federal Research Research Center "Institute of Cytology and Genetics" of the Siberian Branch of the Russian Academy of Sciences, 10 Acad. Lavrentyeva Av., Novosibirsk 630090, Russian Federation

[8]Research Center of Neurology, 80 Volokolamskoe Shosse, Moscow 125367, Russian Federation

[9]Integrated Information Systems Center, Institute of Atmospheric Optics of the Siberian Branch of the Russian Academy of Sciences, 1 Acad. Zuev Sq., Tomsk 634055, Russian Federation

Abstract: The goal of this survey is to analyze the global trends of development of massive data collections and related infrastructures in the world aimed at the evaluation of the opportunities for the shared usage of such collections during research, decision making, and problem solving in various data intensive domains (DIDs) in Russia. The representative set of DIDs selected for the survey includes astronomy, genomics and proteomics, neuroscience (human brain investigation), materials science, and Earth sciences. For each of such DIDs, the strategic initiatives (or large projects) in the USA and Europe aimed at creation of big data collections and the

respective infrastructures planned up to 2025 are briefly overviewed. The information technology projects aimed at the development of the infrastructures supporting access to and analysis of such data collections are also briefly overviewed. The set of large data collections included into the survey and expected to be created soon is planned to be used as a reference point for the design and development of the research infrastructures for data management and analysis making them compatible with the foreign open research infrastructures. In particular, the data collections considered in the survey, the goals of their creation and the researches planned to be accomplished based on them make it possible to proceed to the design and implementation of the advanced components of the research infrastructures, such as, for example, conceptualization facilities of the application domains to be investigated in data intensive research, respective metamodels, components intended for data reuse and reproducing of programs and workflows, etc.

Keywords: fourth paradigm; data intensive domains; research infrastructures; data collections; big data

DOI: 10.14357/19922264160101

Acknowledgments

This survey was partially supported by different grants for groups from participating research institutes: for IPI FRC CSC RAS by RFBR grants 13-07-00579, 14-07-00548, and 16-07-01028; for IOA SB RAS by RFBR grant 13-07-00411; for IMCES SB RAS by RFBR grants 13-05-12034 and 14-05-00502; for IMET RAS by RFBR grants 14-07-00819 and 15-07-00980; for INASAN RAS by RFBR grant 15-02-04053 and by the Presidium of RAS Program P-41; for ICG SB RAS by RSF grant 14-24-00123; for RCN by RFBR grants 15-04-08744 and 15-04-05066; and for SRI (IKI) RAS by RFBR grant 15-02-10203-K.

