CODATA and global challenges in data-driven science

A. Rybkina[1,2], S. Hodson[2], A. Gvishiani[1], P. Kabat[3], R. Krasnoperov[1], O. Samokhina[1], and E. Firsova[1]

Received 13 March 2018; accepted 25 June 2018; published 9 August 2018.

This synthesis report presents the scientific results of the international conference "Global Challenges and Data-Driven Science" which took place in St. Petersburg, Russian Federation from 8 October to 13 October 2017. This event facilitated multidisciplinary scientific dialogue between leading scientists, data managers and experts, as well as Big Data researchers of various fields of knowledge. The St. Petersburg conference covered a wide range of topics related to data science. It featured discussions covering the collection and processing of large amounts of data, the implementation of system analysis methods into data science, machine learning, data mining, pattern recognition, decision-making robotics and algorithms of artificial intelligence. The conference was an outstanding event in the field of scientific diplomacy and brought together more than 150 participants from 35 countries. It's success ensured the effective data science dialog between nations and continents and established a new platform for future collaboration. *KEYWORDS:* Big Data; Open Data; FAIR principles; data-driven science; system analysis methods; data mining; machine learning; pattern recognition; international conference; CODATA.

Citation: Rybkina, A., S. Hodson, A. Gvishiani, P. Kabat, R. Krasnoperov, O. Samokhina, and E. Firsova (2018), CODATA and global challenges in data-driven science, *Russ. J. Earth. Sci., 18,* ES4002, doi:10.2205/2018ES000625.

Introduction

Research Data Management (RDM) is becoming increasingly important as data is growing in unprecedented volumes. Nor is it enough simply to store and preserve data: curation and stewardship are necessary to ensure that information (metadata and provenance information) is added that allows data to be reused and ultimately allows value to be extracted from data. Good research practice, the tangible benefits of data reuse, re-analysis and largescale analysis or integration in meta-studies – all that mean that research institutions need to improve their ability to manage and curate digital data.

Combined with the maxim that research data should be "open by default" or "as open as possible, as closed as necessary", the FAIR principles – building on previous formulations (OECD, Royal Society, G8 Ministers) – have gained acceptance as a useful and effective summary of the attributes, that allow data to be understood, analyzed and reused in various contexts. Further work is required to adapt and develop FAIR Data Policies, to address legal issues, in particular those of legal interoperability (CODATA-RDA Interest Group on Legal Interoperability, Implementation Guidelines https://doi.org/10.5281/zenodo.162241) and those defining the necessary limits of Open Data.

[1]Geophysical Center of the Russian Academy of Sciences, Moscow, Russia

[2]Committee on Data of the International Council for Science (CODATA), Paris, France

[3]International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria

http://rjes.wdcb.ru/doi/2018ES000625-res.html

The international scientific conference titled "Global Challenges and Data-Driven Science" was organized by CODATA (the Committee on Data of the International Council for Science) in St. Petersburg, Russian Federation from 8 October to 13 October 2017 in partnership with the Russian CODATA NMO (National Membership Organization) and the Geophysical Center of the Russian Academy of Sciences (GC RAS). The conference benefited from the support of the Russian Science Foundation, and was the first ever Eurasian regional CODATA conference. Significant scientific contribution was provided to this event, also, by the International Institute for Applied Systems Analysis (IIASA) Laxenburg, Austria and by the Group on Earth Observations (GEO), Geneva, Switzerland.

The conference brought together more than 150 participants from 35 countries: Armenia, Australia, Austria, Bangladesh, Brazil, Canada, China, Egypt, Fiji, Finland, France, Germany, Ghana, Hong Kong, India, Ireland, Israel, Italy, Japan, Jordan, Kenya, Laos, Morocco, Namibia, New Zealand, Nigeria, Poland, Russian Federation, Sweden, Saudi Arabia, South Africa, Sri Lanka, Switzerland, United Kingdom, and the USA. Leading scientists, data managers and experts, as well as Big Data researchers, were among the conference participants.

International multidisciplinary scientific dialogue between representatives of various fields of knowledge was encouraged and facilitated at CODATA 2017. The conference examined and encouraged Open data principles and their substantiation in the FAIR data principles (which argue that in order to have greatest value for science and innovation, data should be Findable, Accessible, Interoperable and Re-usable) [*Wilkinson et al.,* 2016]. A major source of inspiration for this aspect of the conference was the principles and enabling practices that are described in the international accord "Open Data in a Big Data World" [*Science International,* 2015] initiated by CODATA. The accord identifies the opportunities and challenges of the data revolution as one of today's predominant issues of global science policy. Organized in 2015, the accord was the output of the first of a proposed series of annual meetings of the four top-level international scientific bodies: the International Council for Science – ICSU, the InterAcademy Partnership – IAP, The World Academy of Sciences – TWAS and the International Social Science Council – ISSC. Indeed, it is these four international bodies that can aspire to represent the global scientific community in international science policy and diplomacy, along with IIASA – International Institute for Applied Systems Analysis [*Johansson et al.,* 2012].

The St. Petersburg conference covered a wide range of topics related to data science. It featured discussions covering the collection and processing of large amounts of data, the implementation of system analysis methods into data science, machine learning, data mining, pattern recognition, decision-making robotics and algorithms of artificial intelligence. During four days of the conference, more than 160 scientific talks were presented in 25 sessions. On the occasion of the conference several business meetings and workshops were organized and were devoted for example to the Data Citation principals, geomagnetic studies, the role of the geoinformatics in the modern data world and to numerous other topics. Among the topics featured in scientific sessions were Data Science applications in Earth and planetary sciences, data mining for seismic hazard and risk assessment and earthquake prediction, Earth observing systems and data for global energy (oil and gas extraction and carbon dioxide storage), Big Data in mining and metallurgical technologies, geospatial data and applications in Earth sciences, and other topics.

