New Data Access Challenges for Data Intensive Research in Russia

© Leonid Kalinichenko[I] © Alexander Fazliev[II] © Eugene Gordov[III] © Nadezhda Kiselyova[IV]
© Dana Kovaleva[V] Oleg Malkov[V] © I. Okladnikov[III] © Nikolay Podkolodny[VI]
© Natalia Ponomareva[VII] © Alexey Pozanenko[VIII] © Sergey Stupnikov[I] © Alina Volnova[VIII]
[I] Institute of Informatics Problems, FRC CSC RAS, Moscow
[II] Institute of Atmospheric Optics, Siberian Branch of RAS, Tomsk
[III] Institute of Monitoring of Climatic and Ecological Systems, Siberian Branch of RAS, Tomsk
[IV] Institute of Metallurgy and Material Sciences of RAS, Moscow
[V] Institute of Astronomy of RAS, Moscow
[VI] Institute of Cytology and Genetics, Siberian Branch of RAS, Novosibirsk
[VII] Research Center of Neurology , Moscow
[VIII] Space Research Institute, RAS, Moscow
leonidandk@gmail.com faz@iao.ru gordov@scert.ru kis@imet.ac.ru
dana@inasan.ru malkov@inasan.ru igor.okladnikov@gmail.com pnl@bionet.nsc.ru
ponomare@yandex.ru apozanen@iki.rssi.ru sstupnikov@ipiran.ru alinusss@gmail.com

Abstract

The goal of this survey is to analyze the global trends for 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 DID the strategic initiatives (or large projects) in USA and Europe aimed at creation of big data collections and the respective infrastructures planned up to 2025 are briefly overviewed. The IT projects aimed at the development of the infrastructures supporting access to and analysis of such data collections are also briefly overviewed. The paper concludes with an idea of organizing in Russia of a target interdisciplinary program for the development of the pilot project of the distributed infrastructure and platform for the access to various kinds of data in the world, storage of data and their analysis during research in various DIDs. As a part of such infrastructure, the program should also include development of the high performance interdisciplinary center for data intensive applications support in various DIDs. This survey is intended also to serve as a basis for the panel discussion at the International Conference DAMDID/RCDL'2015.

Proceedings of the XVII International Conference DAMDID/RCDL'2015 «Data Analysis and Management in Data Intensive Domains», Obninsk, October 13-16, 2015

1 Introduction

Nowadays scientific research and decision making in various areas of human activity are provided on the basis of data analysis. Volume and variety of data accumulated in the respective domains grow exponentially.

In the book *The Fourth Paradigm – Data Intensive Scientific Discovery* [30] "data intensive sciences" are introduced by referring to ideas formulated by Jim Gray in 2007. Gray distinguishes 4 paradigms of scientific research in the order of their historical appearance: Empirical Science (describing natural phenomena), Theoretical Science (using models to achieve generalizations), Computational Science (simulating complex phenomena) and Data exploration (also unifying theory, experiment and simulation).

According to the Fourth Paradigm data intensive research is an integral part of various areas of science, economics, business. These areas are designated below as *data intensive domains* (DIDs).

Research and development in DIDs are inconceivable without new data obtained through observations and measurements in the nature and society. Process of data extraction, processing and analysis resembles process of mineral mining operations. Minerals are mined, processed and transformed into materials for development of different products. Similarly to minerals data are extracted from the nature for observable phenomena and processes. Data extraction from the nature becomes more and more complicated and sophisticated alongside with the

development of knowledge. This complexity is motivated by the increasing scale of micro and macro phenomena to be investigated. Global projects and missions (including space ones) aimed at data extraction and accumulating with the help of the newest specialized high-technological instruments located on the Earth and in space are organized. Data extraction is very costly process requiring development of specific technologies and huge investments. The process of data extraction during investigation of some kind of phenomena in DID can take many years. Result of data extraction is raw data ("ore") that have to be processed and analyzed. Alongside with the data extraction development the rapid advancements and expansion take place in the following areas:

- methods and tools for data accumulating, processing, analysis and management in different DIDs;
- variety of problems to be solved on the basis of extracted data;
- accumulation of experience of solving of such problems and interdisciplinary usage of their solutions.

Considerations mentioned above motivate the authors to analyze global trends for development of massive data collections in the world and opportunities for the shared usage of such collections during research, decision making and problem solving in various DIDs in Russia.

The main motivation for this work is to initiate systematic analysis of the following topics attracting significant interest in the world:

- development of massive data collections in various DIDs;
- development of infrastructures for accumulating and usage of massive data collections;
- systematizing of experience of problem solving in DIDs; etc.

Some of the pragmatic aims of this analysis include:

- revealing technical, legal and financial issues of access of Russian scientists in various DIDs to accumulated and expected data collections in the world[1];
- determining needs for specific hardware and software infrastructures for access to massive data collections from Russia;
- determining possibilities for Russia to contribute to the "world data treasury", to the creation of infrastructures, methods and tools for data analysis and problem solving.

Preliminary analysis shows that the Western world is highly concerned with the issues caused by the "flood" of DIDs with the big data. The issues include data analysis (for instance, natural language texts analysis as a part of cognition), data accumulation and interdisciplinary usage, design of specific infrastructures aimed at handling with "data flood" caused by installing new facilities and instruments of big data extraction and processing in nearly real time. A lot of activities are undertaken in this direction including organization of joint projects, workgroups and their symposia, conferences, discussion of possible solutions; design and development of new infrastructures; design and testing right now of the fragments of infrastructures oriented on access and analysis of data collections planned to start functioning after 2020; developing of use cases of the problems to be solved.

Situation in Russia is such that without timely and effective access to data (the most important data collections are accumulated abroad) scientific research in various areas of many DIDs will become less and less efficient.

Data collections overviewed in this paper and examples of their usage are time limited: large projects intended for data accumulation and usage in various DIDs carried out up to 2025 are considered. The representative set of DIDs including astronomy, genomics and proteomics, neuroscience (human brain investigation), materials science and Earth sciences was predefined[2] for our overview and analysis. This set includes also the informatics accompanied with a collection of existing data analysis methods (machine learning, data mining, statistics), software and hardware platforms.

Researchers from different institutions of RAS, SB RAS and RAMS participated in this work. The list of institutions includes Federal Research Center "Computer Science and Control", Space Research Institute, Institute of Astronomy, Institute of Cytology and Genetics, Research Center of Neurology, Institute of Metallurgy and Materials Science, Institute of Monitoring of Climatic and Ecological Systems, Institute of Atmospheric Optics.

The paper is structured as follows. For every DID considered in the next sections the following information is provided:

- large strategic initiatives in USA and Europe;
- examples of massive data collections in the world up to 2025;
- known infrastructure projects and data centers;
- comparable projects in Russia..

Selecting initiatives, collections and infrastructures, the authors attempted to choose the projects collecting unique data that are important for scientific research in

[1] Data extraction in many DIDs as well as acquisition of concrete data collections possess high technological complexity and cost. "Import substitution" is out of the question for at least 10 years in this area.

[2] This set has been formed during preparation of the DAMDID/RCDL'2015 conference. In the sequel the set can be enlarged. Physics is not included in the set intentionally. E.g., methods and tools for analysis of data got from hadron collider evolve very fast. The LHC project occupies a distinguished position in Russia, data to be processed and infrastructure are very specific.

Russia. Volume of such collections should be at least dozens of TB. The projects were selected in which access to data within the projects requires to overcome technical, legal and financial issues.

In the last section several western IT-projects are overviewed. The projects are aimed at research and development of infrastructures for access and analysis of data that are planned to be collected several years later..

The paper is concluded with a general evaluation of the state-of-art of data access in data intensive research in Russia as well as with the recommended activities to be performed.

2 Data Collections in Astronomy

Large data collections in astronomy are usually generated either by large observational units (ground-based or space observatories), or by successful large-scale observational projects (similarly, ground- or space-based), or data of specific kind may be gathered by some thematic authorities. Often these data collections are supplied with data interfaces, processing software, other user-support tools.

The typical classes of problems to be solved using such extensive data collections are:

- search for relations predicted by the theory;
- testing theoretical hypotheses;
- creation of datasets having certain characteristics to compare with modeling results;
- search for anomalous objects and events.

Data use treatment can be different. Typically, in astronomy proprietary period while data are available exclusively to observing team is about one year, sometimes less or more. This is reflected in access policy of data collections. Such restrictions prevent those who wish to deal with frontier astronomical large-scale raw data.

2.1 Global Astronomical Missions

Though the examples of projects included into this subsection formally do not belong to the national or world-wide initiatives, by their scale, objectives, expected scientific significance and influence they deserve to take similar position. Entering such projects would open unexplored and very fruitful prospects for the Russian astronomers as well as for computer scientists.

2.1.1 Large Synoptic Survey Telescope (LSST)

With the advent of the Large Synoptic Survey Telescope (LSST, http://www.lsst.org/lsst/about/technology) which will take its first light in 2020, the data flow in optical astronomy will be put to an unprecedented rate. LSST will not only record around 40 billion of objects, but also one of the key features of LSST will be to open the time domain for massive studies of variable sources and rare events: each object will be observed about 1000 times.

The LSST design is driven by four main science themes: probing dark energy and dark matter, taking an inventory of the Solar System, exploring the transient optical sky, and mapping the Milky Way. The current baseline [41], with an 8.4m (6.7m effective) large, high-precision primary mirror, a 9.6 deg^2 field of view, and a 3.2 Gigapixel camera, will allow about 10,000 deg^2 of sky to be covered using pairs of 15-second exposures twice per night every three nights on average. The system is designed to yield high image quality as well as superb astrometric and photometric accuracy. The total survey area will include 25,000 deg^2 and will be imaged multiple times in six bands (ugrizy), covering the wavelength range 320 – 1050 nm. About 90% of the observing time will be devoted to a deep-wide-fast survey mode which will uniformly observe a 18,000 deg^2 contiguous sky region on average 825 times (summed over all six bands) during the anticipated 10 years of operations. These data will serve the majority of the primary science programs.

The rapid cadence of the LSST observations will produce an enormous volume of data, ~15 TB per night, leading to a total data volume for 10 years of 60 PB compressed raw data and 30 PB for the catalog database, a total of 0.5 EB processed data, which will need ~150 TFLOPS total computing power. This corresponds to a final catalog containing 20 billion galaxies and 17 billion stars, with 7 trillion source detection and a total of 30 trillion measurements; the corresponding volume of the final database is expected to be 15 PB. Processing such a large volume of data, converting the raw images into a faithful representation of the Universe, and archiving the results in useful form for a broad community of users is a major challenge.

Two main deliverables are planned to make widely available: the transient event reporting system which will send out alerts to the community within 60 seconds of completing the image readout and yearly data releases which will deliver the most completely analyzed data products of the survey. To make the LSST scientific data available world-wide, the LSST team is working with foreign institutions and governments to share the costs. Institutions joining LSST early will have the customary advantage of deep familiarity with the LSST system and survey. Early collaboration of institutions worldwide will also assure access to additional local computational capability to efficiently search and run scripts on the 30PB database and undertake calculations on the 100PB of images [35]. More details on the CNRS/IN2P3 project aimed at such issues are included into the 7.3.1 section. The Russian scientists are not involved into the LSST community yet.

2.1.2 Square Kilometer Array

Square Kilometer Array (SKA, https://www.skatelescope.org/software-and-computing/) is the most ambitious project in radio astronomy. This radio telescope is to be built in Australia and South

Africa and would have a total collecting area, with thousands of dishes, of approximately one square kilometer. It will operate over a wide range of frequencies, its size will make it 50 times more sensitive than any other radio instrument, and will be able to survey the sky more than ten thousand times faster than ever before. Construction of the SKA is scheduled to begin in 2018 for initial observations by 2020, but the construction budget is not secured at this stage. The SKA will be built in two phases, with Phase 1 (2018-2023) representing about 10% of the capability of the whole telescope. The headquarters of the project are located at the Jodrell Bank Observatory, in the UK.

The computing requirements of the SKA will exceed those of the fastest supercomputers available in 2015, whilst the data processing and amounts of data will compete with that generated by the entire Internet, facilitating the need for a new kind of high speed network. The amount of sensory information collected pose a huge problem in data storing and require real-time signal processing to reduce the information to relevant data. It was estimated the array could generate an EB a day of raw data which could be compressed to around 10 PB [64].

Organizations from eleven countries are currently members of the SKA Organization – Australia, Canada, China, Germany, India, Italy, New Zealand, South Africa, Sweden, the Netherlands and the United Kingdom. Further countries have expressed their interest in joining the SKA Organization which will continue to expand over the coming years. Russia is not in the list.

