# ARCTIC TERRITORY – DATA AND MODELING

*A. Rybkina*[1], *A. Reissell*[2]

[1] Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia),
[2] International Institute for Applied Systems Analysis (IIASA, Austria)

a.rybkina@gcras.ru

International, independent and non-political approach should be provided for the better understanding of the Arctic future and its current needs. It is vital to focus on the complexity of entire region and interconnections of its geological, biological, atmospheric and geographic features. Imbalance within its systems will lead to the catastrophic effects that will reflect on the world systems. Due to its unique geo-systems the integrated systems approach is vitally needed. The presented study will focus on the complex assessment of the status of the Arctic territory and its Future Scenarios. The GIS-base project is developing to combine verified Arctic data to reflect the existing models and identify the gaps. It includes geological, geophysical and geographical data to provide systems approach for further analysis and data interpretation.

# TO THE RE-USING OF DATA ON ROCKBURSTS FOR UP TO DATE RESEARCH OF THE MINING SEISMICITY PROBLEM

*A. Batugin*

National University of Science and Technology “MISiS” (NUST “MISiS”, Russia)

as-bat@mail.ru

Presently, the man-caused seismicity problem is felt in many regions worldwide with an increasingly intensity. Regions with man-caused seismicity have formed in Russia, China, USA, Poland and other countries. The patterns of occurrence and display of the mining seismicity have been insufficiently ascertained and studied and the instructions and guidance documents are mainly focused on the handling of common rockbursts. It creates a situation where the mining seismicity suddenly occurs in increasingly more new regions causing material, social and environmental damage. At the same time, major geodynamic events and sudden seismic intensifications remain statistically seldom events, which renders their study and a search for patterns more difficult. It is also facilitated by the uniqueness of conditions for a manifestation of such geodynamic events and often the information unavailability.
However, rockbursts have been known since the late 19th century and the major ones that would be considered induced earthquakes according to the modern classifications happened as early as in the first half of the 20th century. In the published written works and specialized catalogues, there is some information regarding conditions of their manifestation, but their analysis performed from the viewpoint of the geomechanical advances of that time. It seems that a re-analysis of the data concerning occurrence of strong rockbursts from the modern perspective could be conductive to a better and deeper understanding of the origins and mechanisms of induced earthquakes happening today. In this relation, there arises a task of re-using the rockburst data to search for the patterns of the mining seismicity manifestations that have not been determined as yet.
The presentation analyzes the structure and content of specialized catalogues and global bibliographical guides to rockbursts that were published in the USSR and Russia. For instance, the bibliographical guide to rockbursts by I.M. Batugina and I.M. Petukhov present structured abstracts of over 1,000 articles out of the global scientific literature published 1900 through 1980. The data from a number of published works can be considered as a confirmation of some presently studied patterns that are related to the re-activation of tectonic faults, existence of tectonically stressed areas, and other things. A mechanism of a tectonic rockburst at the South Ural bauxite deposit that was unidentified at the time is looked into as an example.
The following can be stated among the prerequisites to the re-use of the rockburst data:

- existence of a large scope of text and image information concerning the conditions of rockburst manifestations at various deposits;
- some of the rockburst manifestation information has been analyzed and structured;
- scientific center Geodynamics of Subsurface at Moscow Institute of Steel and Alloys holds a collection of all articles dated 1990 through 1980 on rockbursts at the deposits worldwide;
- there is experience of the data re-use at individual deposits (Durrheim R.J., Kozyrev A.A. and others).

The following can be named among the main trends of the data re-use:

- performance of data generalization and pattern detection;
- conduct of studies of geodynamic zoning, analysis of tectonic stress fields, etc. in the areas of strong rockbursts' manifestation;
- data re-analysis taking into account the outcomes of supplementary works.

To conclude, there are opportunities for a re-analysis of data on the rockburst manifestations at the deposits worldwide to extend the statistical basis of the search and determination of patterns of the mined seismicity manifestations.

# APPLIED PROBLEMS IN BIG DATA ANALYSIS FOR METALLURGY

*M. Nezhurina, S. Solodov*

National University of Science and Technology “MISiS” (NUST “MISiS”, Russia)

nezhurinam@gmail.com

IT-solutions implement, related with big data analysis, become more relevant for big full cycle metallurgical companies. A big number of challenges related with increasing operational efficiency and reducing production costs were solved with ERP and MES systems large-scale deployment. However, these platforms cannot allow operational mode to analyze a large amount of information, related with product quality and equipment status. At the moment, this functionality can be provided with SAP HANA systems.

Using this system for product certification quality assessment, allows to compare a set of product requirements and current conditions of the technological process in the real-time scale. Informational providing of the system, is building on the established by the company all-union and republican-registered, sectoral classifiers and standardized documents base. Incoming information - is incoming data from SCADA and MES systems. The system functionality implements the product certification algorithms and reassigned with a non-closed material need order, which allows to reduce customers complains number, to fulfill the order in the stipulated time with a client, based on assessment and quality prediction of steel products and to increase the suitable metal yield. Thanks to the optimal RAM memory usage opportunities and innovative proposals, SAP HANA technologies increasing data processing in-memory speed and give the unique opportunities for analysis and through-control of all products quality in the real-time.