References

1. Hey, T., S. Tansley, and K. Tolle, eds. 2009. The fourth paradigm: Data-intensive scientific discovery. Redmond, WA: Microsoft Research. 284 p. Available at: http://goo.gl/edvr6W (accessed February 1, 2016).
2. Juric, M., and T. Tyson. 2015. LSST data management: Entering the era of petascale optical astronomy. *High. Astron.* 16:675.
3. Taylor, A. R. 2015. Data intensive radio astronomy en route to the SKA: The rise of big radio data. *High. Astron.* 16:677.
4. Fleming, S. W., F. Abney, T. Donaldson, *et al.* 2015. Beyond the Prime Directive: The MAST discovery portal and high level science products. *American Astronomical Society (AAS) Meeting* #225. #336.59.
5. Zhelenkova, O., V. Vitkovsky, and T. Plyaskina. 2010. Electronic archive of observational data of astrophysical observatory. *Russ. J. Digital Libraries* 13(4). Available at: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2010/part4/ZVP (accessed February 1, 2016).
6. Kardashev, N. S., V. V. Khartov, V. V. Abramov, *et al.* "RadioAstron" — a telescope with a size of 300 000 km: Main parameters and first observational results. *Astron. Rep.* 57(3):153–194.
7. Shustov, B. M., A. I. Gomez de Castro, M. Sachkov, *et al.* 2014. WSO-UV progress and expectations. *Astrophys. Space Sci.* 354(1):155–161.
8. Kardashev, N. S., I. D. Novikov, V. N. Lukash, *et al.* 2014. Review of scientific topics for the Millimetron space observatory. *Physics-Uspekhi* 57(12):1199–1228.
9. Why neuroinformatics? International Neuroinformatics Coordinating Facility. Available at: http://www.incf.org/about/why-neuroinformatics (accessed February 1, 2016).
10. Human Brain Project. Available at: https://www.humanbrainproject.eu (accessed February 1, 2016).
11. Human Connectome Project. WU-Minn HCP 500 Subjects Data Release: Reference manual. Available at: http://goo.gl/FsfmUb (accessed February 1, 2016).
12. Hawrylycz, M. J., E. S. Lein, A. L. Guillozet-Bongaarts, *et al.* 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489:391–399.
13. Gomez-Cabrero, D., I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegner. 2014. Data integration in the era of omics: Current and future challenges. *BMC Syst. Biol.* 8(2):I1.
14. Greene, C. S., J. Tan, M. Ung, J. H. Moore, and C. Cheng. 2014. Big data bioinformatics. *J. Cell. Physiol.* 229(12):1896–1900.
15. Herland, M., T. M. Khoshgoftaar, and R. Wald. 2014. A review of data mining using big data in health informatics. *J. Big Data* 1(2). 35 p.
16. Kamesh, D. B. K., V. Neelima, and R. R. Priya. 2015. A review of data mining using bigdata in health informatics. *Int. J. Sci. Res. Publ.* 5(3).
17. Genome 10K community of scientists. 2009. Genome 10K: A proposal to obtain whole-genome sequence for 10 000 vertebrate species. *J. Heredity* 100(6):659–674.
18. Davis-Dusenbery, B., Z. Onder, D. Locke, and D. Kural. 2015. Petabyte-scale cancer genomics in the cloud. *TCGA Symposium Oral Presentations*. 34.
19. Materials Genome Initiative for Global Competitiveness. Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf (accessed February 1, 2016).

20. The Materials Data Facility. Available at: http://www.nationaldataservice.org/mdf/ (accessed February 1, 2016).
21. Versailles Project on Advanced Materials and Standards (VAMAS). Available at: http://www.vamas.org/ (accessed February 1, 2016).
22. Belov, G. V., V. S. Iorish, and V. S. Yungman. 1999. IVTANTHERMO for Windows — database on thermodynamic properties and related software. *CALPHAD* 23(2):173–180.
23. Kiselyova, N. N., V. A. Dudarev, and V. S. Zemskov. 2010. Computer information resources in inorganic chemistry and materials science. *Russ. Chem. Rev.* 79(2):145–166.
24. Copernicus. Observing the Earth. Available at: http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview3 (accessed February 1, 2016).
25. Ramapriyan, H. K., J. Behnke, E. Sofinowski, D. Lowe, and M. A. Esfandiari. 2010. Evolution of the Earth Observing System (EOS) data and Information System (EOSDIS). *Standard-based data and Information systems for Earth observation*. Eds. L. Di and H. K. Ramapriyan. Lecture notes in geoinformation and cartography ser. Berlin–Heidelberg: Springer. 63–92.
26. Schnase, J. L., D. Q. Duffy, M. A. McInerney, *et al.* 2014. Climate analytic as a service. *Conference on Big Data from Space BiDS'14 Proceedings*. Luxembourg: Publications Office of the European Union. 90–93.
27. Dubernet, M. L., V. Boudon, J. L. Culhane, *et al.* 2010. Virtual atomic and molecular data centre. *J. Quant. Spectrosc. Ra. Transfer* 111(15):2151–2159.
28. Rixon, G., M.-L. Dubernet, N. Piskunov, *et al.* 2011. VAMDC — the Virtual Atomic and Molecular Data Centre — a new way to disseminate atomic and molecular data — VAMDC Level 1 Release. *J. Phys. Conf. Ser.* 1344:107–115.
29. National Data Service (NDS). Available at: http://www.nationaldataservice.org/ (accessed February 1, 2016).
30. Gangler, E. 2014. Big data challenge posed by the Large Synoptic Survey Telescope. Big data technology in the service of the Gaia data processing. *Conference on Big Data from Space BiDS'14 Proceedings*. Luxembourg: Publications Office of the European Union. 194-197.
31. Frezouls, B., and P.-M. Brunet. 2014. Big data technology in the service of the Gaia data processing. *Conference on Big Data from Space BiDS'14 Proceedings*. Luxembourg: Publications Office of the European Union. 198–201.