2.1.3 Gaia Mission

Gaia mission (http://www.cosmos.esa.int/web/gaia/release) is a space observatory of the European Space Agency (ESA) designed for astrometry. The mission aims to construct a 3D space catalog of approximately 1 billion astronomical objects, mainly stars. Additionally Gaia is expected to detect thousands to tens of thousands of Jupiter-sized planets beyond the Solar System (exoplanets), 500,000 quasars and tens of thousands of new asteroids and comets within the Solar System. The spacecraft will monitor each of its target stars about 70 times over a period of five years. Gaia will create a precise three-dimensional map of astronomical objects throughout the Milky Way and map their motions, which encode the origin and subsequent evolution of the Milky Way. Gaia is expected to become a milestone of the new view to the Milky Way and the key to many fundamental astronomical problems. Gaia was launched on 19 December 2013. The overall data volume that will be retrieved from the spacecraft during the five-year mission assuming a nominal compressed data rate of 1 Mbit/s is approximately 60 TB, amounting to about 200 TB of usable uncompressed data, stored in the InterSystems Cachè database.

The responsibility of the data processing, partly funded by ESA, has been entrusted to a European consortium (the Data Processing and Analysis Consortium, or DPAC) which has been selected after its proposal to the ESA Announcement of Opportunity released in November 2006. DPAC's funding is provided by the participating countries and has been secured until the production of Gaia's final catalogue scheduled for 2020. Gaia will generate of 40 GB of information flow per day, or 1 PB over the full life of the mission. DPAC brought together more than 400 specialists from throughout the ESA community to provide the required scientific expertise. It will remain in place around 3 years after the end of the mission, up to release of the final product: the Gaia catalogue. Access to raw data is limited to DPAC members, who will be entrusted with the task of converting the telemetry data into scientifically meaningful information. This information will be released in the form of catalogues that will be made available to the worldwide astronomical community [65].

2.1.4 Gravitational Wave Astronomy

LIGO (http://www.ligo.caltech.edu/) as well as Virgo (http://www.mpa-garching.mpg.de/galform/virgo/) projects are aimed for the first time detection of gravitational waves. LIGO Data System contains more than 1 PB of raw data. Archive data are publically accessible, and the S5 Data Release covers data accumulated in 2005-2007. It contains mostly data series in different channels and masks of data quality. A real-time alerting system (Rapid Triggers from LIGO Data) is supporting in a framework the LIGO/Virgo projects [1]. The main goal of the alerting system is detection and crude localization of possible transient sources of gravitational waves and rapid distribution of type of detection and localization areas for astronomical observatories involved in follow-up observation in electromagnetic bandwidth, mostly in optic.

2.2. Data Collections Freely Available for the World Community

Several examples are included into this subsection.

One of the leading astronomical observational projects of recent years, the project with unexampled (number of publications)/(project value) ratio, the Sloan Digital Sky Survey (SDSS) http://www.sdss.org/ has been working for more than 15 years to make a map of the Universe, and will continue for many years to come (SDSS-IV project is planned for 2014-2020). Every night the dedicated 2.5-m wide-angle optical telescope produces about 200 GB of multicolor photometric survey and spectroscopic data about various kinds of objects in Universe, including galaxies, supernovae, stellar population, exoplanet survey, etc. Before the data are released, they are used and processed by SDSS Institution members. Some DRs are for SDSS collaboration use only. The SDSS-IV collaboration is still growing, including a number of institutions from the US, Germany, China, Japan, France, UK, Canada, Korea, Israel, Brazil (but none from Russia).

NED: NASA/IPAC Extragalactic Database (https://ned.ipac.caltech.edu/ui/) is a comprehensive database of multiwavelength data for extragalactic objects, providing a systematic, ongoing fusion of information integrated from hundreds of large sky surveys and tens of thousands of research publications. The contents and services span the entire observed spectrum from gamma rays through radio frequencies. As new observations are published, they are cross-identified or statistically associated with previous data and integrated into a unified database to simplify queries and retrieval. Seamless connectivity is also provided to public data in NASA astrophysics mission archives (NASA/IPAC Infrared Science Archive IRSA, NASA's High Energy Astrophysics Science Archive Research Center HEASARC, Mikulski Archive for Space Telescopes MAST (see below in this text), to the astrophysics literature via the SAO/NASA Astrophysics Data System ADS, and to other data centers around the world [48]. The approximate volume of data in the NED database may be at least 20 TB, judging by volume of data of major incorporated surveys.

Perhaps the most prominent archive of, primarily, space telescope data in the optical, ultraviolet, and near-infrared parts of the spectrum, the Mikulski Archive for Space Telescopes (MAST, https://archive.stsci.edu/index.html) is a NASA funded project to support and provide to the astronomical community a variety of astronomical data archives. The primary focus of the MAST remains, however, the archive of scientific data obtained with the Hubble Space Telescope. The archive includes also the data of such keystone missions as Kepler, IUE, GALEX etc. MAST is located at the Space Telescope Science Institute (STScI). Recently the volume of data collection exceeded hundreds of TB. Data retrieved from the MAST are part of the public domain and need no explicit permission for use, with the exception of proprietary data. These proprietary data are only provided to those who have proprietary rights (i.e. to the participants of relevant observational projects, including Russian scientists if their observational proposal won the competition and has been approved). No explicit permission is needed for the use of MAST tools [22].

The ESO (European South Observatory) Science Archive http://archive.eso.org/cms.html is perhaps the most extensive and well-organized collection of data of one of the largest and most successful ground observatory. It contains data from ESO telescopes at La Silla Paranal Observatory, including the APEX submillimeter telescope on Llano de Chajnantor. In addition, the raw UKIDSS/WFCAM data obtained at the UK Infrared Telescope facility in Hawaii are available, as well as advanced ESO data products. The monthly inflow of data is about 7~8 TB, while the total volume of the archive have exceeded a hundred of TB several years ago. The Principal Investigators of successful proposals for time on ESO telescopes have exclusive access to their scientific data for the duration of a proprietary period, normally of one year, after which the data becomes available to the community at large. All data in the ESO archive remains the property of ESO. ESO reserves the right to restrict or enhance access to the data at any time. The public partition of the ESO archive is accessible to all registered ESO users from within the international astronomical community. There were official steps taken by Russia to enter the ESO community, however it came to nothing so far, due to financial problems and conflict of interests [23].

The GCN project (Gamma-ray Coordinates Network is evolving now to TAN: Transient Astronomy Network), is the first successful project for real time GRB (Gamma-ray bursts) alert distribution. The project is fully open for any subscriber, and has different interfaces for real-time data distribution. The GCN is an important infrastructure project for emerging transient astronomy and high energy astrophysics [5]. The project continues to collect references for all interesting GRB discovered since 1991.

2.3 Examples of Astronomical Missions and Data Collections in Russia

The closest analogue of the ESO archive in Russia is the General Observation Archive of the Special Astrophysical Observatory of the Russian Academy of Science, the largest Russian observational facility (http://www.sao.ru/oasis/cgi-bin/fetch?lang=en). The archive (named OASIS) evidently does not contain so large data volume (back in 2010 it was 250 GB) however still represents the largest Russian observatory data archive providing data interface [72].

In Russia, there are three planned space projects involving relatively large flow of data (expected volume of data is dozens of TB). These are Spectrum-X-Gamma (an international high-energy astrophysics observatory, which is being built under the leadership of the Russian Space Research Institute (http://hea.iki.rssi.ru/SRG/en/index.php) to be launched in 2016 [37], WSO-UV - a major international collaboration with Russia playing the leading role to study the Universe in the 115-310 nm ultraviolet (UV) wavelengths range, which is beyond the reach of ground-based instruments (http://www.wso-uv.org/wso-uv2/index.php?lang=en) - to be launched in 2021 [61] and even more ambitious Millimetron project for sub millimeter observations [73], the predecessor of which is successfully operating in orbit since 2011 Radioastron [27].

2.4 Infrastructure Projects

An example of a completely open source of large-scale astronomical data is represented by a brilliant infrastructure project of the Strasbourg astronomical Data Center, CDS (http://cds.u-strasbg.fr/) including such high scientific level astronomical data products as stellar database SIMBAD (http://simbad.u-strasbg.fr/simbad/), the most complete and popular database of astronomical catalogues VizieR (http://vizier.u-strasbg.fr/viz-bin/VizieR), interactive sky atlas Aladin (http://aladin.u-strasbg.fr/aladin.gml), etc. [55].

Another example is the Astronomical Wide-field Imaging System for Europe (Astro-WISE, http://www.astro-wise.org/portal/index.shtml). This is an environment consisting of hardware and software which is federated over about a dozen institutes all over Europe. Astro-WISE started out as a system geared towards astronomy, but is now also being used for projects outside astronomy. Astro-WISE is an all-in-one system: it allows a scientist to archive raw data, calibrate data, perform post-calibration scientific analysis and archive all results in one environment. The system architecture links together all these commonly discrete steps in data analysis, including the input, output, and the software code used, for arbitrary data volumes. The Astro-Wise system will be connected to the Virtual Observatory via the Euro-VO. At the moment, researchers use the system with 10's TB of image data. 100s of TB of data will start entering the system when the OmegaCAM camera starts operations in Chile. Several large surveys plan to use Astro-Wise. The information system online storage is presently 1.6 PB. It contains scientific data of dozens of projects, some of them are public and some are accessible only to a given project participants [6].

3 Data in Brain Research

Neuroscience is a scientific discipline that studies the structure, function, development, genetics, biochemistry, physiology, pharmacology, pathology of the nervous system, and psychology. Neuroscience is at the frontier of investigation of the brain and mind. The study of the brain is becoming the cornerstone in understanding how we perceive and interact with the external world and, in particular, how human experience and human biology influence each other. It is likely that the study of the brain will become one of the central intellectual endeavors in the coming decades. In this survey we limit ourselves to the area of brain investigation.

The amount of data that can be generated in a typical neuroscience lab grows at an amazing rate. But the collected knowledge about the brain lags behind, since condensing all these data sets into a coherent picture is an increasingly complex task. That is why neuroscience needs neuroinformatics; the collaboration between neuroscientists and computer scientists to make both new and old neuroscience data more accessible and more useful to the research community, and to advance our understanding of the brain at a much faster rate than previously possible [71]. The following strategic initiatives and projects are examples of important endeavors being developed to advance studying of brain as the *most complicated biological structure in the known universe*.

3.1 Strategic Initiatives in Brain Research

3.1.1 Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative

BRAIN initiative has been announced by the White House in April 2013 as a 12 years research strategy. National Institute of Health (NIH) is one of several federal agencies involved in the BRAIN Initiative. Planning for the NIH component of the BRAIN initiative is guided by the long-term scientific plan, “BRAIN 2025: A Scientific Vision” [12], which details main high-priority research areas and calls for a sustained federal commitment of $4.5 billion over 12 years. Recently besides NIH, DARPA and NSF BRAIN initiative was joined by the federal agencies FDA and IARPA.

BRAIN aims at creating for the first time a dynamic understanding of brain function showing how individual cells and complex neural circuits interact in both time and space, in health and disease. The analysis of circuits of interacting neurons is aimed at identifying and characterizing the component cells, defining their synaptic connections with one another, observing their dynamic patterns of activity as the circuit functions in vivo during behavior, and perturbing these patterns to test their significance. It aims also at understanding of the algorithms that govern information processing within a circuit and between interacting circuits in the brain as a whole.

It is expected to produce conceptual foundations for understanding the biological basis of mental processes through development of new theoretical and data analysis tools. Rigorous theory, modeling, and statistics are advancing the understanding of complex, nonlinear brain functions where human intuition fails.

The mammalian brain contains ~108 (mouse) - 1011 (human) neurons. These neurons are not homogeneous, but consist of diverse subpopulations with genetically, anatomically, physiologically, and connectionally distinct properties. Defining these cell types and their functions in different brain regions, and providing methods for experimental access to them in a variety of animals and in humans is essential to generating a comprehensive understanding of neural circuit function.

There is a need to improve spatial resolution and/or temporal sampling of human brain imaging techniques, and develop a better understanding of cellular mechanisms underlying commonly measured human brain signals (fMRI, DW MRI, EEG, MEG, PET) —for example, by linking fMRI signals to cellular-resolution population activity of neurons and glia contained within the imaged voxel, or by linking DW MRI connectivity information to axonal anatomy. Understanding these links will permit more effective use of clinical tools for manipulating circuit activity, such as deep brain stimulation and transcranial magnetic stimulation (TMS).