The conference was opened with welcoming speeches from the leading members of the program committee. The outstanding Russian informatics scientist, Vladimir Vasiliev, rector of the Saint-Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO), served as chairman of the conference program committee. In his short opening speech he emphasized the tremendous impact of modern Data Science and information technologies on all aspects of human life and activity.

Geoffrey Boulton – member of the British Royal Society and President of CODATA; Heide Hackman – Executive Director of ICSU; Alexei Gvishiani – vice-chair of the IIASA council and chair of CODATA-Russia, member of RAS and Academia Europaea; Pavel Kabat – Director General and Chief Executive Officer of the International Institute for Applied Systems Analysis (IIASA);

Alevtina Chernikova – rector of National University of Science and Technology MISIS, awardee of the Government of the RF prize in the field of education; greeted the audience at the opening of the conference by formulating its goals, topics and directions of discussions.

Phenomenon of BIG data nowadays is one of the key issues in the modern scientific community. Extreme growth of data volumes among the wide range of research topics cause new challenges in data collection, storage and processing. These challenges are cross cutting among the conference sessions and panel discussion.

The scientific program of the conference, with a versatile and interdisciplinary format, was organized as a series of sessions thematically clustered over specific areas and topics of the Data Science and its applications to various scientific, social and economic or industrial sectors.

As a rule, morning sessions began with invited plenary lectures. Among the invited speakers the audience met internationally highly recognized scientists from leading national and international research organizations and universities. Professor Pavel Kabat (IIASA, Austria), made a presentation on Data Diplomacy and its role in the modern world [*Kofner et al.,* 2017]. Professor Fred Roberts, Director of CCICADA Rutgers University (USA), the author of more than 200 data scientific papers, formulated the principles of determining Big Data and possible scenarios of developments in this area [*DiRenzo et al.,* 2015; *Nelson et al.,* 2014]. Barbara Ryan, Director of the Intergovernmental Group of Earth Observations (GEO, Switzerland), highlighted the topic of studying the Earth from space and principles of general economic theory in the modern world regarding development of geo observation systems. Renowned Russian economist and rector of the Russian Academy of National Economy and Public Administration (RANEPA, Russia) Vladimir Mau delivered a lecture about current state of data science and its role in the evolution of human cognition [*Mau,* 2015]. Catriona MacCallum, director of Open Science for the Hindawi Publishing Corporation, presented in her lecture the importance of open publications and surveyed current issues in publishing data. CODATA President, Professor Geoffrey Boulton's lecture was called "Symphony of Data". In this "symphony", he attractively presented a narrative and analysis of scientific development and its modern challenges.

Conference speakers highlighted in their talks the importance of the missing aspects and exemplified the usage of the current Open Data strategies as well as the importance role of research data publishing and dissemination. Many speakers discussed relevant problems of data publication from the point of view of new tools and services that make publishing data easier and more effective, as well as self-sustainable [*CODATA,* 2015; *Costello,* 2009; *Hey et al.,* 2009; *Parsons and Fox,* 2013].

As the world is becoming more complex and interdependent, risks are becoming systemic. The conviction that science and technology, if wisely directed, can benefit all humankind; the belief that international co-operation between national institutes promotes co-operation between nations and so the economic and social progress of peoples – this is the major, fundamental and important philosophy of IIASA.

Speakers focused on global environmental change and drew attention to the unprecedented amount of data about health of the planet that provides great opportunities but also poses immense challenges for scientific analysis. A key challenge is to aggregate data from multiple sources with potentially questionable quality and credibility and obtain useful "information" as a result.

Also key issues were highlighted: collaboration at national, regional and international levels is essential – including hyper-partnering, radical sharing (e.g. creating true interdependencies similar to those that exist in healthy (biological) ecosystems); international policy agendas must be supported, leveraged and implemented; and broad, open data policies must be advanced to leverage existing and planned investments in Earth observations and geospatial data.

The major problematic boundaries of open data principles: safety and security, privacy and confidentiality, public/private interface and legal frameworks and principles. To understand the necessity of open data, one should remember the beautiful and apposite quotation from George Bernard Shaw: "If you have an apple and I have an apple and we exchange these apples, then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas."

The most thoroughly covered thematic do-

main of the conference was devoted to general issues of the Data Science. The Science International Accord on Open Data in a Big Data World [*Science International,* 2015], laid out a set of principles and enabling practices in order that Open Data should advance science, particularly in major interdisciplinary research areas, as well as the responsibilities of various stakeholders, including at the national level [*Atkins et al.,* 2003; *Bromley,* 1991; Doldirina et al., Legal Approaches for Open Access to Research Data, https://umaine.edu/scis/wp-content/uploads/sites/269/2017/09/LegalInterop Data.pdf]. Corresponding thematic sessions addressed first of all national policies and international perspectives of Open Data and Open Science program, as well as survey progress towards these objectives in a wide range of countries.

Reflecting these drivers, a growing number of funders and scholarly journals have developed research data policies. The presentations and discussion testified to the growing consensus that research data should be "open by default" and FAIR.

Data Science Applications

An important topic of the conference concerned the effective management and dissemination of historical disaster data. Data is an essential resource for disaster reduction and response, as exemplified by the quick response maps from observed data, the disaster loss information from multidisciplinary data, the knowledge and decision from mass information, the advice to post-disaster construction from stakeholders and the early warning and risk research from data simulation. The multidisciplinary data and records of a given disaster event are the historical documents for further disaster research, just as medical examination data are an essential resource for studying human illness. However, many historical datasets for particular natural disaster events have been lost after decades, even when those events had great destructive impacts with a high profile which attracted a great humanitarian response at the time [*Frolova et al.,* 2010].