# APPLICATION OF MACHINE LEARNING TECHNOLOGIES FOR THE OPTIMIZATION OF PRODUCTION PROCESSES IN METALS AND MINING

*A. Khaitin*

Yandex Data Factory

samuylova@yandex-team.com

Rapid growth in the amounts of accumulated data, availability of computing power, development of machine learning algorithms and practical experience of their applications to commercial and industrial tasks, have led to the possibility of wide-scale use of these new technologies in the metals and mining sector. At the same time, due to the novelty of the technology and lack of recognized examples, the specifics of the technology application remain largely unknown to the practitioners. We argue that metals and mining are among the most attractive sectors for the application of these technologies. This is due to the combination of stable, repetitive processes, that allow both accumulating the necessary data and applying machine learning techniques to the routine decision-making, and importance of optimization that is typical for capital-intensive industries. Building upon the experience of Yandex Data Factory, we will provide an overview of the practical applications of machine learning to metals and mining. One such example includes our collaboration with Magnitogorsk Iron & Steel works. For them, we created a machine learning service that recommends the optimal combination of ferroalloys needed to produce specific steel grades with international standard chemical compositions, and at the lowest cost for each specific smelting. The service uses the seven years of historical steelmaking records. Use of machine learning techniques allows forecasting the deviations of traditional physics-based models, and significantly increases the accuracy of the forecast. The pilot tests in a production environment have demonstrated that the use of machine learning-based recommender system allows decreasing the use of ferroalloys by up to 5%. Similarly, machine learning techniques can be applied to other tasks in metals and mining, augmenting traditional statistical approaches and physics-based modelling. Such possible applications include: optimisation of raw material use, such as use of cyanide during ore beneficiation, dynamic control of process parameters during both upstream stages and further processing, virtual sensing, and others. For the successful choice of the use cases it is essential to identify the intersection between technology applicability and expected economic benefits. Not to be overlooked, the use of machine learning techniques also brings certain challenges related to the organizational and managerial aspects. Since machine learning models are not interpretable to humans due to their complexity, it is only possible to measure the results of their applications through carefully organized experiments. This experimental approach should be integrated into existing business culture, with its ability to correctly identify success metrics, design experimental procedure, carry out measurements etc. In addition, automation of decision-making brings new managerial questions, such as how the responsibilities should be assigned. Such matters should not be excluded from the discussion of the practical use of machine learning technologies in favour of more technical aspects, since to a large extent the successful integration of the new technologies in the industrial sectors relies on the ability of organization to adapt their usual business practices.

# BUILDING MODERN DATA PLATFORM BASED ON OPEN SOURCE PROJECTS

*S. Zolotarev*

IBS Company

SZolotarev@ibs.ru

Today more and more enterprises consider use of 100% open source components as the wise choice for building a modern Data Platform for Analytic and Data Analysis . Legacy analytical platform or database are costly, slow and not able to scale at the speed modern business requires.

# NEW QUALITY OF PREDICTIVE ANALYTICS BY USING NEW DATA SOURCES: SOLUTIONS AND EXPERIENCE

*M. Ageykin, V. Shvey, S. Shvey*

JSC "EC-leasing"

mageykin@ec-leasing.ru

According to IBM strategic forecast, all companies in the next 5 years will be divided into winners and losers depending on quality of making corporate decisions. Research and case studies provide evidence that a well-designed and appropriate computerized decision support system can encourage fact-based decisions, improve decision quality, and improve the efficiency and effectiveness of decision processes. There is resource that we all have aplenty: a large amount of open data, both structured and unstructured. This report introduces the concept of acquiring data from big data sources such as social media, news, mobile and smart devices, weather information, and information that is collected via sensors and using this data to get new quality of predictive analytics. Predictive analytics today in many companies is perceived as an evolutionary step in business analytics and is used primarily to build a forecast based on the same data on which reports are built. Nevertheless, this does not take into account the enormous importance of external factors in forecasting and nowcasting. In this report will be shown cases from various areas where external data are the basis for predictive analytics and allow gain results unattainable to the forecast based only on enterprise data. Multiple scenarios for storing, handling, preprocessing, filtering, and exploring big data in predictive analytics are described. Techniques are described that help to form homogeneous (homogeneous) groups of data, based on data from various sources. An important part is working with text documents and gaining information for use in predictive modeling. At the end of the report you can see examples of predictive models based on external data which allow the companies using them to generate additional profits and outrun their competitors.

# USING BIG DATA IN PROCESS INDUSTRIES WITH SAP

*A. Knjazhev*

SAP Data Science

alexey.knjazhev@sap.com

SAP is worldwide IT leading company with revenue more than 22 Bln euro. In Russia SAP operating since 1997. Most of largest Russian companies are using SAP Business Applications. Specializing on business applications It has large experience in practical using of Big Data in process industries.