Well-curated, public data platforms with common data standards, seamless user accessibility, and central maintenance would make it possible to preserve, compare, and reanalyze valuable data sets that have been collected at great expense. Valuable lessons and best-practices can be learned from existing public datasets, which include the Allen Brain Atlas, the Mouse Connectome Project, the Open Connectome Project, the CRCNS data sharing project, ModelDB, and

the Human Connectome Project, as well as datasets generated by the physics, astronomy, climate science, and technology communities. A first unifying attempt, the Neuroscience Information Framework (NIF) sponsored by NIH, provides a portal to track and coordinate multiple sites, but the myriad genetic, anatomical, physiological, behavioral and computational datasets are difficult to manage because of their heterogeneous nature. The NIH Big Data to Knowledge (BD2K) Initiative offers opportunities to neuroscientists to develop new standards and approaches. Brain related research is provided by the following BD2K centers of excellence: Big Data for Discovery Science Center, Center for Causal Modeling and Discovery of Biomedical Knowledge from Big Data, ENIGMA Center for Worldwide Medicine, Imaging and Genomics.

BRAIN Initiative will require infrastructure for integrating and sharing relevant datasets and data analysis methods. The infrastructure for maintaining common databases will require dedicated resources, which may be provided by the NIH BD2K Initiative or the NIH Blueprint for Neuroscience Research to support the BRAIN initiative.

Google is building the software tools and supporting infrastructure needed to analyze PB-scale datasets generated by the BRAIN Initiative and the neuroscience community to better understand the brain's computational circuitry and the neural basis for human cognition. Google is working closely with the Allen Institute for Brain Science to develop scalable computational solutions to advance scientific understanding of the brain [21].

3.1.2 Human Brain Project (HBP) of the EU

In 2013, the European Union announced the Human Brain Project (HBP), a €1 billion project with an emphasis on information computing technology infrastructure for neuroscience [31]. The HBP is a EC Flagship that aims to accelerate our understanding of the human brain, make advances in defining and diagnosing brain disorders, and develop new brain-like technologies.

The HBP is organized in thirteen subprojects, spanning strategic neuroscience data, cognitive architectures, theory, ethics and society, management and the development of six new informatics-based platforms including:

- Neuroinformatics (searchable atlases and analysis of brain data);
- Brain Simulation (building and simulating multi-level models of brain circuits and functions);
- Medical Informatics (analyzing clinical data to better understand brain diseases);
- Neuromorphic Computing (brain-like functions implemented in hardware);
- Neurorobotics (testing brain models and simulations in virtual environments);
- High Performance Computing (providing the necessary computing power).

HBP is expected to become a major driver of ICT in Europe. A key goal of HBP is to construct realistic simulations of the human brain – this will require molecular and cellular information and from that it will be possible to model and understand biological and medical processes. In addition, this will make possible to use that information to design and implement new kinds of computers and robotics, to translate the results obtained into technology (neuromorphic processors).

It is planned to build data driven models that capture what have been learned about the brain experimentally: its deep mechanics (the bottom up approach) and the basic principles it uses in cognition (the top-down approach). The brain models will be developed with learning rules that are as close as possible to the actual rules used by the brain and couple such models to virtual robots that interact with virtual environments. It is expected that such models will learn the same way the brain learns and that they will exhibit the same kind of intelligent behaviour.

Potentially it is possible to build models that are orders of magnitude larger than the models we can actually simulate on a given supercomputer. For instance, in the next few years it is planned to build very detailed molecular-level models that would require *hundreds of EB of memory to simulate*. But even in 2020, it is expected that supercomputers will have no more than 200 PB. Therefore it is planned to build fast random-access storage systems next to the supercomputer, store the complete detailed model there, and then allow the multi-scale simulation software to call in a mix of detailed or simplified models (models of neurons, synapses, circuits, and brain regions) that matches the needs of the research and the available computing power. It is expected that such pragmatic strategy will allow to build more detailed models, while keeping the simulations to the level of detail that can be supported with the current supercomputers.

The major obstacle that hinders our understanding the brain is the fragmentation of brain research and the data it produces. The most urgent need is thus a concerted international effort that can integrate this data in a unified picture of the brain as a single multi-level system. It is planned to build an integrated system of ICT-based research platforms, which without resolving all open problems, would allow neuroscientists, medical researchers and technology developers to accelerate the pace of their research.

Initially the HPC facilities at 4 centers can be used for the purposes of the HBP: Jülich (6 PFLOPS peak, 450 TB memory, 8 PB scratch file system) allowing simulations up to 100 Mio neurons (scale of mouse brain), Swiss CSCS (836 TFLOPS peak, 64 T, 4 PB) in particular for software development and optimization, Barcelona SC (1 PFLOPS peak, 100 TB) for molecular-level simulations, CINECA (2 PFLOPS, 200 TB, 5 PB) mainly for data analytics. In addition KIT Karlsruhe provides 3 PB of storage. All centers are linked with 10

Gbit/s. In the neuromorphic area SpiNNaker chips are being used that have 18 cores and share 128 MB RAM allowing to simulate 16.000 neurons with 8 Mio plastic synapses with 1 W energy budget.

3.2 Data Collections in Neuroscience

3.2.1 Human Connectome Project (HCP)

Structural (anatomical) connectivity of brain can be mapped at several levels, which have been termed "macro-(at the level of centimeters to millimeters), meso-(at millimeters to micron scale), and micro-scale (at micron to nanometer resolution)". If to store a human genome it is required several GB, to store connectome on the subcellular level we would need trillion GB. Current human connectomic efforts are only at a macro-scale, they are based on noninvasive MRI-based methods. DW MRI exploits the anisotropy of water diffusion along myelinated axons to map white matter tracts. fMRI measures correlations in spontaneous activity across areas in resting subjects. It provides an indirect measure of anatomical connectivity based on the assumption that correlations are greatest amongst connected regions.

The Human Connectome Project (HCP) [32] is an NIH-funded effort started in 2010 and planned to be completed in 2015 that is aimed to chart a comprehensive map of neuronal connections and its variability in healthy adults (on the macro-scale). The resulting data are freely available to the scientific community using a powerful, user-friendly informatics platform. The HCP has developed customized MRI equipment to comfortably image the large subject population. All subjects in the main cohort are being scanned on a dedicated 3 Tesla (3T) scanner optimized for HARDI (High Angular Resolution Diffusion Imaging) and for R-fMRI (Resting-state fMRI), and T-fMRI (task-evoked fMRI). A subset of subjects will also be scanned using 7T MRI and, possibly, an additional set of subjects at 10.5T. The HCP aims to study 1,200 subjects studied by the fall of 2015. Participants from families with twins and non-twin siblings are being scanned on the same equipment using the same protocol for every subject. Subsets of 50-100 same-sex twin pairs are undergoing additional scans using resting-state/task magnetoencephalography (MEG) combined with electroencephalography (EEG).

A robust and reliable platform to provide a stable, well-structured database for storing vast amounts of HCP data is used. A powerful supercomputer with dedicated time for the HCP is enabling complex analyses of connectivity data and other analyses to be carried out efficiently and incorporated into this database. A user-friendly platform for data mining, analysis, and visualization is being built to enable investigators around the world to capitalize on these enormously rich datasets. The total volume of data in the existing and planned data releases have an order of several dozens of TB. The HCP data can be downloaded from ConnectomeDB or by ordering pre-formatted hard disc drives.

3.2.2 Neuroscience Databases

A number of online neuroscience databases are available which provide information regarding gene expression, neurons, macroscopic brain structure, and neurological or psychiatric disorders. Some databases contain descriptive and numerical data, some to brain function, others offer access to "raw" imaging data, such as postmortem brain sections or 3D MRI and fMRI images. Some focus on the human brain, others on non-human.

As the number of databases that seek to disseminate information about the structure, development and function of the brain has grown, so has the need to collate these resources themselves. As a result, there now exist databases of neuroscience databases, some of which reach over 3000 entries. E.g., the Neuroscience Information Framework (NIF) is a dynamic inventory of Web-based neuroscience resources: data, materials, and tools accessible via any computer connected to the Internet. This is a meta database of neuroscience-relevant data incorporating over 100 databases [53].

Another kind of meta database in neuroscience is Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) [51] that facilitates finding and comparing neuroimaging resources for functional and structural neuroimaging analyses. Now NITRC offers a cloud-based computing environment to help researchers manage and analyze large volumes of brain imaging data. NITRC provides enhanced services such as virtual computing and data storage; and broaden the range of scientific domains from MR to PET, SPECT, CT, MEG/EEG, optical imaging, and digital atlasing, genetic imaging, clinical neuroinformatics, computational neuroscience, electrophysiology, computational neuroscience, and neuroimaging genomics and genetics.

Several examples of the neuroscience databases are shown below.

Allen Brain Atlases. The Allen Mouse and Human Brain Atlases [28] are projects within the Allen Institute for Brain science which seek to combine genomics with neuroanatomy by creating gene expression maps for the mouse and human brain. They will help advance various fields of science, especially those surrounding the understanding of neurobiological diseases. The atlases are free and available for public use online. The different types of cells in the central nervous system originate from varying gene expression. A map of gene expression in the brain allows researchers to correlate forms and functions. The Allen Brain Atlas lets researchers view the areas of differing expression in the brain which enables the viewing of neural connections throughout the brain. Viewing these pathways through differing gene expression as well as functional imaging techniques permits researchers to correlate between gene expression, cell types, and pathway function in relation to behaviors or phenotypes. The atlas can show which genes and particular areas are effected in neurological disorders; the action of a gene in a disease can be evaluated in conjunction with general expression

patterns and this data could shed light on the role of the particular gene in the disorder.

Open Access Series of Imaging Studies (OASIS). OASIS [54] is a project aimed at making MRI data sets of the brain freely available to the scientific community. By compiling and freely distributing MRI data sets, it becomes possible to facilitate future discoveries in basic and clinical neuroscience. OASIS is made available by the Washington University Alzheimer's Disease Research Center, Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN).

Neural ElectroMagnetic Ontologies (NEMO). NEMO is an NIH funded project that aims to create EEG and MEG ontologies and ontology based tools. These resources will be used to support representation, classification, and meta-analysis of brain electromagnetic data. The three pillars of NEMO are: DATA, ONTOLOGY, and DATABASE. NEMO data consist of raw EEG, averaged EEG, event-related brain potentials (ERPs), and ERP data analysis results. NEMO ontologies [50] include concepts related to ERP data (including spatial and temporal features of ERP patterns and properties), data provenance, and the cognitive and linguistic paradigms that were used to collect the data. The NEMO database portal is a large repository that stores NEMO consortium data, data analysis results, and data provenance.

3.3 Data in Neuroscience Research in Russia

Several Russian Research Centers (Research Center of Neurology, IHNA&NPh RAS, IHB RAS, and others) have accumulated large collections of data on microscopic and macroscopic brain structure, brain genetics and chemistry, structural and functional MRI , EEG, and evoked potentials in normal development, aging as well as in neurological and psychiatric diseases. At present most of these databases are available in the research centers where they were developed and in those research centers with which they cooperate. The development of open access collections would be a useful next step for facilitating effective research in this area.

4 Data in Genomics and Proteomics

Molecular genetics in the post-genomic era is characterized by the appearance of qualitatively new opportunities in research based on the use of high-performance experimental technologies, such as massive-parallel DNA sequencing, multilocus genotyping, multi-parameter gene expression profiling using a DNA chip, ChIP-chip or ChIP-seq technologies, proteomics and metabolomics technologies and others. Rapid improvement of high-performance experimental technologies for genomics, transcriptomics, proteomics, metabolomics, molecular genetics, cell biology, biomedicine and other experimental approaches led to unprecedented massive amounts of experimental data and knowledge [26]. These methods used are for the comparative analysis of genomes, searching for genetic variations and biomarker applications in the field of biotechnology, agriculture, pharmacology, clinical research, forensic, etc.

Now there are more 7400 high-throughput genome sequencers in operation situated in 1027 centers all over the world. In Russia, only 14 high-performance genome sequencers in the 6 sequencing centers are used. Therefore, most of the data accumulates in the foreign centers sequencing, in particular, in the US, Europe, China, South Korea, etc.

A huge amount and complexity of molecular biological information requires the use of new technology for processing massive data and integrating to uncover large hidden values from large datasets that are diverse, complex, and of a large scale. The absence of special technologies for searching, sharing, storage, curation, transfer, integration, computing and visualization of massive data, and special analytical software make it difficult to analyze, systematize and apply big data for specific tasks of bioinformatics, biotechnology, agriculture, pharmacology, clinical and personalized medicine, and others [25][36][29][39].