In recent years, there was a surge in the attention given to the collection, preservation and data sharing of historical disaster data. More and more such datasets can be discovered and made accessible. But still a relatively small proportion of disaster datasets are published to make them really trustful, copyright-clear, and cited. CODATA Task Group of Linked Open Data for Global Disaster Risk Research works to promote data publishing of event-oriented disaster datasets, as an approach to manage the disaster related data.

Another important topical issue that was discussed is collaboration of stakeholders within the Data Science community on the regional level. The issues of Big Data, as well as Small Data, are common among different projects in regional collaborations. Therefore, the database construction for Data Science will be of common practice for regional groups. However, such databases constructed through regional collaboration may be quite specific to the region, discipline or language under consideration and, thereby, there will be a significant problem on how the regional efforts can be coordinated with a larger-scaled international collaboration along with funding issues.

Artificial intelligence (AI) [*Gvishiani et al.,* 2002], such as machine learning, text and factographic data mining, as well as deep learning and IOT (Internet of Things) / IOE (Internet of Every-thing) have become prevalent as crucial versatile methodologies and technologies for conducting Data Science. Because these are in the middle of mainstream developments, sharing the most advanced technologies may be faced by a variety of problems such as digital divide between regions and challenges of reconciling Intellectual property (IP).

The speakers at the relevant session "Regional Collaboration for Data Science" presented case studies for regional collaborations with the aim of identifying the challenges in conducting Data Science through various levels of regional collaboration such as intra-domestic regions, bilateral nations, and much larger regions like the whole Pacific Rim. The session demonstrated the increasing importance of well-established scientific network on a global scale, where CODATA national member organizations and regional committees could play a key role and provide an efficient tool for its collaboration [*Aitsi-Selmi et al.,* 2016; *Karmen et al.,* 2017].

A significant thematic domain covered by the conference, were the modern applications of Data Science. Nowadays, solutions to many problems, that could not be solved earlier, have become pos-

sible by means of Data Science. One of such problems is creation of a new sort of Knowledge-Based System (KBS) [*Rajendra et al.,* 2009]. The new understanding of a KBS means the calculation tool that possesses the following abilities: contains all relationships between all variables of the object; allows to calculate the values of one part of variables through others; allows to calculate solutions of direct and inverse problems; allows to predict characteristics of an object, that has not been investigated yet; allows to predict technology parameters to construct an object with desired characteristics. An ensemble of multifactor quality, quantity and computational models are the base of new sort of KBS [*Sheremet,* 2013].

Formulations of the problems of a creation of the new sort of KBS, the methodology, mathematical methods, informatics techniques, and tools for a creation of KBS was discussed in detail at a special thematic session, entitled "Data Driven Knowledge-Based Systems for Basic and Applied Sciences: Combustion, Detonation, Nanotechnology, Renewable Energetics, etc." The best practice and examples of KBS created in various areas of basic and applied research of combustion, nanotechnology, Materials Genome, solar energetic, socio-economic systems, etc. were presented and discussed. Two analytical platforms "Deductor" [https://basegroup.ru/deductor/description] and "PolyAnalyst" [http://www.megaputer.ru/polyanalyst.php], that provide necessary tools for KBS creation were presented at the session. These examplar platforms demonstrate the possibilities of offered by Data Science methods, as applied in the generalization of the connections between the object experimental variables, as well as in forecasting of "new experimental results" without real experiments [*Abrukov et al.,* 2007; *Guhr et al.,* 1998; Bobyl A. et al., 2016, Generalized Radon–Nikodym Spectral Approach. Application to Relaxation Dynamics Study, https://arxiv.org/abs/1611.07386].

Another session discussed the applications of Data Science and the Coordination of Data Standards and Interoperability in Agricultural Research. Current efforts to define, implement and coordinate agricultural research data standards are driven by issues related to interoperability, cost and quality. In addition, the aspiration to obtain greater value from existing agricultural datasets, agricultural productivity concerns, and desires to accelerate the transfer of agricultural research findings to the user community are critical factors that call for effective coordination. Recent opportunities in agricultural research data to drive change in the next decade, coupled with the current emphasis on adoption of Big Data solutions and the Data Cubes concept in the agriculture sector, underscore the urgent need for coordination of data standards and interoperability in agricultural research. The speakers of the panel session "Coordination of Data Standards and Interoperability in Agricultural Research: Gaps, Over-laps, Challenges and Future Directions" reviewed the motivations and requirements for standardization of agricultural research data, and the current state of standards development, interoperability issues and adoption – including gaps and overlaps – in the agriculture sector [*Kondrashov,* 2015].

In addition, multidisciplinary sessions were organized, dealing with the principles of citation and publication of data, data policy and diplomacy, smart cities, regional co-operation in the science of data, management of research data at universities, co-ordination of data standards, open data in education [*Costello,* 2009; Rauber A. et. al., 2015, Data Citation of Evolving Data – Recommendations of the Working Group on Data Citation, Research Data Alliance. Available: https://www.rd-alliance.org/system/files/RDA-DC-Recommendations_151020.pdf (Accessed 25 July 2017)]. Several sessions were devoted to the discussion of the use of Big Data in commerce, medicine, and social sciences [*Abrukov et al.,* 2007; *Agrawal et al.,* 2017; *Amato,* 2017; *Anand and Mohanty,* 2011; *Johansson et al.,* 2012; *Khater et al.,* 2017; *Medema and Fischbach,* 2015; *Metz,* 2005; *Wang et al.,* 2015; *Zhang et al.,* 2015].