Today 90%-95% of Big Data use cases connected to predictive scenarios. There are a lot of other scenarios:- Predictive Maintenance;- Predictive Quality;- Human Capital Management (predictive of candidate compatibility with Company culture, predictive of compatibility of brigade members, predictive of outflow of employees, education recommendations);- Contractors (predictive of paying delays, predictive of supply delays);- Fraud;- Sales forecasting;- Health and safety;- Inventory Optimization.

The most useful in Process industries are two scenarios: Predictive Maintenance and Predictive Quality. The main reasons of that: - There is large competition today between companies on the market. Buy proposing better quality companies receive competitive advantages.- Process industries are resource-intensive industries.- If you know something before you can do it more effectively.- Historically key machines are good equipped by PLCs.- There is good statistics of downtimes, maintenance and quality issues presented.

SAP will present several Predictive Maintenance and Predictive Quality projects which was realized as in Russia as in other countries:- Predictive of defects of slabs after casting machines;- Predictive of breaking of paper on paper producing machine;- Planning of maintenance in transport company.

These cases were realized with SAP Predictive Analytics – special solution which helps to define relationship between goal variables and PLCs data. As result analyst receives ready-to-use model with regression analyses of variables, correlation analyses of variables, estimation of probability and stability of designed model. Based on SAP HANA technology solution allows to do analyses fast for large models independently on quantity of variables.

Predictive of defects of slabs after casting machines allows to select slabs with the most probability of defect to control. Also SAP Predictive Analytics allows to predict as defect itself as type of defect with good probability. Using recommendations of system allows to decrease time of quality control on 25% and detect more defects by selecting right slabs. To predict of breaking of paper on paper producing machine information are using from more than 500 PLCs. With SAP Predictive Analytics quantity of breaks was decreased on 50%.By using comparing information from PLCs with etalon condition of device (accumulator battery) transport company decreased maintenance time on 40%. SAP will also present their vision how to organize Big Data project to receive right result in right time.

# MATHEMATICS FOR/OF BIG DATA

*V. Zaborovskiy, L. Utkin*

Polytechnic University, St.Petersburg

vlad@neva.ru

One of new cultural and technical phenomena of modern Society can be named as datamani or tsunami of data of fourth industrial revolution. This "big wave" tackle a wide range of social, economic, industrial and scientific challenges, already covered the hi-tech companies and will soon can swallowed up all the inhabitants of the planet. Somebody think that leading source of Big Data is the Internet that added 1021 bytes (a zettabyte) of information every year and causes new noumena (thing-in-itself) called as "Big Data's Mysteries". This new essence seems to be mathematical in nature, but reflects deep feature of matter and can be "mining" by computer tools and technologies. One of the mysteries associated with these phenomena is that Machine learning systems works spectacularly well, but specialists - both programmers and mathematicians cannot explain why. Modern Internet search engine is estimated to seek information from around 15 exabytes (1015 bytes) of data. People judges the importance of a data on the basis of the importance of events that caused or linked to it but what is "clever" algorithm are used for this search machine, how computers calculates the relative importance of a data , and ignoring all other? How Data can be transfer into knowledge - the key questions of AI paradigm. In the presentation we discusses modern approaches that using now to extracting meaningful knowledge from the available data. This is not a trivial task and represents a severe challenge in which Mathematics plays an important role through techniques of statistical learning, signal analysis, distributed optimization, compress sensing etc. We are focused our attention on basic ideas of Mathematics for Big Data applications and explains why Big Data itself is a source of new Mathematics that base on several not trivial formal essence like dimension of data set, volume of data, depth of data (like median) and operations like sum of the clusters if data is or nor Gaussian, representing groups of points by their sufficient statistics and extract these points from RAM. At the end of the discussion we stress that opportunities to educate and train professionals to use Big Data paradigm open the gate to new business, engineering and science opportunities.