Alongside with the variety of the heterogeneous and distributed data in bioinfomatics, their variability and inconsistency, veracity of data (quality of the data being captured influencing on the accuracy of analysis), multimodal, multilevel and multiscale nature of systems in medicine and biology, we can specifically highlight the volume and velocity features of the data in the field of bioinformatics and computer systems biology, which require new approaches to handle massive data:

- A huge volume of experimental data. There are estimations of total genomic data on all projects, which indicate that by 2018 the annual operating volume of new data can reach more than 3300 PB, one third of which is the original sequence data and two thirds are the expansion of data due to their interpretation and analysis (Fig. 1).
- High-speed data accumulation and data update. The amount of data, including the results of sequencing the genomes is tripled annually. The large growth in the number of heterogeneous and distributed data sources is also an important feature of the subject area. In particular, the NAR online Molecular Biology Database Collection includes descriptions of more 1500 molecular biology databases (http://www.oxfordjournals.org/nar/database/a/).

4.1 Genome Data Collections

Every year many new projects are started in which it is planned to obtain molecular genetic experimental data that characterize the work of various cells, tissues, organisms with different genetic characteristics in different conditions and with different effects. The following are some interesting examples of projects that produce huge amounts of data.

The 1000 Genomes Project (http://www.1000genomes.org/) is executed and coordinated by an international consortium of 75 companies and organizations. The aim of the project is the creation of the most detailed catalogue of the genetic variability of the human genome based on the results of sequencing the genomes of more than 2600 people from 26 populations around the world. The first part of the data (more than 1700 genomes) reached in 2012 the project volume of 200 TB. It is currently available for analysis on Amazon EC2, or Elastic MapReduce (http://aws.amazon.com/publicdatasets).

The 1001 Genomes Project (http://www.1001genomes.org) started in early 2008 to discover the whole-genome sequence variation in strains of the reference plant Arabidopsis thaliana, which serves as a model plant for a detailed study of the molecular and genetic mechanisms of plants. The information allows to develop new possibilities of genetics, which identifies alleles that underlie phenotypic diversity of the entire genome of a species, and at various levels, including the biochemical, metabolic, physiological, morphological, and the level of the whole plant. The results of the project will be important for the evolution of plant breeding and medicine. Currently, more than 1100 genetic lines (organisms) have been sequenced and analyzed.

Genome 10K Project (https://genome10k.soe.ucsc.edu/), includes a collection of more than 16000 sequences of the genomes of vertebrate species, including those living in the present and recently extinct mammals, birds, reptiles, amphibians, fish, and many of which are threatened or endangered. The total amount of data is expected to grow to more than 1.0 PB [24].

Encyclopedia of DNA Elements (ENCODE: http://www.genome.gov/10005107) is a description of the functional DNA elements in the human genome. The total amount of data in the present time is more than 15 TB.

The aim of the Human Microbiome Project is to characterize microbial communities found at multiple human body sites and to look for correlations between changes in the microbiome and human health. Currently, the volume of data over 18 TB.

The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov) contains a study of genomes of patients suffering from more than 33 types of cancer. Currently, it provides information on more than 7000 variants of cancer (the amount of data is about 1 PB). The amount of data as expected will be about 2.5 PB at the end of the project. Simply downloading the complete TCGA repository would require several weeks with a highly optimized network connection. Once downloaded, integrated analysis of this data remains out of reach for any researcher without access to the largest institutional compute clusters. The Cancer Genomics Cloud (CGC) Pilots project seeks to directly address these challenges by co-localizing data with the computational resources to analyze it, without the need to wait in a queue [16].

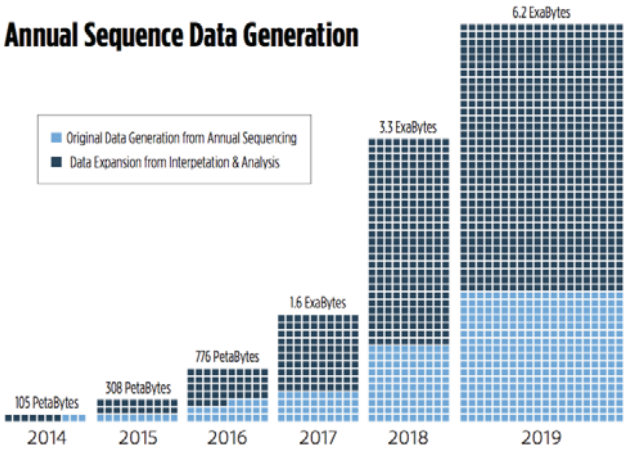

Fig. 1. Estimation of the annual volume of use of new data on genome sequencing and expand them in the analysis and interpretation (http://www.onrampbioinformatics.com/our-story/#big-data)

European Molecular Biology Laboratory is one of the world's largest biology-data repositories, currently stores 20 PB. The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size [39].

National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) - contains a large number of different databases for genomics, transcriptomics, proteomics, interactomics and other – omics data [11][34][40]. The total volume is more than 15 PB.

4.2. Charting the Human Proteome: Understanding Disease Using a Tissue-Based Atlas

A decade on from the completion of the Human Genome, the Human Protein Atlas, a multinational research project supported by the non-profit Knut and Alice Wallenberg Foundation, recently launched (November 6, 2014) is an open source tissue-based interactive map of the human proteome (Royal Institute of Technology (KTH), Stockholm, Sweden). A team of multidisciplinary researchers with expertise spanning biotechnology, information technology, and medicine have used a combination of several 'omics technologies to map proteins down to the single cell level, showing both proteins restricted to certain tissues — such as the brain, heart, or liver — and those present in all tissues. It has taken this team over 1,000 person years to compile a searchable, open source database (www.proteinatlas.org) comprising 13 million annotated images of human tissues. The interactive database is aimed at researchers interested in basic research into human biology as well as those working in translational medicine.

4.3 ELIXIR – European Life Science Infrastructure for Biological Information

Building a sustainable European infrastructure for biological information, supporting life science research

and its translation to medicine, agriculture, bioindustries and society, ELIXIR unites Europe's leading life science organizations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information.

ELIXIR will provide the facilities necessary for life science researchers - from bench biologists to cheminformaticians - to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built.

ELIXIR is an inter-governmental organization, which builds on existing data resources and services within Europe. It follows a hub-and-nodes model, with a single Hub located in a new building alongside EMBL-EBI in Hinxton, Cambridge, UK and a growing number of nodes located at centers of excellence throughout Europe. Governments and ministries of countries are members of the ELIXIR consortium, and the scientific community in each member country develops their national node. Russia is not an ELIXIR member. The goal of ELIXIR is to orchestrate the collection, quality control and archiving of large amounts of biological data produced by life science experiments. Some of these datasets are highly specialized and would previously only have been available to researchers within the country in which they were generated.

ELIXIR is a Special Project of the European Molecular Biology Laboratory (EMBL). ELIXIR interface with Research Data Alliance is an ELIXIR pilot action for 2015.

4.3.1 Interoperable Controlled-Access Big Data Transfer for ELIXIR - Expanding EGA Collaboration

Building upon the existing ELIXIR based collaboration within the European Genome-phenome archive (EGA) from the European Bioinformatics Institute (EMBL-EBI) in the UK and the Center for Genomic Regulation (CRG) in Spain, this pilot project aims to solve limitations on computing, network bandwidth, and storage buffer areas that affect the current EGA data delivery, and to produce a general data transfer solutions applicable across ELIXIR.

Existing UDT/FTP encrypted data stream transfer technology from the EGA developed at the EBI will be extended and made portable for use at the CRG, and a second method aimed at very large scale data transfer in the 100 TB range using GridFTP/Globus will be developed and tested between the EBI and CRG.

4.3.2 BILS-ProteomeXchange Integration Using EUDAT Resources

This pilot action aims to integrate the raw data repositories for mass spectrometry (MS) proteomics data run by BILS (Sweden) and ProteomeXchange (via the PRIDE database, EMBL-EBI, UK), using the European infrastructure EUDAT, and will serve as an example to connect national data storage services and international repositories through ELIXIR.

4.4 BD2K Initiative

The Big Data to Knowledge (BD2K) (BD2K, https://datascience.nih.gov/bd2k) enables the use of biomedical Big Data to advance human health through the creation, indexing, and dissemination of methods, tools, and training materials. The BD2K initiative (started in 2012) addresses four major aims that, in combination, are meant to enhance the utility of biomedical Big Data:

- To facilitate broad use of biomedical digital assets by making them discoverable, accessible, and citable.
- To conduct research and develop the methods, software, and tools needed to analyze biomedical Big Data.
- To enhance training in the development and use of methods and tools necessary for biomedical Big Data science.
- To support a data ecosystem that accelerates discovery as part of a digital enterprise.

BD2K has been established to develop new approaches, methods, software, tools, and related resources and provide training to advance data science in the context of biomedical research. 185 institutions involved, 11 BD2K centers of excellence [7] include: Big Data for Discovery Science Center, Center for Big Data in Translational Genomics, Center for Causal Modeling and Discovery of Biomedical Knowledge from Big Data, Center for Expanded Data Annotation and Retrieval, Center for Predictive Computational Phenotyping, Center of Excellence for Mobile Sensor Data-to-Knowledge, A Community Effort to Translate Protein Data to Knowledge: An Integrated Platform, ENIGMA Center for Worldwide Medicine, Imaging and Genomics, KnowEng, a Scalable Knowledge Engine for Large-Scale Genomic Data, Patient-Centered Information Commons, The National Center for Mobility Data Integration to Insight.

5 Data in Materials Science

Advanced materials are essential to economic security and human well-being and have applications in multiple industries, including those aimed at addressing challenges in clean energy, national security, and human welfare. The peculiarity of information systems (IS) in inorganic chemistry and materials science is extensive (hundreds PB) volumes of information that is result of "row" initial experimental data processing and systematization. Creating a digital data infrastructure that not only stores a wide range of data but is easily and reliably searchable is a challenge faced by materials science and engineering. Challenges facing the materials community include making users aware of the tools and data available; defining and implementing a widely accepted governance structure; balancing security requirements with data usability and

discoverability; and generating standards for describing data and assessing data quality. It is required to provide broad and open access to the data and tools generated by the materials community across the materials development continuum to allow both the reuse of individual data sets and the application of data analytics techniques to examine the aggregation of large volumes of data from many disparate sources.

5.1 Materials Genome Initiative (MGI)

On June 24 2011 USA President announced the *Materials Genome Initiative* [43], to double the speed with which new materials can be discovered, developed, and manufactured. Accelerating the development of advanced materials is critical for achieving global competitiveness. Just as the open sharing of DNA sequence data accelerated the Human Genome Project and fueled a rapid economic expansion in biomedical applications, the MGI aims to speed the creation and deployment of new materials through enhanced public-private coordination and greater access to instrumentation, modeling and simulation tools, and pre-competitive data that describe materials properties and behavior. A primary goal of the MGI is to catalyze greater collaboration across the advanced materials workforce, including Federal agencies, industry, professional societies, and academia. MGI will help position the U.S. for sustained leadership across the many sectors that utilize advanced materials from energy to electronics and defense to health care. MGI aims to capitalize on recent breakthroughs in materials modeling, theory, and data mining to significantly accelerate discovery and deployment of advanced materials while decreasing their cost. At the heart of MGI is the Materials Innovation Infrastructure, a framework of seamlessly integrated advanced modeling, data, and experimental tools that will be used to attain the MGI vision. On December 4, 2014 the Materials Genome Initiative Strategic Plan has been published [44]. NSTC (National Science and Technology Council) Subcommittee on the Materials Genome Initiative includes now NIST, DOE, DOD, NSF, NASA, NIH, USGS, DARPA in coordination with the Nanotechnology Knowledge Infrastructure (NKI).

Deeper integration of experiment, computation, and theory planned in MGI, as well as the routine use of accessible digital materials data, represents a shift in the usual way research is conducted in materials science and engineering. A major challenge facing MGI is how to establish mechanisms that will facilitate a flow of knowledge across the materials development continuum through deeper collaborations not only between theorists and experimentalists, but between academia and industry, and with manufacturers as well.

5.1.1 The Materials Data Facility

In June 2014, the National Data Services Consortium announced its first pilot project, The Materials Data Facility (MDF) [42]. It served as a response to the White House's MGI to accelerate the process for creating new materials. Being able to share data readily through the materials development chain will be critical to achieving this acceleration. The MDF will provide material scientists a scalable repository for lab and computational data alike, a place to publish data with links to associated literature. The MDF will leverage a national infrastructure for data sharing and reuse to connect to a variety of resources for materials science research, including specialized databases and analysis capabilities. The capabilities needed by MGI scientists mirrors closely the broader NDS vision: sharing data privately before publication, creating data collections, publishing, linking with the literature, and connecting with other data resources and databases in the world. Facility is to make the data from materials research more widely and easily available for re-use, re-analysis, and verification.