Data Science and Earth and Planetary Studies

Data Science plays a pivotal role in the Earth and planetary sciences. In recent decades, the demand for information about our planet has driven the creation of new highly capable observing and collection systems. Environmental observations are critical for forecasting weather and climate, monitoring geophysical fields, volcanoes, seismicity, tsunamis, etc. and in assessing the recovery from disasters [*Bondur,* 2016; *Frigg et al.,* 2015; *Fuss et al.,*

2014; *Janssen,* 2010; *Reissell,* 2016; *Zlotnicki et al.,* 2005]. In this scenario, Earth observation technologies are developing rapidly to collect data from diversified locations over shorter periods of time. In turn, the datasets generated can be combined and analyzed to gain new scientific insights.

How is it possible to create an integrated system for Earth and environmental observation, collection and analysis in order to manage the increasing volume of data allowing easy access by the research and civil community? Each of the topic areas requires a range of measurements derived from a variety of platforms including satellite, airborne and in situ sources. Merging observed data with geospatial analysis allows the generation of better knowledge of natural processes and risk management.

The relevant thematic sessions of the conference provided cutting edge insights into creation of integrated systems for Earth and environmental observations, their data collection and analysis in order to manage efficiently the increasing data volumes and provide easy access to the research and civil communities. The speakers considered the state-of-the-art and perspectives in data science relevant to Earth observations and environmental research.

Among other topics, concerning Data Science applications in studying our planet, the aspect of geospatial information was covered. Geospatial information and corresponding technologies are essential in a wide range of applications and research sectors, supporting planning and decision making in the academic, governmental, commercial, and non-profit domains. To foster the growing demand of geospatial data, tools, technologies, and expertise scientific and governmental institutions across the globe are developing reliable geospatial information infrastructures and implementing appropriate policies. The thematic session "Geospatial data and applications in Earth's sciences" focused on the topical issues in the area of geospatial data and technologies in Earth sciences. Among other issues, this session explored the problems of organization and management of the vast arrays of geospatial information, which is acquired at many levels and that has a variety of potential uses [*Odintsova et al.,* 2017; *Rybkina et al.,* 2016].

A special scientific session was devoted to the application of Big Data in mining and metallurgy. It was entitled "Big Data in Mining and Metallurgical Technologies: Applications and Prospects". The spectrum of Big Data applications in these sectors is rather extensive. It covers both the flows of billions of discrete particles with individual properties, requiring description and testing, and the need to create specialised digital databases and models. The latter are required for processing optimization for natural and technogenic mineral raw materials, as well as for artificial ones. This session encompassed a discussion of the current tasks and prospects for collecting, storing, processing and analyzing Big Data sets and making of important management and production-related decisions in the mining and metallurgical industries based on respective studies. The session speakers discussed in detail the applicable platform solutions offered by the leading IT-companies for implementing sectoral tasks, as well as the existing methods and tools for analyzing Big Data, considering characteristic features of these industries [*Vaisberg,* 2015].

Among the Earth sciences applications, covered by the conference sessions, a session was devoted to the issues of data mining and systems analysis aimed at seismic hazard and risk assessment, earthquake prediction and to the data that are needed for these purposes. The session examined and discussed the results obtained by the creation and development of the phenomenological, system, and geoinformational approaches to the multivariate seismic hazard assessment based on artificial intellect algorithms. Also considered were the results of carrying out the seismic hazard assessment for specific tectonically active regions by means of these approaches and representation of the results using GIS technologies [*Gvishiani et al.,* 2013].

Thematic sessions on disaster risk research also included the session on study of natural hazards and risk assessment. In particular, the results of the Russian Science Foundation project No. 15-17-30020 "Application of system analysis for estimation of seismic hazard in the regions of Russia, including the Caucasus–Crimea and Altai–Sayan–Baikal region" were reported. This project was aimed at solving the problem of adequate seismic hazard assessment in seismically active regions of Russia (the Caucasus–Crimea and the Altai–Sayany–Baikal), as well as the development of new methods of seismic hazard assessment and the improvement of existing ones. FCAZm (Fuzzy Clustering And Zoning modernized), an advanced