# PRACTICAL ASPECTS OF MACHINE LEARNING ALGORITHMS APPLICATION TO BIG DATA ANALYSIS

*A. Terekhov*

SPbU

a.terekhov@spbu.ru

The article is devoted to show how some practical tasks in different target domains are solved with wide usage of machine learning algorithms for big data analysis. The main problem that faces the researcher in this field is to choose from the variety of known approaches one which suits the specified task and adopt it to gain maximum efficiency. In addition with the cases which were solved some open problems are described. The article consists of several chapters which reveal the results of the investigations done on the Software Engineering Chair of SPbSU.The first one concerns well known task of medical data analysis. Nowadays many medical institutions store a large amount of knowledge (history, diagnoses) which were accumulated for years or even decades. This data can be used for various practical purposes - for example, to verify appointments, to make recommendations, to train the system to check new patients for any known diagnoses, etc. One of the important problems is the data mining from the human language. Diagnoses, anamnesis, etc. are usually written in free form. Here just the algorithms of machine learning are applied. Such solutions are currently a very hot topic, in the western countries there are hundreds of companies developing software in this area. But it's hard to transfer it in our country because of Cyrillic alphabet used in our clinics.
One more important medical problem is described in the next chapter. Deep learning methods are effectively used for various tasks of image analysis. When analyzing fluorography, specialists process a large volume of images, more than 95% of which do not contain pathologies. There is a task - to filter out images that do not exactly require specialist's attention. With the help of deep learning algorithms we could check if there are suspicious areas in the picture, where potentially pathology can exist.
The next chapter observes certain use cases of big data implementation from another domain. There were several experiments with data that mobile operator has at its disposal. The first use case includes the analysis of most popular clients' routes during special days. This analysis helps to optimize advertising placement and increase the effect of advertisement campaigns for the Operator. This case is successfully solved and implemented. Two other cases are at an initial stage. The first of them – is a case that studies the dependence of clients satisfaction index (CSI) and average data/voice traffic and monthly revenue of a client. The second one – is a case that studies the clients clustering depending upon their data traffic usage by applications.
Finally the problem of verifying the authorship of manuscripts is described. The problem is very common now. The expert approach, widespread in the verification of the authorship of historical manuscripts, has an obvious lack of objectivity. In particular, Professor Noar Gardiner at the University of Michigan library found a copy of the third part of the work "Description of Egypt" (al-Hitat) of the ancient Egyptian historian and geographer of the period of Mamelukes Takiyuddin al-Makrizi. In his work, Professor Gardiner made the assumption that the found manuscript could be a draft copy written by Al-Makrizi himself. Later, a well-known expert on the works of al-Makrizi, Professor Frederick Bauden from the University of Liege, identified the autograph as the final version of the third part of al-Hithat and, thus, the only final version of al-Hithat found so far. Proceeding from this, the task of developing automatic algorithms for verifying the authorship of manuscripts seems to be very relevant. Now automatic methods for verifying the authorship of Arabic manuscripts have focused on the calculation of various kinds of features on the basis of the manuscript image, which were then fed to the classification algorithm. An obvious drawback of such approaches is the need to

manually select the method for calculating traits, which, firstly, is not optimal, and secondly, it requires considerable time. An alternative approach based on convolutional neural networks automatically selects the most optimal features due to the parameterization of their calculation and, thus, is devoid of the indicated drawbacks. In practice, this approach is advanced in many areas related to image processing, where it surpasses in quality methods based on manual choice of the method of calculating traits. Thus, the method based on the convolutional neural network architecture AlexNet with a fixed moving window showed positive results on the verification of the authorship of the manuscripts of al-Makrizi.

# TOOLS FOR AD-HOC ACCESS, VISUALIZATION AND ANALYSIS OF SPACE MISSION TELEMETRY AND SCIENCE DATA

*C. Pankratz, F. Sanchez, D. Lindholm, A. Wilson, R. Christofferson, T. Baltzer, D. Osborne*

Laboratory for Atmospheric and Space Physics,
University of Colorado, Boulder, CO, 80020, USA

chris.pankratz@lasp.colorado.edu

The ability to access, view, and analyze both telemetry and science data originating from spacecraft and space instruments is essential to enable the success of a space mission and extracting the full benefit of its science results. Mission engineers and operators need to understand the health, safety, and performance of flight systems, and scientists need efficient and powerful ways to discover, access, survey and understand science results. Telemetry data originate from many different types of subsystems and take many different forms, from sensor measurements to onboard software status. Similary, science products can take many different forms that may challenge science community members to easily use the data within their chosen analysis environments. The variety and abundance of these data creates challenges for efficient access and assessment, and demands the creation of tools for analysis to serve all aspects of space flight operations and research. LASP (the Laboratory for Atmospheric and Space Physics) has developed a number of approaches and tools that address many of these issues. These take the form of versatile, browser-based, project-independent tools that can be resident on institutional servers or hosted in the Cloud, and can be configured to unify access to disparate data sets that may exist in other data centers. This presentation will describe general approaches and three such tools. The first is called WebTCAD (Web-based Telemetry, Checking, Analysis, and Display) that enables ad-hoc interactive visualization and analysis of telemetry data from spacecraft and instruments, including engineering and state parameters. The second is an integrated web site called the LASP Interactive Solar Irradiance Datacenter, LISIRD, http://lasp.colorado.edu/lisird, which allows web-based exploration and access to solar irradiance and related data sets using convenient, interactive or scriptable, standards-based interfaces. The third tool is a "middleware" layer called LaTiS, which provides an enabling technology for flexible, standards-based serving of data to web sites, applications, and user tools. LaTiS maps data sets to a common model, enabling the rapid time-based queries for scientific analysis, permitting access to disparate data at potentially geographically separated locations.