5.2 VAMAS Program

Versailles Project on Advanced Materials and Standards (VAMAS) [70] is an international collaborative program which was set up to promote the underpinning research and development that is required to be undertaken as a precursor to the preparation of new standards relevant to advanced materials. It is intended that the program leads to agreed international collaboration standards which will aid trade world-wide. *Pre-standards research* — also called pre-normative research — is necessary in the case of advanced materials because traditional tests are simply not suitable for these materials. The pace at which they enter the market place requires the rapid development of novel approaches to measurement and performance evaluation. VAMAS exists to remove barriers to trade in new technologies by addressing this need for research on which to base international standards.

5.3 Data projects in Materials Science in Russia

All developments of IS in materials science in Russia are initiative of developers.

5.4 World-wide Data collections in Materials Science

Several examples of such collections follow. Besides data on materials properties, such sites contain also data analysis systems:

- data collections at the National Institute of Standards and Technology (NIST, USA) [46] contain information on plastics, carbon nanotubes, high-strength alloys, artificial bone and joint replacements and other of the emerging materials for which the NIST develops testbeds, defines benchmarks, and develops formability measurements and models;
- data collections at the National Institute for Materials Science (NIMS, Japan) [52] making Materials Information Hub. Databases of NIMS are a world-wide center of materials information, which ranges from the dimensions of atoms and molecules, through the nanoscopic and metallographic sizes, to the scales of components, structures and social infrastructures. NIMS is

carrying out fundamental research and generic/infrastructural technology research and development in the field of materials science, and with improving the level of materials science and technology;

- network of scientific and technical information STN (Germany) [63] is an online database service that provides global access to published research, journal literature, patents, structures, sequences, properties, and other data. With STN, the patent and sci-tech information needed to make business-critical decisions can be found. Paid access to these data is possible. In Russia there is a center of STN data access in the IOC of RAS;
- data collections at the Springer Materials (Germany) [62] support the curated data on 3000+ physical and chemical properties of 250,000+ materials and chemical systems. Data sources currently include the Landolt-Börnstein New Series, the Linus Pauling Files and specialized databases on thermophysical properties, polymer thermodynamics, adsorption isotherms, and 32,000+ substance profiles. Paid access to these data was organized by the RFBR grants.

5.4.1 Foreign Commercial Databases Most Popular among the Russian Inorganic Materials Scientists

The list of foreign collections of materials science data being of the most interest for the Russian researchers is given in the following table. The table contains also references to the Russian analogues of such collections as well as to the open access systems containing similar data collections.

Table 1

Commercial IS	Analogous Noncommercial IS (in Russia, open-access)
SGTE [59]	IVTANTHERMO [8] NIMS Thermodynamic DB [52]
Alloy Phase Diagram data (information about tens thousands of phase diagrams of binary and ternary systems) [3]	AtomWork [52]
MSI-Phase Diagram Centre (MSIT-PDC) (information about tens thousands of phase diagrams of binary, ternary, and more complicated systems) [45]	AtomWork
Springer Materials [62]	Phases [56] AtomWork
FIZ/NIST Inorganic Crystal Structure Database (ICSD) (information about crystal structures of inorganic substances) [17]	AtomWork
Database PDF (information about powder diffractograms of inorganic and organic substances) [33]	AtomWork

5.5 International communities and working groups

Several examples of such groups include:

- Scientific Group Thermodata Europe (SGTE) [59] is a consortium of centers engaged in the development of thermodynamic databanks for inorganic and metallurgical systems and their application to practical problems. SGTE provides, maintains and expand of high quality databases, that enable the user to undertake complex calculations of chemical equilibrium efficiently and reliably. SGTE centers are located in Canada, France, Germany, Sweden, United Kingdom, USA;
- Asia Materials Data Committee (AMDC) [4]. Development and integration of the information resources of chemistry and materials science developed in Asian countries (Japan, South Korea, China, India, etc.) is a task of this group. Russia is represented by A. A. Baikov Institute of Metallurgy and Materials Science (IMET);
- working team of developers of information systems in inorganic chemistry and materials science that joins NIMS (Japan) and IMET (Russia). The aim of this team is integration of the information resources of NIMS and IMET with the purpose of providing users with consolidated data on the properties of substances and development of the systems of calculation and prediction of material properties.

The detailed review of the information resources of inorganic chemistry and materials science is given in [38].

5.6 Materials Science Infrastructures

Scientific Group Thermodata Europe (SGTE) [59] is a consortium of centers engaged in the development of thermodynamic databanks for inorganic and metallurgical systems and their application to practical problems. SGTE provides, maintains and expand of high quality databases, that enable the user to undertake complex calculations of chemical equilibrium efficiently and reliably. SGTE centers are located in Canada, France, Germany, Sweden, United Kingdom, USA.

6 Data Collections and Infrastructures in the Earth Sciences

The object of the study in the Earth science is the planet Earth with its atmosphere. Comprehensive studies of lithosphere, atmosphere, hydrosphere, biosphere and cryosphere processes are aimed to a better understanding of the Earth functioning as a system. The main feature of the Earth science is a complex hierarchy of thematic areas including basic and applied sciences. This hierarchy imposes strict

limitations on the data from different thematic areas, as well as on integrated data structures aimed at usage in meteorology, climatology, oceanology, and ecology applications.

Data sets in the Earth science are the results of local and remote observations and numerical modeling of the studied processes. The volume of the corresponding archives, for example, for remote sensing amounts up to tens of PB, and for climatic numerical experiments it is up to several PB (CMIP5 (http://cmip-pcmdi.llnl.gov/cmip5/), ERA-CLIM (http://www.era-clim.eu/)). In the field of climatology main efforts are now focused on finding out the causes and consequences of the current and possible in the future global climate changes. The main application of this research is a creation of "climate services" similarly to weather services. The data generation tools here are represented by a network of weather stations, networks of oceanic buoys, networks of ground observation systems monitoring local climatic and environmental characteristics, a set of satellite-borne instruments monitoring the atmosphere and surface, and climate models.

The main source of large amounts of data is satellites. In the United States peta-scale data collection are formed, maintained and serviced by specialized national agencies (NASA, NOAA, DoE, etc.), while in Europe it is done by thematic supranational structures (such as ECMWF (http://www.ecmwf.int/) and ESA (http://www.esa.int/)) or by consortiums of leading core universities and research centers. These structures are actively involved in the implementation of the above-mentioned investigations. It should be mentioned that regular funding for the creation, development and support of these collections and the infrastructure goes through the budgets of respective Agencies and Programs and it reaches billions of dollars (or euros) a year.

Preparing this review authors sought to include materials that characterize present and near-future projects. In particular, projects for development of various components of Earth Observations (EO) program and its integration with modeling results, adequately represented in the proceedings of the Conference on Big Data from Space (BiDS'14).

6.1 Examples of Major Data Acquisition Projects in the Earth Science

In Earth science the most significant progress in providing a full range support to activities associated with large volumes of data is achieved in the area of remote sensing. Examples of relevant programs are presented below.

Copernicus [13], the most ambitious Earth observation program to date, is the new name for the Global Monitoring for Environment and Security program, previously known as GMES. This initiative is headed by the European Commission (EC) in partnership with the European Space Agency (ESA). ESA coordinates the delivery of data from upwards of 30 satellites. The EC, acting on behalf of the European Union, is responsible for the overall initiative, setting requirements and managing the services. The Copernicus Space Component comprises two types of satellite missions, ESA's families of dedicated Sentinels and missions from other space agencies, called Contributing Missions. A unified ground segment, through which the data are streamed and made freely available for Copernicus services, completes the Space Component. ESA is developing a new family of satellites, called Sentinels, specifically for the operational needs of the Copernicus program. The Sentinels will provide a unique set of observations, starting with the all-weather, day and night radar images from Sentinel-1A, launched in April 2014 (a TB per day is collected). Sentinel-2 will deliver high-resolution optical images for land services and Sentinel-3 will provide data for services relevant to the ocean and land. Sentinel-4 and Sentinel-5 will provide data for atmospheric composition monitoring from geostationary and polar orbits, respectively. Sentinel-6 will carry a radar altimeter to measure global sea-surface height, primarily for operational oceanography and for climate studies. In addition, a Sentinel-5 Precursor mission is being developed to reduce data gaps between Envisat, in particular the Sciamachy instrument, and the launch of Sentinel-5.

The Space Component [66] is managed by ESA and serves users with satellite data available through the Sentinels and the Copernicus Contributing Missions at national, European and international levels. The Copernicus Contributing Missions are divided into five Mission Groups, based upon the mission type (SAR/Optical/Atmospheric) and, for the Optical missions, per resolution class. The ground segment, facilitating access to Sentinel and Contributing Mission data, completes the Copernicus Space Component. The Space Component forms the European contribution to the worldwide Global Earth Observation System of Systems (GEOSS).

Copernicus will provide services for a range of different applications such as air-quality forecasting, flood warnings, early detection of drought and desertification, early warnings of severe weather, oil-spill detection and drift prediction, sea-water quality, crop analysis, forest monitoring, land-use change, agriculture, food security and humanitarian aid – to name but a few. These services fall into six main categories: land management, the marine environment, atmosphere, emergency response, security and climate change.

The Data Access Portfolio Document (DAP) [14] defines the datasets that are made available to the Copernicus Users in response to their Earth Observation data requirements, for a certain period of time, and the conditions (i.e. data licensing, product types available, delivery timelines, data access mechanisms) under which they are accessible. The categories of Copernicus Users eligible for accessing data from the Copernicus Space Component are defined and maintained by the European Commission (EC) in the Data Warehouse

requirements document [15]. Seven categories for which the legally binding definitions are part of the ESA-User License have been identified [69].

At the same time, Copernicus (receiving 5 billion Euros from 2014 to 2020 alone), represents a step-change in the European capacity to deliver high quality EO data with a policy that is based on full and open access to the data (up to 8 TB a day when fully operational). Copernicus can provide a major breakthroughs. Providing high resolution data about land, ocean and the atmosphere and allowing for derived products, Copernicus has the potential to drive science into new spheres and collaborations through the step change in data volume and quality but also by the integration potential of the geospatial reference framework it creates.

Copernicus, is comprised of multiple systems that collect data to monitor the Earth from different sources including in situ ground stations, airborne and sea-borne sensors, and Earth Observation (EO) satellites known as Sentinels. The Sentinel data will need to be ingested at a sustained rate of nearly 8 TB per day with total data projections at nearly 3 PB per year presenting new challenges for delivering data to science users in the United States (U.S). NASA is implementing for a Sentinel data mirror archive in the U.S. including potentially redistributing the data to other U.S. Government Agencies.

NASA's Earth Observing System (EOS) is a coordinated series of polar-orbiting and low inclination satellites for long-term global observations of the land surface, biosphere, solid Earth, atmosphere, and oceans, enabling an improved understanding of the Earth as an integrated system. EOS information infrastructure comprises 12 national data centers in the USA which store and provide continuous access to a wide variety of geophysical information about Earth and Space: polar and land processes; upper atmosphere, global biosphere, atmospheric dynamics and geophysics; physical oceanography, radiation budget, tropospheric chemistry, clouds and aerosols; global snow and ice distribution; cryosphere; biogeochemical dynamics; human interactions in the environment; hydrological cycle; climate and weather; solid Earth geophysics, marine geology and geophysics, solar terrestrial physics, paleoclimatology; and satellite remote sensing. В США создана the Earth Observing System Data and Information System (EOSDIS) [57].

At present, this experience is widely used for creation of national segments of the global information system and infrastructure of the designed for several decades ahead international project GEOSS (Global Earth Observation System of Systems) aimed at monitoring the Earth from space. The GEOSS Portal is the main entry point to Earth Observation data from all over the world (http://www.geoportal.org/web/guest/geo_home_stp). Targeted funding amounts up to tens of billions of dollars.

6.2 Infrastructure Projects Examples

Several examples of the infrastructure projects that are under development in USA and Europe follow.

Data Observation Network for Earth (DataONE, https://www.dataone.org/) is the foundation of new innovative environmental science through a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data. DataONE itself does not store data. It connects multiple data repositories in a federated network to provide integrated search and discovery and to provide replication services to the data repositories within the federation. Supported by the U.S. National Science Foundation as one of the initial DataNets, DataONE will ensure the preservation, access, use and reuse of multi-scale, multi-discipline, and multi-national science data via cyberinfrastucture elements and a broad education and outreach program.

