

Аннотация работы, подаваемой на конкурс ИКИ РАН.

1. Авторы. Л. А. Калиниченко (ИПИ РАН, ВМиК МГУ), А. А. Вольнова (ИКИ РАН), Е. П. Гордов (ИМКЭС СО РАН), Н. Н. Киселева (ИМЕТ РАН), Д. А. Ковалева (ИНАСАН), О. Ю. Малков (ИНАСАН), И. Г. Окладников (ИМЭКС СО РАН), Н. Л. Подколodный (ИЦиГ СО РАН), А. С. Позаненко (ИКИ РАН), Н. В. Пономарева (НЦ Неврологии), С. А. Ступников (ИПИ РАН), А. З. Фазлиев (ИОА СО РАН).

2. Название ПРОБЛЕМЫ ДОСТУПА К ДАННЫМ В ИССЛЕДОВАНИЯХ С ИНТЕНСИВНЫМ ИСПОЛЬЗОВАНИЕМ ДАННЫХ В РОССИИ

3. Обзор опубликован в журнале Информатика и её применения, т. 10, вып. 1, стр. 3-23, 2016 г.

4. 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, протеомику, нейронауку. 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. 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 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. Также выявляются технические, правовые и финансовые проблемы на пути обеспечения доступа ученых России в различных ОИИД к уже накопленным и ожидаемым коллекциям данных в мире, а также способности России заметным образом участвовать во вкладе в мировую сокровищницу данных, в создание соответствующих инфраструктур, методов и средств решения задач анализа данных.

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 обзоре формулируются проблемы доступа российских учёных к данным больших проектов, работа которых уже идёт или запланирована на ближайшее будущее: Обзорный широкоугольный 8-метровый телескоп (Large Synoptic Survey Telescope, LSST), Массив квадратного километра (Square Kilometer Array, SKA), Космическая обсерватория Гая

(Gaia), проект LIGO (Laser Interferometer Gravitational-Wave Observatory). Также рассматриваются проблемы обработки и анализа общедоступных коллекция данных проектов Sloan Digital Sky Survey (SDSS), NASA/IPAC Extragalactic Database (NED), Mikulski Archive for Space Telescope (MAST).

Основным отличием работы российских научных сообществ в области хранения, обработки и анализа больших массивов данных от подобных процессов в США и ЕС является децентрализованность подобной деятельности и отсутствие согласованных исследований на эту тему в различных институтах, изучающих информацию и работу с ней. В качестве результата обзора инициативный авторский коллектив учёных из различных научных сфер сформулировал необходимость организации организации 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 использования (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.