# GEOMAGNETIC SURVEY FOR DIRECTIONAL DRILLING OF DEEP WELLS IN THE ARCTIC REGION

*R. Lukianova, A. Gvishiani*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

a.gvishiani@gcras.ru

Modern technologies allow drilling wells which are almost horizontally entering the oil-bearing layer. Such wells can reach reservoirs located a few kilometers from the starting point of drilling, including those under the seabed at a considerable distance from the shore. Directional drilling requires precise control of the drill string orientation underground. To measure the geographic azimuth telemetric systems based on economical magnetometric technology are often used. Precise following the assumed direction of drilling, correction of deviations and verification of the magnetometric data of the downhole magnetometer-inclinometer is an important task to ensure the specified drilling parameters. Minimization of errors due to different-scale variations of the Earth's magnetic field is achieved by applying the modern models of the main magnetic field and local magnetic anomalies as well as the aeromagnetic surveys. In the Arctic region, additional problems arise with geomagnetic survey for directional drilling due to the global structure of the Earth's magnetic field and the influence of solar activity factors. In the high latitudes, large sporadic perturbations of the geomagnetic field occur due to powerful (up to millions Amperes) electric currents which are developed in the Earth's ionosphere during magnetic storms. Geomagnetic disturbances caused by magnetic storms introduce a greater uncertainty in the records of downhole inclinometers. To solve the problem it is necessary to conduct a parallel independent geomagnetic monitoring at observatories located in the vicinity of the area of conducting drilling operations with the aim of filtering the storm time magnetic disturbances. Using the conventional methods for calculation the actual borehole profile, model evaluations of the effect of sporadic disturbances of magnetic declination observed during magnetic storms on the displacement azimuth and the intensity of the borehole curvature in the specified locations were performed. It is shown that if geomagnetic disturbances are not filtered on the basis of data from parallel ground based observations, they can lead to unacceptably large errors in borehole inclinometric measurements and deviation of the trajectory of the borehole by a magnitude exceeding the permissible values. The need for magnetic survey for directional drilling of deep wells dictates the need to deploy full-scale magnetic observatories on the coast of the Arctic Ocean.

# MODELING OF TECHNO-SOCIAL SYSTEMS: SEMANTIC APPROACH

*S. Maruev*

Russian Presidential Academy of National Economy and Public Administration (RANEPA)

maruev@ranepa.ru

Nowadays, most of the digital content is generated within techno-social systems where people are connected to other people as well as to artefacts such as documents and concepts. These networks provide a rich context for understanding the role of particular nodes. This paper focuses on network models of techno-social systems which are suitable for aggregation of heterogeneous information and processing by graphs theory methods. We describe a life cycle of models and methods of analysis as well as developing systems for guidance and navigation. We give a generalization of the multidimensional network as a formal framework of representing reality; as a framework, particularly suitable for representing bottom-up created knowledge, shifting the focus of representation from specification and conceptualization to capturing empirical data about the objects at the instantiation level. The conceptualization is substituted with limited abstractions of phenomena taken from a particular techno-social system. Graph-based methods, introduced in this paper, can be perceived as a kind of "fuzzy inferencing", allowing the solving of certain problems based on noisy and incomplete data. The sheer volume of the data in techno-social systems enable to make conclusions as viable as conclusions achieved by the traditional logical tools. A few use cases demonstrate the use of the method for building a thesaurus, determining the goods, grouping the interests of users of the social network. Some of them are "the polar" use cases in terms of the essential properties of the algorithms providing accurate prediction. For recommended social systems most reliable features should dominate the results trumping less reliable predictors; which in terms of fuzzy logic operations could be expressed as follows: the results of the operation AND should be highly skewed towards the value of the bigger operand. The logical aggregation skewed towards the value of the bigger operand spectacularly fails for the other use cases, which is the prediction of the item based on the words in the textual description. Authors demonstrate that for this use case, significant number of weak predictors easily overrules the judgment based on strong predictors. Since the volume of data in most of the techno-social system falls into the category of Big Data, in this paper we provided a short overview of only those network mining methods which are scalable. We demonstrated that computationally efficient and scalable algorithms usually are based on the iterative computational scheme where on each iteration only connections between neighbour nodes are used. We show how to use graph models of techno-social business environment for problems of natural language understanding and for new objectives such as recommended social systems to reduce users' information overload by relevant content recommending.

# SITE EFFECT IN ARCHAEOLOGICAL CITY JRASH IN JORDAN

*W. Eidolimat*

Jordan Seismological Observatory

w.olimat@gmail.com

In this study we determined the local site effect by means of the horizontal-to-vertical (H/V) spectral ratios. The Nakamura's concept (Nakamura Y. 1989) is applied in the important historical city JERASH in order to determine the resonance frequencies and amplification factors for, finding the dynamic characteristics for structural engineering purposes. Results obtained in this study shows that; dominant frequencies F varies between 2.2 Hz and 3.19 Hz in the city area, while, the amplification factor A varies between 0.87 and 23.47. This means that structural culture in most localities of the study area might be seriously affected by any of eventually major short periodic earthquakes released by the nearby seismologic active sources westward, except of localities characterized by long periodic dominant frequencies of the study areas, considering that most structures of the study area are characterized by one to three story profiles.

At the investigated area, two maps were prepared that show the spatial variations of the predominant period and seismic amplification according to Nakamura's technique. The analysis shows that areas with thick layer of sediments in these cities have relatively high predominant periods and high seismic amplification compared to the areas with thin layer of sediments. This result is in line with the theory of topographic effects on seismic amplification and also confirmed the suitability of the H/V spectral ratio of ambient noise as a geophysical exploration tool in seismic hazard assessment.