DataONE will create lightweight and easily deployed "Slender Node" software and develop DataONE compatibility for common repository software systems (e.g. DSpace and others) that are already deployed in hundreds of high-value repositories worldwide. These new features include:

- measurement search to leverage semantic technologies and enable highly precise data discovery and recall of data needed by researchers;
- tracking the data through creation, all transformations, and analyses (provenance) to enable more reproducible science by storing and indexing provenance trace information that can be used to both reproduce scientific data processing and analysis steps and to discover specific data sources by examining the documented workflows; and
- data extraction, sub-setting and processing services to enable researchers at any location to more easily participate in "big data" initiatives (e.g. working with data from large environmental observatories and participating in broad-scale synthesis and modeling endeavors).

DataONE currently hosts three Coordinating Nodes that provide network-wide services to enhance interoperability of the Member Nodes and support indexing and replication services. Coordinating Nodes provide a replicated catalog of Member Node holdings and make it easy for scientists to discover data wherever they reside, also enabling data repositories to make their data and services more broadly available to the international community. DataONE Coordinating Nodes are located at the University of New Mexico, the University of California Santa Barbara and at the University of Tennessee (in collaboration with Oak Ridge National Laboratory). 25 member nodes are currently involved.

Satellite observations for climate modeling. New projects devoted to integration of satellite information

and modeling data are initiated by NASA. A next generation cyberinfrastructure to support comparison of satellite observations with climate models under development. Its architecture and software are funded by NASA's Computational Modeling, Algorithms and CyberInfrastructure (CMAC) program [47]. Publishing remote sensing data alongside of climate model output encourages better comparisons and understanding that, in turn, more completely inform decision makers, states, and federal government and (inter-)national stakeholders who make critical policy decisions involving the weather, future climate, state and regional level tourism, water resources and food management based on this information. It should be added that NASA role in the Intergovernmental Panel on Climate Change (IPCC) is growing and has solidified in the form of the obs4MIPs project, whose goal is twofold:

- the identification of NASA's key remote sensing observations that are readily comparable with the model output datasets part of the World Climate Research Program's (WCRP) Coupled Model Intercomparison Project (CMIP) and IPCC's Assessment Reports (AR) and
- the static publication of a handful of those datasets to the DOE funded Earth System Grid Federation (ESGF), the home for all IPCC generated AR data, a next generation cyberinfrastructure that overcomes these challenges.

The cyberinfrastructure provides automatic conversion of NASA HDF-EOS/HDF datasets into NetCDF/CF datasets compatible with the ESGF; the ability to perform model checking on those converted datasets using the Climate Model Output Rewriter (CMOR-2) checker; and the ability to automatically publish remote sensing data into the ESGF.

Next step is to transform climate analytics into a service [SCH14]. CAaaS combines high-performance computing and data-proximal analytics with scalable data management, cloud computing virtualization, the notion of adaptive analytics, and a domain-harmonized API to improve the accessibility and usability of large collections of climate data. MERRA Analytic Services (MERRA/AS) provides an example of CAaaS. MERRA/AS enables MapReduce analytics over NASA's Modern-Era Retrospective Analysis for Research and Applications (MERRA) data collection. The MERRA reanalysis integrates observational data with numerical models to produce a global temporally and spatially consistent synthesis of key climate variables. The effectiveness of MERRA/AS has been demonstrated in several applications. In our experience, CAaaS is providing the agility required to meet our customers' increasing and changing data management and data analysis needs.

Earth System Grid Federation (ESGF) is also oriented on climate modeling. ESGF is an international collaboration with a current focus on serving the World Climate Research Programme's (WCRP) Coupled Model Intercomparison Project (CMIP) and supporting climate and environmental science in general. It is a Gateway to scientific data collections, which may be hosted at sites around the globe. Gateways are web portals that allow one to register for access and discover data, and have potential access to the entire ESGF network of Gateways and Data Nodes. The U.S. Earth System Grid Center for Enabling Technologies (ESG-CET) project is a primary contributor of cyberinfrastructure for the ESGF effort, especially the gateways and nodes. The Earth System Grid (ESG) is developing a virtual collaborative environment based on Grid technologies such as Globus tools (The Globus Project) to facilitate analyzing the impacts of global climate change at national laboratories, universities and other laboratories. ESG will provide access to data produced by earth and climate science simulations through a Web portal. Through that portal climate scientists and researchers utilize distributed resources to discover, access, select, and analyze model data produced and stored in archives. Historical and future projections of climate model output for the recent IPCC report were primarily distributed through the Earth System Grid Federation (ESGF). For the latest set of experiments, nearly 2 PB were archived at ESGF nodes around the world. Projections for the amount of data for the next set of experiments are expected to exceed 100 PB. Increasingly, climate projection data are used by nonexperts for policy and management decision support purposes.

European projects. In Europe, a coordinated approach to the creation of a global data infrastructure was developed within the FP7 project «Global Research Data Infrastructures: The Big Data Challenges» [68]. The elaborated Program is aimed at development in Europe collaborative data infrastructure that will enable researchers and other stakeholders from education, society and business to use, re-use and exploit research data to the maximum benefit of science and society. Projects implementing this Program can be divided into three categories: projects of supranational structures like ESA and ECMWF, national projects of which there are very stingy information, and international projects within the EU RTD programs (http://cordis.europa.eu/). In particular, the list of infrastructure projects under the 7th Framework Programme includes more than 350 projects. Not less than a tenth of these projects are related to the Earth science.

The *Jasmin facility* (http://www.jasmin.ac.uk/) is a "super-data-cluster" which delivers infrastructure for data analysis. In technical terms it is half super-computer and half data-center and as such provides a globally unique computational environment. The JASMIN infrastructure [9] provides compute and storage linked together by a high bandwidth network in a unique topology with significant compute connected with much greater bandwidth to disk than is typical of a normal data center. JASMIN provides four basic services to the community: Storage (including disk and tape), Batch Computing, Hosted Computing, and Cloud Computing.

JASMIN key services include the academic component of the facility for Climate and

Environmental Monitoring from Space (CEMS) and the Centre for Environmental Data Archival (CEDA, including the British Atmospheric Data Centre, BADC). JASMIN will offer a 15 PB storage infrastructure. A range of NERC science community services are run in the JASMIN infrastructure, one of which is the academic component of the facility for Climate, Environment and Monitoring from Space (CEMS), hosting data and services specifically for the Earth Observation science community.

The FP7 thematic infrastructure project VAMDC (Virtual Atomic and Molecular Data Center) [18]. A number of Earth science applied fields are based on results of basic sciences. In particular, the important role is played by quantitative data obtained in such basic sciences as spectroscopy, atmospheric chemistry reactions, etc. Given that the number of molecules considered in solving such problems as an air quality forecasting in the region is almost a thousand (over two thousand counting their isotopes), the volume of the spectral data and the costs of their quality assessment (taking into account the constant flow of new data for the new spectral ranges) make these tasks extremely labor intensive. VAMDC is one of the completed projects in Europe related to basic sciences [18][58]. It was oriented on many research groups and institutes within the European Research Area (ERA) playing a central role in the production of a vast range of atomic and molecular (AM) data that is of critical importance across a wide range of applications. The VAMDC has aims to build a secure, documented, flexible and interoperable e-science environment-based interface to the existing AM data. Nowadays VAMDC becomes a European legal entity.

6.3 Comparable Projects in Russia

In the area of information resources development for the Earth science there are no research infrastructure projects in Russia comparable by the scale with Cyberinfrastructure or e-Science projects. There is a large departmental project ESIMO (http://portal.esimo.ru/portal), which provides integrated information support for maritime activities via an access to the resources of marine information systems. Also there are small-scale projects related to spatial data for the Russian Federation subjects. Projects in the field of IT support to the Earth science funded by the Russian Fund for Basic Research (RFBR, http://www.rfbr.ru) and the Russian Science Foundation (RSF, http://rscf.ru) are not a part of any long-term state programs.

6.4 International Communities and Working Groups

Major international communities in Earth Data Intensive Sciences are Earth and Space Sciences Division (ESSI) of AGU and ESSI Division of EGU. As an example of the formation and support of the community can serve the VAMDC project. Upon its completion the VAMDC consortium has been created. The RAS Institute of Astronomy represents Russia in this consortium. The purpose of this community is to support a distributed information system containing more than two dozens of bases of spectral data, preparation (together with international organizations) of spectral data standards for their description, and transfer and dissemination of these data in Africa, Asia and South America. Russian spectral data of the project are stored at the Institute of Astronomy RAS, the Institute of Atmospheric Optics SB RAS and the Federal Nuclear Center (Snezhinsk, Russia). The problem of participation in the consortium is necessity to pay an annual fee in the foreign currency. In Russia, the mechanism of such contributions and their legal liability are not well developed yet.

7 Examples of Data Infrastructures and Projects Preparing for the Forthcoming Advanced Data Sources Access and Analysis

The European Commission is supporting the development of a pan-European multi-disciplinary data infrastructure through Horizon 2020. A strong coordination effort at European and global levels and the promotion of global interoperability of data infrastructures through community led initiatives are required. Just a few guiding principles to follow.

Federation. A broad trend towards working in data federations for various purposes is being developed. These federations are networks of data repositories and centers that offer processing frameworks and that act based on agreements about legal and ethical rules, interface and protocol specifications and a stack of common services for handling data. Increasingly often such centers are members of multiple federations: a climate modelling center, for example, is a member of the climate modelling data provider federation, as well as being a member of the EUDAT data federation and also a member of the European AAI federation. This trend is likely to continue and will lead to even more federation arrangements. A coordinated approach is planned where each center creates a description of its capabilities, so that for each federation the same description can be (re)used to extract the information needed. Such approach will provide for making research data “open” by default and help change the current research culture to promote data sharing.

Open sharing of the data. Researchers belong to both institutions and to disciplines. All disciplines are international in nature, and so it is critical that there are coordinated international approaches to reducing barriers to data exchange and re-use.

Main technological impediments to data sharing/reuse to be overcome include:

- Heterogeneity of Data Representations and Query Languages;
- Understandability and Discoverability of data;
- Movement of Data across semantic boundaries between multiple contexts;
- Data Mismatching problems (e.g., quality mismatch, data incompleteness mismatch; data

abstraction mismatch – such as conceptual, spatial, temporal, etc.).

European collaborative data infrastructure EUDAT might be considered as an initial step in these directions. [20] It is a pan-European solution to the challenge of data proliferation in Europe's scientific and research communities. EUDAT comprises 25 European partners, including data centers, technology providers, research communities and funding agencies from 15 nations. EUDAT offers common data services, supporting multiple research communities as well as individuals, through a geographically distributed, resilient network connecting general purpose data centers and community-specific data repositories. infrastructure offers solutions for finding, sharing, storing, replicating, staging and performing computations with primary and secondary research data. Community-specific data repository managers can join the data infrastructure to archive, replicate, process and catalogue data on behalf of their community.

7.1 The National Data Service Framework (NDS)

While some communities are making progress in developing discipline-specific data services, the U.S. and international scientific communities lack a unified framework and supporting services for storing, sharing, and publishing data; for locating data; or for verifying data. More specifically, they are lacking standard means of accessing data, software, tools, metadata, and other project materials that can span across disciplines. These capability gaps make it difficult to build on prior research or to reproduce the results of a scientific publication. The nation urgently needs an open framework that supports an integrated set of national-scale services to individually and collectively enable the efficient, convenient, and secure storage, sharing, publication, discovery, verification, and attribution of data by individuals, groups, and large collaborations. This framework and services will constitute a National Data Service (NDS) [49]. Finally it should be possible to get a research environment where access to and citation of data is as useful and necessary as it is for published literature.

NDS is planned as an international federation of data providers, data aggregators, community-specific federations, publishers, and cyberinfrastructure providers. It builds on the data archiving and sharing efforts under way within specific communities and links them together with a common set of tools.

These tools and services would cover four basic capabilities:

- searching for data;
- moving data, both between different repositories and between repositories and computing platforms;
- sharing and publishing data;
- creating, maintaining, and tracking links between data and literature.

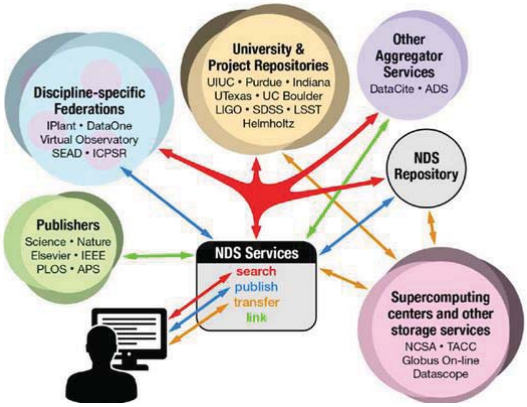

Fig. 2. NDS Environment.