Received December 2, 2015

Contributors

Kalinichenko Leonid A. (b. 1937) — Doctor of Science in physics and mathematics, professor; Head of Laboratory, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; professor, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; leonidandk@gmail.com

Volnova Alina A. (b. 1986) — scientist, Space Research Institute of the Russian Academy of Sciences, 84/32 Profsoyuznaya Str, Moscow 117997, Russian Federation; alinusss@gmail.com

Gordov Evgeny P. (b. 1946) — Doctor of Science in physics and mathematics, Head of Siberian Center for Environmental Research and Training, Institute of Monitoring of Climatic and Ecological Systems of the Siberian Branch of the Russian Academy of Sciences, 10/3 Akademicheski Av., Tomsk 634055, Russian Federation; gordov@scert.ru

Kiselyova Nadezhda N. (b. 1949) — Doctor of Science in chemistry, Head of Laboratory, A. A. Baikov Institute of Metallurgy and Materials Science of the Russian Academy of Sciences, 49 Leninsky Av., GSP-1, Moscow 119991, Russian Federation; kis@imet.ac.ru

Kovaleva Dana A. (b. 1973) — Candidate of Science (PhD) in physics and mathematics, scientist, Institute of Astronomy of the Russian Academy of Sciences, 48 Pyatnitskaya Str., Moscow 119017, Russian Federation; dana@inasan.ru

Malkov Oleg Yu. (b. 1961) — Doctor of Science in physics and mathematics, Head of Department, Institute of Astronomy of the Russian Academy of Sciences, 48 Pyatnitskaya Str., Moscow 119017, Russian Federation; malkov@inasan.ru

Okladnikov Igor G. (b. 1978) — Candidate of Science (PhD) in technology, senior scientist, Siberian Center for Environmental Research and Training, Institute of Monitoring of Climatic and Ecological Systems of the Siberian Branch of the Russian Academy of Sciences, 10/3 Akademicheski Av., Tomsk 634055, Russian Federation; oig@scert.ru

Podkolodnyy Nikolay L. (b. 1952) — Head of Center for Bioinformatics, Federal Research Research Center "Institute of Cytology and Genetics" of the Siberian Branch of the Russian Academy of Sciences, 10 Acad. Lavrentyeva Av., Novosibirsk 630090, Russian Federation; pnl@bionet.nsc.ru

Pozanenko Alexey S. (b. 1962) — Candidate of Science (PhD) in physics and mathematics, Head of Laboratory, Space Research Institute of the Russian Academy of Sciences, 84/32 Profsoyuznaya Str, Moscow, Russian Federation; apozanen@iki.rssi.ru

Ponomareva Natalya V. (b. 1956) — Doctor of Science in medicine, Head of Group and leading scientist, Research Center of Neurology, 80 Volokolamskoe Shosse, Moscow 125367, Russian Federation; ponomare@yandex.ru

Stupnikov Sergey A. — Candidate of Science (PhD) in technology, senior scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; sstupnikov@ipi.ac.ru

Fazliev Alexander Z. (b. 1953) — Candidate of Science (PhD) in physics and mathematics, Head of Integrated Information Systems Center, Institute of Atmospheric Optics of the Siberian Branch of the Russian Academy of Sciences, 1 Acad. Zuev Sq., Tomsk 634055, Russian Federation; faz@iao.ru