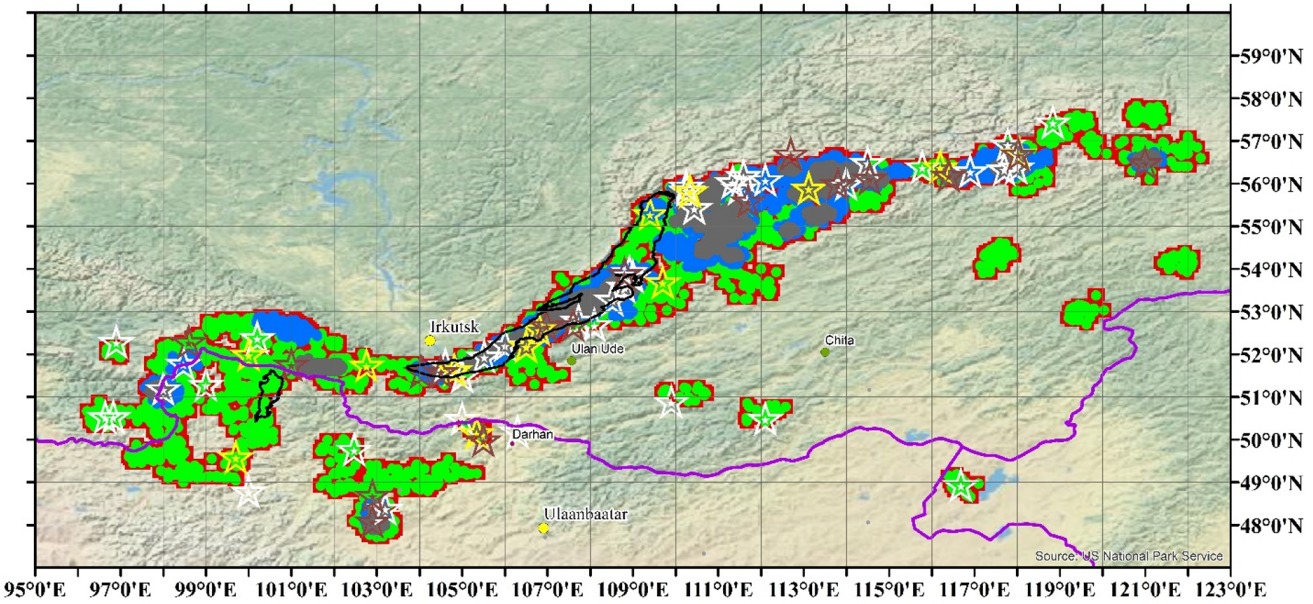

Figure 1. Results of seismic assessment for Altai–Sayan–Baikal region. Recognized highly-seismic FCAZ-zones are shown with filled color for magnitudes: red and green – $M \geq 5.5$; blue – $M \geq 5.75$; grey – $M \geq 6.0$. Earthquake epicenters are shown with stars; each color corresponds to a certain magnitude: white – $M \geq 5.5$; yellow – $M \geq 5.75$; brown – $M \geq 6.0$

.

version of the algorithmic system for determining earthquake-prone areas was introduced. It includes elements of artificial intelligence and pertains to the methodology of advanced system analysis. Morphostructural zoning and the FCAZ algorithmic system were used for integrated comparison and systems analysis of the results of strong earthquake prone areas recognition, obtained for the Altai–Sayany–Baikal region (Figure 1). These results were compared with the ones obtained for the Caucasus–Crimea region. The main conclusion from this analysis is that the identification of strong earthquake prone areas allows us to refine significantly the seismic hazard assessment for tectonically active regions. It facilitates the creation of the information basis for developing measures of damage reduction for civil and industrial infrastructure objects from seismic impacts [*Gvishiani et al.*, 2016, 2017].

Among other questions discussed were the integrated research on earthquake disaster risk, the scientific and educational aspects of such risk reduction, new approaches to seismic hazard assessment, observing and modeling capabilities to reduce uncertainties in hazard assessment, a contribution of hazard and vulnerability to earthquake risk, scientific, economic and political factors as well as the factors of awareness, preparedness and risk communication, which brought about the humanitarian tragedies of the early 21st century, and trans-disciplinary system approaches to disaster risk research and assessment. The session also fostered a broad forum to study the great earthquakes and tsunami occurring in subduction zones using the modern GPS observation in the framework of the keyboard model of deformation cycles of frontal seismogenic blocks of the island arcs and active continental margins [*Ismail-Zadeh et al.*, 2016; *Lobkovsky*, 1982].

The Earth and Planetary part of the conference scientific program included the session “Earth observing systems and data for global energy, oil and gas extraction and carbon dioxide storage”. The subject of session was especially relevant in the context of the continuous development of the oil and gas industry. The session speakers in their talks touched the aspects of creation of geospatial database for oil and gas deposits, drilling perspectives and geomagnetic survey for directional drilling, and digital geological exploration.