The international partners, particularly the Research Data Alliance, will help NDS to ensure transparent access data from around the world. The NDS Consortium has been formed to link NSF DataNet (Data Conservancy, DataONE, SEAD), DIBBs (NCSA Brown Dog) and other major disciplinary initiatives (e.g. ICPSR, ADS); MREFCs (IceCube, LIGO, LSST, NEON), universities, and national organizations and services that connect them (Globus, Internet2, XSEDE, SHARE); publishers (e.g., APS, Elsevier, Nature, Science); and important international efforts (e.g., RDA, Helmholtz, EUDAT, OpenAire) (Fig. 2). Strong partnerships with US and international research organizations and publishers will drive impact.

To illustrate how a scientist might use the NDS, consider the following use case for the astronomy research.

In 2021 the LIGO gravitational wave observatory detects a strong "transient" burst event with an unknown source; an alert is issued. Across the US, physicists and astronomers (who have never worked directly together) engage NDS discovery services to find relevant data from other instruments, leading them to detections from the IceCube neutrino observatory, further isolating the originating portion of the sky. NDS discovery services connect the researchers to the federated discovery tools of the Virtual Observatory to collect data by sky position from large surveys like DES and LSST to look for electromagnetic precursors. Through literature searches, they find publications describing characteristics of similar detections; recent publications and an arXiv preprint supporting NDS data linking lead them to the data underlying the analyses.

They use NDS data transfer services to migrate previous detection data as well as simulation data held at the Blue Waters supercomputing system, containing previously unpublished neutrino emission predictions, to DataScope, a specialized computing platform to compare observations with theoretical models. From this analysis, a crucial insight suggests a new class of stellar object. Using NDS transfer tools, they pull together the LIGO data, corresponding IceCube detections, image cutouts from LSST, and analyses of simulation data into their private space in an NDS

repository. NDS metadata generation tools help them organize a new collection.

Soon, a paper is submitted to a journal, including identifiers for the new data collection. Once the paper is accepted, the NDS data collection is sent to a campus archive for longer-term curated management. With the new publication, readers have direct access to the underlying data, enabling them to verify and extend the results. Results and data are further available to educators, who bring the discovery to a broad audience by updating astronomy e-textbooks.

7.2 The Research Data Alliance

The Research Data Alliance (RDA) created to enable data to be shared across barriers was started in 2013 by a core group of interested agencies – the European Commission, the US National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation. RDA has grown from a core group of committed agencies to a community that now comprises more than 2600 members from more than 90 countries, all dedicated to pragmatically removing the barriers to data sharing and raising awareness of those challenges among regions, disciplines, and professions. RDA still is in a state of discussion and clarification of its objectives. This work is organized in multiple working and interest groups. Twice a year in various locations worldwide RDA holds Plenary meetings to provide the RDA community an opportunity to network and collaborate with peers in various disciplines, and hear from industry experts and world leaders on topics related to research data sharing and exchange. To get an idea of the RDA activity a list of the Large Scale Data Projects discussed at the RDA's 5th Plenary Meeting that was held in San Diego, California, March 8 -11, 2015 (https://rd-alliance.org/plenary-meetings/rda-fifth-plenary-meeting.html) included the following:

- EPOS: A large scale distributed Heterogeneous Research Infrastructure for GeoScience;
- EUDAT: towards a pan-European Collaborative Data Infrastructure;
- Chandra Data Archive: Data Linking and Data Mining;
- DataONE: A DataNet Perspective of RDA; National Earth System Science Data Sharing Platform of China; Towards a Google for Data;
- CLARIN: The Human Brain Project;
- Sustainable Environments Actionable Data (SEAD): Lessons Learned in Data Stewardship;
- SeedMe: Building Blocks for Sharing Preliminary and Transient Results;
- Program "Supercomputing & Big Data" of the German Helmholtz Association; European Environmental Research Infrastructures - ENVRI and ENVRIPLUS;
- CINERGI: Community Inventory of EarthCube Resources for Geoscience Interoperability;
- ELIXIR: the European Life-science Infrastructure for Biological Information;
- OpenTopography: An NSF Earth Science Facility for High Resolution Topography Data;
- The National Data Service: Production Data Infrastructure for the US.

7.3 Preparing for Data Access in Astronomy

Various projects (missions) in different domains have recently started to collect data or are planned to start before or after 2020. In various countries the researchers in respective X-informatics in various countries have initiated (or started to prepare) investigating infrastructures supporting data access, analysis and management over the data collected in such projects. We choose astronomy here to show a couple of such investigations related to LSST project and Gaia mission.

7.3.1 Preparing for LSST Data Access

A Memorandum of Agreement (MOA) [67] between Institut National de Physique Nucléaire et de Physique des Particules (IN2P3), LSST Corporation, LSST Project Office, and NCSA (National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign) signed in March 2015 established a partnership between the parties to enable IN2P3 to contribute to the LSST Data Release Processing during the survey operations of the facility. In exchange, a set of French scientists will collaborate in the scientific exploitation of the LSST database. Under this agreement IN2P3 will contribute to the data processing operations of LSST through the provision of communications, processing hardware, and labor necessary to support LSST annual data release processing as a satellite processing site to the NCSA LSST Archive Center. This agreement solidifies the partnership between CNRS/IN2P3, LSSTC, and NCSA that will be effective and remain in full force and effect through the end of the LSST survey operations.

The goal of the project supported by CNRS/IN2P3 [19] is to collaborate in making LSST data products available to the public and scientists around the world, with two main deliverables: a) the transient event reporting system which will send out alerts to the community within 60 seconds of completing the image readout and b) yearly data releases which will deliver the most completely analyzed data products of the survey.

The annual release catalog consist of more than 100 tables, the most important of which being the object catalog summarizing for each physical source all information acquired during the project lifetime, and the source catalog giving access to each individual measurement data of a single object on a single exposure. Sky coordinates, magnitudes, object morphology and their associated uncertainties will be derived for each individual measurement, and summary data including light curve properties will be stored per object. For each kind of measurement, the automated

pipeline will employ several algorithms in order to cover the broadest range of astronomer needs. This will ensure that for widely accepted measurement algorithms there will be no need to get back to individual image pixels. Thus the end user as well as data mining algorithms will focus on knowledge extraction from the catalogs accessible through a database, rather than dealing with original pixels.

The baseline design for LSST is to provide access to its database though a distributed system located in a data access center. The LSST baseline for the database architecture is to support massive user queries in a massively parallel relational database composed of a single-node non-parallel DBMS, a distributed communications layer, and a master controller, all running on a shared-nothing cluster of commodity servers with locally attached disk drives. The architecture is driven by the variety and complexity of anticipated queries, ranging from single object lookups to complex $O(n^2)$ full-sky correlations over billions of elements. Given the current state of RDBMS and Map/Reduce market, an RDBMS solution is a better fit to the requirements, primarily due to features such as indexes, schema, speed and available spatial libraries. As no off-the-shelf reasonably priced solution meets LSST requirements nowadays, LSST developed a prototype of the baseline architecture, called Qserv, based on open-source software.

After many tests to determine the design configuration, MySQL was chosen as the baseline single-node DBMS and XRootD as an elastic distributed, fault-tolerant messaging system. To mitigate adherence to the underlying RDBMS, Qserv pays close attention to minimizing exposure to vendor specific features. In addition, many key features including the scalable dispatch system and a 2-level partitioner have been implemented at the prototype level and integrated with these two underlying components. Scalability and performance have been successfully demonstrated on a variety of clusters ranging from 20-node-100TB cluster to 300-node-30TB cluster, tables as large as 50 billion rows and concurrency exceeding 100,000 in-flight chunk-queries. Required data rates for all types of queries (interactive, full sky scans, joins, correlations) have been achieved and basic fault tolerance recovery mechanisms were demonstrated.

7.3.2 Preparing for Gaia Mission Data Access

The ESA Gaia satellite launched on December 19th 2013 is scanning the sky from the Lagrange L2 point in order to build the largest, most precise three-dimensional map of our Galaxy by surveying more than a thousand million stars. The Gaia spacecraft downlinks everyday an average of 40 GB of data (reaching 120 GB per day when scanning along the galactic plane) during the 8 hours visibility period from the two ESA ground stations (Cerberos and New Norcia) [10].

The Gaia data processing is handled by a European Scientific Consortium, namely DPAC and relies on six Data Processing Centres (DPC) distributed all around Europe (Madrid (DPCE), Toulouse (DPCC), Cambridge (DPCI), Torino (DPCT), Barcelona (DPCB) and Geneva (DPCG). Each one has its own peculiarities but they all face the same challenge: being able to handle dozens billion rows in database tables. The volume of the end-of-mission DB is expected to exceed 1PB disregarding intermediate data generated in each DPC.

There are finally 2 families of DPCs:

- the ones using the “infrastructure” package coupled with a centralized file system or dedicated DBMS: DPCE, DPCT, DPCB, DPCG;
- the ones using Hadoop as the job management system and the data management: DPCC and DPCI.

For DPCC (that is in charge of Spectroscopic processing) Hadoop was chosen because it is a scalable solution allowing an incremental purchase of the hardware in order to follow the growing needs in terms of volume and processing power over the 5 years of mission. The 6000 cores are expected at the end of the mission. The foreseen final DPCC operational cluster is a set of 8 racks, each of which consists of 64 servers embedded into 16 enclosures. Each server is connected by 2 Gbit attachments, solution that combines good performance and good reliability (redundancy of paths). The global network has a two levels tree topology. While rack consists of two 1Gbit links, the inter-rack backbone network is made of two 10Gbit links.

8 Conclusion

Practically in all data intensive domains (DID) data become a strategic resource affecting all areas of activity of people and determining competitiveness, level of advancement of science, industry, healthcare, defense capacity of the country.

Analysis of the five representative scientific domains in the survey has shown the following.

The novelty of the situation consists in the all over the world development of the process of creation of the PB data collections as the result of application of new high technological ground or space missions in the large research programs (initiatives) intended for the study of the diverse surrounding phenomena in different DIDs. In some domains massive data collections are produced as the result of integration of the large number of relatively small databases collected in various research laboratories around the world. In USA the production of PB data collection often is the natural results of strategic initiatives declared at the level of the US President. In the European Union such programs are the intergovernmental onesIn Russia large inter-agency research programs that would require creature of the newest instruments for studying natural phenomena as well as of the large international information infrastructures for data storage and analysis (e.g., with the BRICS countries) practically do not exist. The majority of the research projects are being organized by initiative in the network of interpersonal academic or university relationships. The need in scientific data is not determined in a systemic way by the governmental scientific bodies, such bodies do not regulate the

processes of duplication of actions of various agencies, scientific institutes and universities in the area of data storage, standardization and quality control.

As the result, contribution of Russia in the world wide data collections is insignificant. It is quite difficult to forecast changing of the situation in the next 10 years due to the underdevelopment of the related technologies in the country and the lack of a possibility of creation of the adequate programs requiring significant funding level. Thus one of the most important problems of preserving the level of scientific research in Russia is providing the efficient access of the Russian research institutes to the data accumulated in the world. Access to the data centers located at the territories of the foreign countries requires to solve a number of technical problems as well as to overcome political and financial restrictions (requiring sometimes to make the international agreements). By efficient access we understand a possibility of carrying out the data analysis as soon as the data become open for the scientists worldwide. We should understand also that it is not sufficient to create methods just for solving typical classic problems (such as methods of statistics, machine learning, data mining, etc.). The experience shows that in specific DID every specific data analysis problem, particularly for big data analysis, requires conducting of the special researches and experiments for creation of a specific approach for the problem solving building on the typical methods if possible.

Analysis shows that in contrast to Russia, abroad we can watch active preparation for usage of the forthcoming data sources (examples of such projects are given in the section 7), including also projects for broad discussions and planning of the new information infrastructures (such as, e.g., NDS, EUDAT, RDA, DataONE, MDF, ELIXIR, etc.). We can watch development of elements and creation of such infrastructures, for instance, for analysis of data that start coming in very soon (e.g., in Gaia mission) or in five years from now (e.g., LSST telescope). In every large-scale project the major international interdisciplinary communities of specialists are formed containing working groups for specification of new functions to be supported by the new infrastructures.

To make an efficient access of the research organizations in Russia to the data accumulated in the world possible, it seems reasonable to organize a national target interdisciplinary program for the development of the pilot project of the distributed infrastructure for the access to and storage of data and their analysis to be compatible with the foreign scientific open infrastructures. It is assumed that such program should include solving of the following main tasks:

- study and choice of the decisions for infrastructures and platforms supporting analysis of big data in various DIDs and providing their access to various kinds of data in the world as well as a shared multidisciplinary usage of data (alongside with the technical problems (including the communication problems) it is assumed that it will be possible to find an international solution of the problems caused by the specific data access restrictions imposed on the particular collections of data);
- formation of communities and working groups in various DIDs, providing measures for establishing contacts with the international communities having analogous designation;
- development of the high performance interdisciplinary center for data intensive usage (ICDIU) for the researchers and practitioners from various DIDs, accumulation of experience of problem solving in ICDIU, development of the proposals for the replication of ICDIUs in the country and their placement as a part of the distributed infrastructure.