# DATA DRIVEN KNOWLEDGE-BASED SYSTEMS: TOWARDS DEVELOPMENT OF THE MULTIFACTORIAL COMPUTATIONAL MODELS OF THE ENERGETIC MATERIALS COMBUSTION

*V. Abrukov[1], A. Lukin[2], M. Kiselev[3], D. Anufrieva[1]*

[1]Chuvash State University,
[2]Western-Caucasus Research Center,
[3]Megaputer Intelligence

lukin@wcrc.ru

Nowadays a solution of many problems that could not be solved early has become possible by means of Data Science. One of such problems is a development of a new sort of Knowledge-Based System (KBS). Under the new sort of KBS we mean the calculation tool that includes the following advantages: - contains all relationships between all variables of the object; - allows to calculate the values of one part of variables through others; - allows to solve direct and inverse problems; - allows to predict the characteristics of an object that have not been investigated yet; - allows to predict technology parameters to obtain an object with desired characteristics. The base of the new sort of KBS is an ensemble of multifactor quality, quantity and computational models. In the present paper we demonstrate possibilities of the Data Science Methods application in the generalization of the connections between the variables of combustion experiments as well as in forecasting of "new experimental results" with using the Artificial Neural Networks (ANN), which are one of the most promising methods of the Data Science. We provide the review of previous modeling efforts with using of the Data Science Methods, [1-6]. The ANN allows modeling temperature profiles in propellant burning waves, predicting burning rate of various propellant compositions for different ranges of pressure and initial temperature, determining propellant composition providing a necessary burning rate for various pressures and temperatures. The preliminary results, earlier obtained by our research team shows that ANN can be considered as a good approximation tool for experimental functions of several variables for the study of combustion behaviors, as a more affordable way to receive additional novel experimental results, and as a good tool to present to scientific community the experimental results obtained. Performance of the energetic materials can be enhanced significantly with the addition of catalysts. When the catalysts are in the nano size range, they offer unique advantages which can eventually enhance performance of of energetic materials. In this connection, the first multifactorial computational models of the combustion of energetic materials with nano additives will be presented. The models are based on the experimental results presented in the following review paper [7].

Acknowledgment
The reported study was funded by the Russian Foundation of Basic Researches (RFBR) according to the research project No. 16-53-48010.

References

1. Abrukov V.S., Troeshestova D.A., Chernov A.S. et al. Application of Artificial Neural Networks for Solution of Scientific and Applied Problems for Combustion of Energetic Materials, Advancements in Energetic Materials and Chemical Propulsion, Ed. by K.K. Kuo, J.D. Rivera, Begell House Inc. of Redding, 2007. P. 268-283.
2. Abrukov, V.S., Malinin, G.I., Volkov, M.E., Makarov, D.N., and Ivanov P.V. Application of artificial neural networks for creation of "black box" models of energetic materials combustion, 2008, In K.K. Kuo and K. Hori, Eds., Advancements in Energetic Materials and Chemical Propulsion, USA, Connecticut: Begell House Inc. of Redding, pp. 377-386.

3. Abrukov V.S., Karlovich E.V., Afanasyev V.N., Semenov Y.V., Abrukov S.V. Creation of propellant combustion models by means of data mining tools. International Journal of Energetic Materials and Chemical Propulsion. 2010. v. 9. No. 5, pp. 385-396.
4. Abrukov V.S., Karlovich E.V., Abrukov S.V. Development of computing models of propellant combustion by means of data mining // 2010 Research Bulletin of the Australian Institute of High Energetic Materials, Vol.2, April, 2011, pp. 129-144.
5. Abrukov V.S., Troeshestova D.A., Abrukov S.V., Karlovich E.V., Polykarpov, A.I., New Combustion's Experiment: Data Mining for Modeling and Creation of Knowledge Base. Conference Paper (electronic format): Tenth International Symposium on Special Topics in Chemical Propulsion (10 - ISICP), At ENSMA - Poitiers (France), 2014, 21 pp.
6. Victor Abrukov, Valery Kochakov, Alexander Smirnov, Sergey Abrukov, Darya Anufrieva. Knowledge-Based System is a Goal and a Tool for Basic and Applied Research. Conference Proceedings of 9th International Conference on Application of Information and Communication Technologies – AICT (14-16 October 2015, Rostov-on-Don, Russia), The Institute of Electrical and Electronics Engineers, Inc., 2015, pp. 60-63.
7. Qi-Long Yan, Feng-Qi Zhao, Kenneth K. Kuo, Xiao-Hong Zhang, Svatopluk Zeman, Luigi T. DeLuca. Catalytic effects of nano additives on decomposition and combustion of RDX-, HMX-, and AP-based energetic compositions. Progress in Energy and Combustion Science 57 (2016) 75–136.