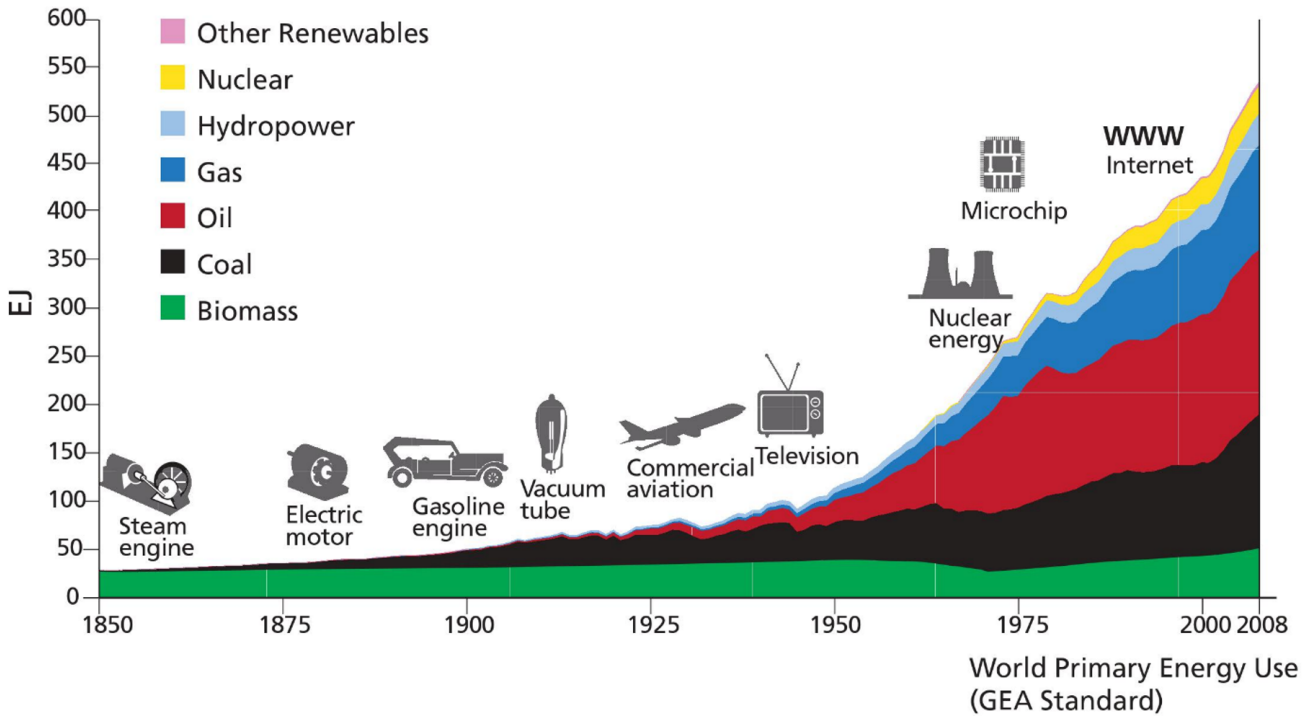

Figure 2. Evolution of primary energy shown as absolute contributions by different energy sources (EJ). Biomass refers to traditional biomass until the most recent decades, when modern biomass became more prevalent and now accounts for one-quarter of biomass energy. New renewables are discernible in the last few decades [*GEA,* 2012].

One of the topics touched within the session was an inextricable link between rapid development of petroleum industry and rapid deterioration of the environmental situation. Solving these problems is important for the future in the context of minimization of adverse impact on environment and reduction of natural and social disasters.

The further large-scale development of the oil and gas industry implies vast social and structural transformations. According to the Global Energy Assessment (GEA), in 2005, about 78% of the world's energy was based on fossil energy sources, which fully provided more than half of the world's population with energy resources, the cost of which was lower than ever. Figure 2 reflects the explosive growth of global primary energy with two distinct stages of development. The first one is characterized by transition from traditional energy sources (such as lumber) to coal, and then to oil and gas. Hydropower, biomass and nuclear energy have a common share of almost 22% over the past decades, while renewable energy sources such as solar and wind energy are still barely distinguishable in the figure. Despite the rapid growth in total energy consumption, more than three billion people still rely on solid fuels such as traditional biomass, waste, charcoal and coal for domestic cooking and heating. The resulting air pollution leads to more than two million premature deaths per year, mostly among women and children. In addition, approximately 20% of the global population still does not have access to electricity [*GEA,* 2012].

The conference scientific program also included a session "Geological data-driven science of the Arctic". The speakers discussed the issues of new geological data, including geophysical, stratigraphic-paleontological, isotope-geochronological, and tectonic, for structural-geological studies of the Arctic region and the Arctic Ocean [*Belov,* 2017; *Reissell,* 2016]. The discussion of these problems and directions promoted the growth of mutual understanding between geologists of different countries and various geological schools in developing a common position on the tectonic structure of this complex and inaccessible region of the planet. This is also important in the context of the existing disputes around the political delimitation of the Arctic territory. The session stimulated the exchange of infor-

mation and technologies between geological surveys and national academies of sciences and examined the driving forces for the development of Earth sciences, including for the development of our knowledge gained from the international project "Atlas of geological maps of the Circumpolar Arctic in 1 : 5 M".

Synthesis report based on the scoping phase of an emerging Arctic territory project, "Arctic Territory – Geological data and modeling" was presented in the framework of Arctic session. It reflected a variety of data related to the Arctic territories including Arctic coastal States' submissions to the United Nations Commission on the Limits of the Continental Shelf (the Commission). This initiative aims to provide a complex, multidimensional, and interdisciplinary overview of the challenges affecting the Arctic territory from a geological, economic, and political perspective, with a special focus on submissions to the Commission regarding territory over 200 nautical miles (nmi) from the coastline of the Arctic coastal States.

During other conference sessions the audience also paid considerable attention to the global energy and hydrocarbon production, terrestrial observation systems and interoperability, extraction of minerals and prospects for the application of new methods.

Conclusions

As a result of the conference, new international scientific groups were formed for further cooperation in the field of Big Data. The wide geographical coverage represented by the conference participants attests to the high profile, quality and scientific significance of this event. The conference was the first, experimental and successful attempt to bring together data specialists and scientists on the regional scale from various domains. Data issues in the modern world are characterized by their increasing importance and their growing complexity. Such international events provide higher visibility for existing studies and confront the community with the new goals and challenges.

The conference was an outstanding event in the field of scientific diplomacy and brought together more than 150 participants from 35 countries. It's success ensured the effective data science dialog between nations and continents and established a new platform for future collaboration. As major international research projects have frequently required diplomatic assistance to progress, the use of science to progress diplomatic objectives is evolving. Primarily science diplomacy is about advancing national interests, but these can be framed in three major dimensions.

First, where the primary diplomatic objective is to promote national needs through various objectives that may vary according to country size and the state of development. Secondly, there may be bilateral or regional issues where science must be part of the relationship management – for example management of cross-boundary resources or environments, agreeing standards, or crisis management. Thirdly, there is a growing number of issues where the national interest must be embedded within a commitment to the global interest. The governance of ungoverned spaces, for example the Arctic, require science to be at the base of framing governance relationships.

We are now in a position to build an effective system for integrating and managing research needs. Growing utilitarian importance of science diplomacy is reflected in various international science activities and CODATA international conference in St. Petersburg played important role in this dimension.

More information about the conference and presentations of the speakers are available on the official web site http://codata2017.gcras.ru.