Acknowledgement

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; for IOA SB RAS by RFBR grant 13-07-00411; for IMCES SB RAS by RFBR grants 13-05-12034, 14-05-00502; for IMET RAS by RFBR grants 14-07-00819, 15-07-00980; for INASAN RAS by RFBR grant 15-02-04053, by grant of Presidium of RAS Program P-41; for ICG SB RAS by RSF grant 14-24-00123; for RCN by RFBR grants 15-04-08744, 15-04-05066; for SRI (IKI) RAS by RFBR grant 15-02-10203-K.

References

[1] LIGO Scientific Collaboration, Virgo Collaboration: J. Aasi et al. Prospects for Localization of Gravitational Wave Transients by the Advanced LIGO and Advanced Virgo Observatories. LIGO-P1200087, VIR-0288A-12, 2013. - http://arxiv.org/abs/1304.0670

[2] C. P. Ahn, R. Alexandroff, C. A. Prieto et al. The Tenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Apache Point Observatory Galactic Evolution Experiment. The Astrophysical Journal Supplement, 211(2), 2014. DOI: 10.1088/0067-0049/211/2/17.

[3] ASM Alloy Phase Diagram Database. - http://www.asminternational.org/asmenterprise/apd/help/About.aspx

[4] Asia Materials Data Committee. - http://amd-c.org/index.html

[5] S. Barthelmy. GCN and VOEvent: A status report, Astronomische Nachrichten, 329(3), 2008, p. 340-342.

[6] A. N. Belikov, F. Dijkstra, J. A. Gankema et al. Information systems playground – the target infrastructure. Scaling Astro-WISE into the petabyte range. Experimental Astronomy, 35(1-2), 2011, p. 367-389.

[7] BD2K centers. - https://datascience.nih.gov/bd2k/funded-programs/centers

[8] G. V. Belov, V. S. Iorish, V. S. Yungman. IVTANTHERMO for Windows — database on thermodynamic properties and related software. CALPHAD, 23(2), 1999, p. 173-180.

[9] V. Bennett, P. Kershaw, M. Pritchard et al. EO science from big geo data on the JASMIN-CEMS infrastructure. Proc. of the Conference on Big Data from Space BiDS'14. European Space Agency-ESRIN, 2014.

[10] B. Frezouls, P.-M. Brunet. Big data technology in the service of the Gaia data processing. Proc. of the Conference on Big Data from Space BiDS'14. European Space Agency-ESRIN, 2014.

[11] H. Binder, L. Hopp, K. Lembcke, H. Wirth. Personalized disease phenotypes from massive OMICs data. Big Data Analytics in Bioinformatics and Healthcare. IGI Global, 2015.

[12] BRAIN 2025: A Scientific Vision. - http://braininitiative.nih.gov/2025/BRAIN2025.pdf

[13] Copernicus. Observing the Earth. - http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview3

[14] Copernicus Space Component Data Access Portfolio: Data Warehouse 2014 – 2020. Prepared by B. Hoersch, V. Amans. Reference COPE-PMAN-EOPG-TN-15-0004. 2015.

[15] Data Warehouse Requirements V2.0 – Copernicus Data Access Specifications of the space-based Earth Observation needs for the period 2014-2020.

[16] B. D. Dusenbery, Z. Onder, D. Locke, K. Blair1, D. Kural. Petabyte-Scale Cancer Genomics in the Cloud. TCGA Symposium 2015 Poster Presentation.

[17] DBs of NIST. - http://www.nist.gov/chemistry-portal.cfm

[18] M. L. Dubernet, V. Boudon, J. L. Culhane et al. Virtual atomic and molecular data centre. Journal of Quantitative Spectroscopy and Radiative Transfer, 111(15), 2010, p. 2151-2159.

[19] E. Gangler. Big data challenge posed by the Large Synoptic Survey Telescope. Proc. of the 2014 conference on Big Data from Space BiDS'14. European Space Agency-ESRIN, 2014.

[20] EUDAT – European Data project. - http://www.eudat.eu/

[21] Fact Sheet. 2014. - https://www.whitehouse.gov/sites/default/files/microsites/ostp/brain_fact_sheet_9_30_2014_final.pdf

[22] S. W. Fleming, F. Abney, T. Donaldson et al. Beyond The Prime Directive: The MAST Discovery Portal and High Level Science Products. American Astronomical Society, AAS Meeting #225, #336.59, 2015.

[23] N. Fourniol, J. Lockhart, D. Suchar et al. News from ESO Archive Services: Next Generation Request Handler and Data Access Delegation. Proceedings of Astronomical Data Analysis Software and Systems XXI Conference. ASP Conference Series, V. 461. Edited by P. Ballester, D. Egret, and N.P.F. Lorente. San Francisco: Astronomical Society of the Pacific, 2012, p.669.

[24] Genome 10K community of scientists. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. Journal of Heredity, 100(6), 2009, p. 659–67.

[25] C. S. Greene, J. Tan, M. Ung, J. H. Moore, C. Cheng. Big Data Bioinformatics. Journal of Cellular Physiology, 229(12), 2014, p. 1896–1900.

[26] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, J. Tegnér. Data integration in the era of omics: current and future challenges. BMC Systems Biology, 8(2:I1), 2014.

[27] N. S. Kardashev, V. V. Khartov, RadioAstron collaboration. RadioAstron — A Telescope with a Size of 300 000 km: Main Parameters and First Observational Results. Astronomy Reports, 57, 2013, p. 153-194.

[28] M. J. Hawrylycz, E. S. Lein, A. L. Guillozet-Bongaarts et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature, 489, 2012, p. 391–399.

[29] M. Herland, T. M. Khoshgoftaar, R. Wald. A review of data mining using big data in health informatics. Journal of Big Data, 1(2), 2014.

[30] Tony Hey, S. Tansley, K. Tolle. The Fourth Paradigm - Data Intensive Scientific Discovery. 2009. - http://goo.gl/edvr6W

[31] Human Brain Project. - https://www.humanbrainproject.eu

[32] Human Connectome Project. WU-Minn HCP 500 Subjects Data Release: Reference Manual. 2014.

[33] International Centre for Diffraction Data. - http://www.icdd.com/.

[34] N. T. Issa, S. W. Byers, S. Dakshanamurthy. Big data: the next frontier for innovation in therapeutics and healthcare. Expert Rev. Clin. Pharmacol. 7(3), 2014, p. 293–298.

[35] M. Juric, T. Tyson. LSST Data Management: Entering the Era of Petascale Optical Astronomy. Highlights of Astronomy, 16, 2015. p. 675.

[36] D. B. K. Kamesh, V. Neelima, R. R. Priya. A Review of Data Mining using Bigdata in Health Informatics. International Journal of Scientific and Research Publications. 5(3), 2015.

[37] I. Khabibullin, S. Sazonov, R. Sunyaev. SRG/eROSITA prospects for the detection of GRB afterglows. Monthly Notices of the Royal Astronomical Society, 426(3), 2013, pp. 1819-1828.

[38] N. N. Kiselyova, V. A. Dudarev, V. S. Zemskov. Computer information resources in inorganic chemistry and materials science. Russ. Chem. Rev. 79(2), 2010, p. 145-166.

[39] D. Kumar, R. Kumar. Impact of Biological Big Data in Bioinformatics. International Journal of Computer Applications, 101(11), 2014.

[40] J. W. Lichtman, H. Pfister, N. Shavit. The big data challenges of connectomics. Nature Neuroscience, 17, 2014, p. 1448–1454.

[41] LSST and Technology Innovation. - http://www.lsst.org/lsst/about/technology

[42] The Materials Data Facility, http://www.nationaldataservice.org/mdf/

[43] Materials Genome Initiative for Global Competitiveness. 2011. - http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf

[44] Materials Genome Initiative Strategic Plan. 2014. http://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan_-_dec_2014.pdf

[45] Materials Science International GmbH. - http://www.matport.com/phase-diagram-center/buy-online/purchase/selectElements.

[46] Materials Science Portal. - http://www.nist.gov/materials-science-portal.cfm

[47] C. A. Mattmann. Next generation cyberinfrastructure to support comparison of satellite observations with climate models. Proc. of the Conference on Big Data from Space BiDS'14. ESA – ESRIN, 2014.

[48] J. M. Mazzarella, P. M. Ogle, D. Fadda et al. Explosive Growth and Advancement of the NASA/IPAC Extragalactic Database (NED). American Astronomical Society, AAS Meeting #223, #302.04. 2014.

[49] National Data Service (NDS). - http://www.nationaldataservice.org/

[50] NEMO ontologies. - http://purl.bioontology.org/ontology/NEMO

[51] Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC). - http://www.nitrc.org/include/about_us.php

[52] NIMS Materials Database. - http://mits.nims.go.jp/db_top_eng.htm

[53] Number of entries using search query 'database'. neuinfo.org. NIF. Retrieved 25 Jan 2015. - https://neuinfo.org/mynif/search.php?q=database&first=true&t=indexable&nif=nlx_144509-1

[54] OASIS. - http://www.oasis-brains.org/

[55] E. Perret, T. Boch, F. Bonnarel et al. Working Together at CDS: The Symbiosis Between Astronomers, Documentalists, and IT Specialists. Proc. of the Open Science at the Frontiers of Librarianship Conference. ASP Conference Series, Vol. 492. San Francisco: Astronomical Society of the Pacific, 2015, p. 13.

[56] Phases Database. - http://phases.imet-db.ru.

[57] H. K. Ramapriyan, J. Behnke, E. Sofinowski, D. Lowe, M. A. Esfandiari. Evolution of the Earth Observing System (EOS) Data and Information System (EOSDIS). In Standard-Based Data and Information Systems for Earth Observation: Lecture Notes in Geoinformation and Cartography, Liping Di and H.K. Ramapriyan, Eds. Springer: Berlin-Heidelberg, 2010.

[58] G. Rixon, M.-L. Dubernet, N. Piskunov et al. VAMDC - The Virtual Atomic and Molecular Data Centre - A New Way to Disseminate Atomic and Molecular Data - VAMDC Level 1 Release. Journal of Physics: Conference Series, 1344, 2011, p. 107-115.

[59] Scientific Group Thermodata Europe. - http://www.met.kth.se/sgte/

[60] J. L. Schnase, D. Q. Duffy, M. A. McInerney et al. Climate Analytic as a Service. Proc. of the Conference on Big Data from Space BiDS'14. ESA – ESRIN, 2014.

[61] B. M. Shustov, A. I. Gomez de Castro, M. Sachkov et al. WSO-UV progress and expectations. Astrophysics and Space Science, 354(1), 2014, p. 155-161.

[62] Springer Materials. - http://materials.springer.com/

[63] STN. - http://www.stn-international.de

[64] A. R. Taylor. Data Intensive Radio Astronomy en route to the SKA: The Rise of Big Radio Data. Highlights of Astronomy, 16, 2015, p. 677.

[65] P. de Teodoro, A. Hutton, B. Frezouls et al. Data Management at Gaia Data Processing Centers. Astrostatistics and Data Mining, Springer Series in Astrostatistics, V. 2. Springer Science+Business Media New York, 2012.

[66] The Copernicus Space Component: Sentinels Data Products List. ESA, Copernicus Space Component Ground Segment team. Reference COPE-GSEG-EOPG-PD-14-0017. 2014.

[67] The LSST-French Connection: Signed and Tweeted! - http://www.lsst.org/News/enews/french-connection-201504.html

[68] Towards a 10-year vision for global research data infrastructures. GRDI2020 Final Roadmap Report. 2012, 108 p. - http://www.grdi2020.eu

[69] User Categories. - https://spacedata.copernicus.eu/web/cscda/copernicus-users/user-categories

[70] Versailles Project on Advanced Materials and Standards (VAMAS). - http://www.vamas.org/

[71] Why neuroinformatics? International Neuroinformatics Coordinating Facility. - http://www.incf.org/about/why-neuroinformatics

[72] O. Zhelenkova, V. Vitkovsky, T. Plyaskina. Electronic archive of observational data of astrophysical observatory., Digital Libraries, 13(4), 2010. - http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2010/part4/ZVP

[73] N. S. Kardashev, I. D. Novikov, V. N. Lukash et al. Review of scientific topics for the Millimetron space observatory. Physics-Uspekhi, 57(12), 2014, p. 1199-1228.