# DEVELOPMENT OF THE MULTIFACTORIAL COMPUTATIONAL MODELS OF THE ENERGETIC MATERIALS COMBUSTION BY MEANS OF DATA SCIENCE METHODS

*V. Abrukov*[1], *A. Lukin*[2], *M. Kiselev*[3], *D. Anufrieva*[1]

[1]Chuvash State University,
[2]Western-Caucasus Research Center,
[3]Megaputer Intelligence

abrukov@yandex.ru

In the present paper we demonstrate possibilities of the Data Science Methods application in the generalization of the connections between the variables of combustion experiments as well as in forecasting of "new experimental results" with using the artificial neural networks, that are one of the most promising methods of Data Science. A review has been carried out for the previous modeling efforts. The artificial neural networks allows modeling temperature profiles in propellant combustion waves, predicting burning rate of various propellant mixtures for different ranges of pressure and initial temperature, determining propellant mixture providing a necessary burning rate for various pressures and temperatures. The preliminary results, earlier obtained by our research team depict that artificial neural networks can be considered as a good approximation tool for experimental functions of several variables for the study of combustion behaviors, as a more affordable way to receive additional novel experimental results, and as a good tool to present to scientific community the experimental results obtained.
Acknowledgment The reported study was funded by the Russian Foundation of Basic Researches (RFBR) according to the research project No. 16-53-48010.

# IMPLEMENTATION OF BUSINESS INTELLIGENCE SOLUTION IN SUPPORT OF DATA ANALYSIS, INFORMATION AND KNOWLEDGE MANAGEMENT

*R. Mabaso*

Air Traffic and Navigation Services

refiloema@atns.co.za

Business Intelligent Tool is designed with the intention to provide a unified communications platform for sharing tacit and explicit knowledge derived from Business Intelligence (BI) and Knowledge Management (KM) systems. In these systems, data, documents, videos, knowledge experts and decision models can be identified, mapped and targeted to address new situations and provide information that will lead to making informed decisions. BI has a significant and vital role to play in the application of KM initiatives. BI applications analyze business operations and produce information to help business users understand, improve and optimize business operations. There is a distinct explanation between BI and KM, Gartner explains BI as a set of all technologies that gather and analyze data to improve decision making. In BI, intelligence is often defined as the discovery and explanation of hidden, inherent, and decision-relevant contexts in large amounts of business and economic data. KM is described as a systematic process of finding, selecting, organizing, distilling and presenting information in a way that improves individual's comprehension in a specific area of interest. KM supports an organization to gain insight and understanding from its own experience and utilizing knowledge for such things as problem solving, dynamic learning, strategic planning and decision making. KM is becoming an increasingly important business resource not only in the development and innovation process, but also in securing an organization's competitiveness and survival in the environment. In view of the fact that there is no full consent about adopting a universal definition, knowledge management can be regarded as a concept encompassing methods of simplifying, enhancing, sharing, distributing and creating knowledge within an organization. The dominant approach of contemporary business operations is based on BI, which, including Data Warehouse as its integral part, represents a highly significant component in KM. As the application of information technologies provides the key support in conducting all business activities, from operative to highly complex ones, this presentation will outline the significance and functional application of BI in the KM System. BI activities should provide knowledge improvement

# OPEN GOVERNMENT DATA (OGD) USAGE AMONG WORKING PROFESSIONALS FROM PRIVATE SECTOR AND NGOS IN INDIA

*S. Saxena*

Central University of Haryana

stuti.razia@gmail.com

While the implementation of Open Government Data (OGD) initiatives in different countries remains asymmetric across the globe, the present study underscores the significance of re-use of OGD in a developing country's context. India is an emerging stage of OGD implementation. However, the usage of OGD in the form of data or information has been active to some extent among the end-users. Conceding that one of the prime objectives of OGD initiative is to facilitate the re-use of data sets by the users to create public value in the form of innovation, this study undertakes a qualitative study to investigate the manner in which OGD is used by the users in India. Specifically, we tap the working professionals from private sector and NGOs to appreciate the utility of OGD for them in their work. Also, we underline the challenges indicated by the working professionals in tapping OGD for their professional purposes. Findings show that while OGD is harnessed for different purposes (report writing and annual statement compilation, for instance), there are many working professionals who do not perceive any utility of OGD for reasons like incomplete, incoherent and inconsistent data sets. The present study is a significant contribution towards understanding the "supply" side of OGD conceding that the "demand" side of OGD has been extensively covered in extant literature.