Acknowledgments. We are grateful to all the conference participants, speakers and lecturers. The conference would not have taken place without the support from the following organizations: Russian Science Foundation (RSF), Committee on Data for Science and Technology (CODATA), Geophysical Center of RAS (GC RAS), International Council for Science (ICSU), CSA innovative group, Mekhanobr-Tekhnika Research and Engineering Corporation, Institute of Earthquake Prediction Theory and Mathematical Geophysics of RAS (IEPT RAS), National University of Science and Technology MISIS, International Institute for Applied Systems Analysis (IIASA), International Union of Geodesy and Geophysics (IUGG), International Social Science Council (ISSC), National Geophysical Committee of RAS (NGC RAS). We are also grateful to Iain Stewart (Head of the External Relations, Communications, and Library Department, IIASA) for his comments. The

authors are grateful to the reviewers for their discussion and valuable comments that contributed to the improvement of the submitted materials. The research was conducted in the framework of budgetary funding of GC RAS, adopted by The Federal Agency for Scientific Organizations (FASO Russia).

References

Abrukov, V. S., et al. (2007), Application of Artificial Neural Networks for Solution of Scientific and Applied Problems for Combustion of Energetic Materials, *Advancements in Energetic Materials and Chemical Propulsion, K. K. Kuo and J. D. Rivera (eds.)* p.268–283, Begell House Inc., Redding.

Agrawal, P., S. Khater, M. Gupta, N. Sain, D. Mohanty (2017), RiPPMiner: A bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links, *Nucleic Acids Res.*, *45,* No. W1, W80–W88, **Crossref**

Aitsi-Selmi, A., et al. (2016), Reflections on a Science and Technology Agenda for 21st Century Disaster Risk Reduction, *International Journal of Disaster Risk Science*, *7,* No. 1, 1–29, **Crossref**

Amato, G., F. Carrara, F. Falchi, C. Gennaro, C. Meghini, C. Vairo (2017), Deep learning for decentralized parking lot occupancy detection, *Expert Systems With Applications*, *72,* 327–334, **Crossref**

Anand, S., D. Mohanty (2011), Computational methods for identification of novel secondary metabolite biosynthetic pathways by genome analysis, *Handbook of research on computational and systems biology: Interdisciplinary applications, Limin Angela Liu, Dongquing Wei and Yixue Li (eds.)* p.380–405, Medical Information Science Reference (IGI-Global), Hershey, PA, USA. **Crossref**

Atkins, D., et al. (2003), Revolutionizing Science and Engineering Through Cyberinfrastructure, Report of the Blue-Ribbon Advisory Panel on Cyberinfrastructure, National Science Foundation, Washington, DC. (Bermuda Principles, 1996)

Belov, S. Yu. (2017), Monitoring of parameters of coastal Arctic ecosystems for sustainability control by remote sensing in the short-wave range of radio waves, *The Arctic Science Summit Week 2017* p.161, Czech Polar Reports, Prague. (ISBN 978-80-906655-2-1)

Bondur, V. G., A. S. Ginzburg (2016), Emission of Carbon-Bearing Gases and Aerosols from Natural Fires on the Territory of Russia Based on Space Monitoring, *Doklady Earth Sciences*, *466,* No. 2, 148–152, **Crossref**

Bromley, A. (1991), *Policy Statements on Data Management for Global Change Research, Global Change Research Program*, Office of Science and Technology Policy, Washington, DC, US.

CODATA (2015), The Value of Open Data Sharing, Paper commissioned by the Group on Earth Observations, Group on Earth Observations, Geneva, CH.

Costello, M. J. (2009), Motivating Online Publication of Data, *Bioscience*, *59,* 418–427, **Crossref**

DiRenzo, J., D. A. Goward, F. S. Roberts (2015), The Little-known Challenge of maritime cyber security (with), *Proceedings of the 6th International conference on Information, Intelligence, Systems and Applications (IISA)* p.1–5, IEEE, USA. **Crossref**

Frigg, R., E. Thompson, C. Werndl (2015), Philosophy of Climate Science Part I, *Observing Climate Change*, *12,* 953–964.

Frolova, N., V. Larionov, J. Bonnin (2010), Data Bases Used in Worldwide Systems for Earthquake Loss Estimation in Emergency Mode: Wenchuan Earthquake, *Proc. TIEMS 2010 Conference* p.4–26, TIEMS, Beijing, China.

Fuss, S., et al. (2014), Betting on Negative Emissions, *Nature Climate Change*, *4,* No. 10, 850–853, **Crossref**

GEA (2012), *Global Energy Assessment – Toward a Sustainable Future*, 93 pp. Cambridge University Press and the International Institute for Applied Systems Analysis, Cambridge, UK and New York, USA, and Laxenburg, Austria.

Guhr, T., A. Müller-Groeling, H. A. Weidenmüller (1998), Random-matrix theories in quantum physics: common concepts, *Physics Reports*, *299,* No. 4, 189–425, **Crossref**

Gvishiani, A., J. Dubois (2002), *Artificial Intelligence and Dynamic Systems for Geophysical Applications*, 350 pp. Springer-Verlag, Paris. **Crossref**

Gvishiani, A. D., S. M. Agayan, B. A. Dzeboev, I. O. Belov (2017), Recognition of Strong Earthquake – Prone Areas with a Single Learning Class, *Doklady Earth Sciences*, *474, Part 1,* 546–551, **Crossref**

Gvishiani, A. D., B. A. Dzeboev, S. M. Agayan (2016), FCAZm intelligent recognition system for locating areas prone to strong earthquakes in the Andean and Caucasian mountain belts, *Izvestiya. Physics of the Solid Earth*, *52,* No. 4, 461–491, **Crossref**

Gvishiani, A., et al. (2013), Fuzzy-based clustering of epicenters and strong earthquake-prone areas, *Environmental yanEngineering and Management Journal*, *12,* No. 1, 1–10.

Hey, T., S. Tansley, K. Tolle (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery, 1 edition (October 16, 2009)*, 284 pp. Microsoft Research, Redmond, Washington.

Ismail-Zadeh, A., A. Korotkii, I. Tsepelev (2016), *Data-Driver Numerical Modelling in Geodynamics: Methods and Applications*, Springer-Nature, Switzerland. (http://www.springer.com/gp/book/9783319278001)

Janssen, K. (2010), *The Availability of Spatial and*

Environmental Data in the European Union: At the Crossroads Between Public and Economic Interests, 617 pp. Kluwer Law International, USA.

Johansson, T. B., A. Patwardhan, N. Nakićenović, L. Gomez-Echeverri (2012), *Global Energy Assessment*, Cambridge University Press, Cambridge.

Karmen, P., M. F. Montserrat, D. G. Tom, C. Ian (2017), *Science for Disaster Risk Management 2017: Knowing Better and Losing Less*, Publications Office of the European Union, Luxembourg.

Khater, S., M. Gupta, P. Agrawal, N. Sain, J. Prava, P. Gupta, M. Grover, N. Kumar, D. Mohanty (2017), SBSPKSv2: structure-based sequence analysis of polyketide synthases and non-ribosomal peptide synthetases, *Nucleic Acids Res.*, *45,* No. W1, W72–W79, **Crossref**

Kofner, J., P. Balás, M. Emerson, P. Havlik, E. Rovenskaya, A. Stepanova, E. Vinokurov, P. Kabat (2017), High-level consultation meeting on Eurasian Economic Integration, IIASA project "Challenges and Opportunities of Economic Integration within a Wider European and Eurasian Space" Executive Summary, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Kondrashov, D., M. Chekroun, M. Ghil (2015), Data-driven non-Markovian closure models, *Physica D*, *297,* 33–55, **Crossref**

Lobkovsky, L. I. (1982), The model of seismic gaps and catastrophic earthquakes in island arcs, *Proceedings of 5th School of marine geology. A. P. Lisitsin (ed.), Vol. 2* p.41–42, P. P. Shirshov Inst. of Oceanology RAS, Moscow.

Mau, V. (2015), Economic Crises in the Recent History of Russia, *Economic Policy*, No. 2, 9–32.

Medema, M. H., M. A. Fischbach (2015), Computational approaches to natural product discovery, *Nature Chemical Biology*, *11,* No. 9, 639–648, **Crossref**

Metz, B., O. Davidson, H. De. Coninck, M. Loos, L. Meyer (2005), IPCC Special Report on Carbon Dioxide Capture and Storage, Intergovernmental Panel on Climate Change, Working Group III, Geneva, Switzerland.

Nelson, Ch., et al. (2014), ACCAM global optimization model for the USCG aviation air stations, *Proceedings of 2014 IIE Industrial and Systems Engineering Research conference (ISERC2014)* p.1–10, Institute of Industrial & Systems Engineers, USA.

Odintsova, A., et al. (2017), Dynamics of oil and gas industry development in the 20th century using the world's largest deposits as an example: GIS project and web service, *Geoinformatics*, No. 4, 2–6. (in Russian)

Parsons, M. A., P. A. Fox (2013), Is Data Publication the Right Metaphor?, *Data Science Journal*, *12,* WDS32–WDS46, **Crossref**

Rajendra, A., S. Priti (2009), *Knowledge-Based Systems, 1 edition*, 354 pp. Jones & Bartlett, USA.

Reissell, A. (2016), IIASA Arctic Futures Initiative and Finland, Country of/on Extremes?, *Geoinformatics Research Papers*, *4,* BS4002, **Crossref**

Rybkina, A., et al. (2016), Development of geospatial database on hydrocarbon extraction methods in the 20th century for large and super large oil and gas deposits in Russia and other countries, *Russian Journal of Earth Sciences*, *16,* No. 6, ES6002, **Crossref**

Science International (2015), *Open Data in a Big Data World*, 4 pp. International Council for Science (ICSU), International Social Science Council (ISSC), The World Academy of Sciences (TWAS), InterAcademy Partnership (IAP), Paris.

Sheremet, I. A. (2013), *Augmented Post Systems: The Mathematical Framework for Knowledge and Data Engineering in Network-Centric Environment*, 215 pp. EANS, Berlin.

Vaisberg, L. (2015), Mehanika of loose media under vibration effects: methods of description and mathematical modeling, *Enrichment of Ores*, *4,* 21–31.

Wang, H., K. Sivonen, D. P. Fewer (2015), Genomic insights into the distribution, genetic diversity and evolution of polyketide synthases andnonribosomal peptide synthetases, *Curr. Opin. Genet. Dev.*, *35,* 79–85, **Crossref**

Wilkinson, Mark D., et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data*, *3,* 160018, **Crossref**

Zhang, Q., J. R. Doroghazi, X. Zhao, M. C. Walker, W. A. Van der Donk (2015), Expanded natural product diversity revealed by analysis of lanthipeptide-like gene clusters in Actinobacteria, *Applied and Environmental Microbiology*, *81,* No. 13, 4339–4350, **Crossref**

Zlotnicki, J., J. L. Le Mouel, A. Gvishiani (2005), Automatic fuzzy-logic recognition of anomalous activity on long geophysical records: Application to electric signals associated with the volcanic activity of La Fournaise volcano (Reunion Island), *Earth And Planetary Science Letters*, *234,* No. 1–2, 261–278, **Crossref**

E. Firsova, A. Gvishiani, R. Krasnoperov, A. Rybkina and O. Samokhina, Geophysical Center of the Russian Academy of Sciences, 3 Molodezhnaya St., 119296 Moscow, Russia. (a.rybkina@gcras.ru)

S. Hodson, Committee on Data of the International Council for Science (CODATA), 5 rue Auguste Vacquerie, 75016 Paris, France

P. Kabat, International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1 A-2361 Laxenburg, Austria