# GENERIC SPACE SCIENCE VISUALIZATION IN 2D/3D USING SDDAS

*J. Mukherjee*

Southwest Research Institute

joey@swri.org

The Southwest Data Display and Analysis System (SDDAS) is a flexible multi-mission / multi-instrument software system intended to support space physics data analysis, and has been in active development for over 20 years. For the Magnetospheric Multi-Scale (MMS), Juno, Cluster, and Mars Express missions, we have modified these generic tools for visualizing data in two and three dimensions. The SDDAS software is open source and makes use of various other open source packages, including VTK and Qwt. The software offers interactive plotting as well as Python and Lua modules to modify or export the data before plotting. In theory, by writing a Lua or Python module to read the data, any data could be used. Currently, the software can natively read data in IDFS, CEF, CDF, FITS, SEG-Y, ASCII, and XLS formats. We have integrated the software with other Python packages such as SPICE and SpacePy. Included with the visualization software is a database application and other utilities for managing data and the software can natively retrieve data from the Cluster Active Archive, Space Physics Data Facility at Goddard/NSSDC, as well as other IDFS archives. Other archives site could be incorporated through our generic data server architecture as well. Line plots, spectrograms, geographic, orbit plots, volume rendering, strip charts, etc. are just some of the types of plots one can generate with SDDAS. In addition to visualizing the data, one can interact with the data through a GUI allowing one to modify the color scales, subset the data, or slice in multiple dimensions. Other methods of slicing the data include slicing based on magnetic field information (parallel or perpendicular cuts).This paper will present the data model behind the design of SDDAS and discuss some of the strategies used, and present a standard for doing additional generic 2D/3D data analysis.

# UNRAVELLING OCEAN DATA – AN INTRODUCTION TO SOUTH AFRICA'S FLOATING UNIVERSITY PROGRAMME "SEAMESTER"

*I. Ansorge*

University of Cape Town

isabelle.ansorge@uct.ac.za

South Africa's Department of Science and Technology's 10-year Global Change Grand Challenge programme requires platforms to 'attract young researchers to the region and retain them by exciting their interest in broad aspects of ocean data analysis' "SEAmester" – South Africa's recently established Floating University – achieves just that. "SEAmester" introduces marine science as an applied and cross-disciplinary field to students who have shown an affinity for core science disciplines. It identifies with the South African government's National Development Plan on education, training and innovation – critical to the country's long-term development and investment in this sector. "SEAmester" has a long-term vision aimed at building capacity within the marine sciences by coordinating and fostering cross-disciplinary research projects, access to a wide range of marine datasets and their handling -while achieving this goal through a highly innovative programme. The strength of "SEAmester" is that postgraduate students combine theoretical classroom learning with the application of this knowledge through ship-based, and more importantly, hands-on research. This paper provides an overview of the South African "SEAmester" Floating University programme and gives examples of how students are exposed through ship-board training steams to an extensive range of marine science datasets and the calibration steps necessary to ensure their accuracy. This paper also identifies, not only the advantages of exposing students to raw data, but also the challenges and limitations faced by these students.

# FORECASTING OF WORK OF SOLAR POWER PLANTS BY MEANS OF ARTIFICIAL NEURAL NETWORKS

*V. Abrukov*

Chuvash State University

abrukov@yandex.ru

Methods and technologies of creating of computational systems for forecasting of solar power plants (SP) work in dependence of meteorological conditions are presented. Artificial neural networks which are included into the analytical platform "Deductor", which has a wide range of methods for the pre-treatment of data, methods for preliminary data analysis, and data modeling, were used to create forecasting systems. The results of the unique continuous four-year multi-parameter monitoring of SP characteristics and meteorological conditions serves of the base of the computational systems created. The following variables: voltage, current in the load circuit and power, solar radiation, external temperature, humidity, dew point, wind speed, wind direction, wind chill index, heating index, index of "temperature + humidity + wind", index of "temperature + humidity + wind + solar radiation", atmospheric pressure, ultraviolet index, and index of evaporation have been registering by the monitoring system. Two characteristics of the SP work: the power density (the value of power related to the total area of the solar panels) and the value of the conversion coefficient of solar energy into electrical energy (the ratio of power density to solar radiation) were determined additionally. Two variants of the created multifactor computational models for forecasting the power density and the conversion coefficient in dependence of meteorological conditions are described. The first variant uses the full set of recorded meteorological conditions. The second variant uses the limited set of meteorological conditions available from the forecast of the Hydro meteorological Center of the Russian Federation. Examples of application of both variants of forecasting systems are given. The obtained forecasting systems can be used not only for the direct prediction of the SP work, but also for the regionalization of the territory of the Russian Federation from the point of view of the perspectives of the building of SP. The importance of this approach to forecasting lies in the fact that it is not a question of the generally accepted zoning of the territory of the Russian Federation from the point of view of the level of solar radiation, but of zoning in terms of power density and the conversion coefficient of solar energy into electrical energy. The proposed approach can be used to forecast of wind power plants work, work of thermodynamic installations using solar energy, and other devices of alternative energetic.

# THE KNOWLEDGE-BASED SYSTEM IS A GOAL AND A TOOL FOR BASIC AND APPLIED RESEARCH

*V. Abrukov[1], A. Lukin[2], M. Kiselev[3]*

[1]Chuvash State University,
[2] Western-Caucasus Research Center,
[3]Megaputer Intelligence

abrukov@yandex.ru

There are several important issues related to analysis and modeling of experimental data independently of the field of science and technology where they were obtained. These problems are as follows:

- What is the best way to generalize experimental data?
- What is the best way of generalization which allows solving direct and inverse problems?
- What is the best way of generalization which allows predicting results of an experiment which was not carried out?
- What is the best way of generalization which allows determining a technology of object creation with required properties and characteristics?
- What is the best way to represent the experimental results for a scientific community?

We do believe that all these issues can be solved by creating a Knowledge-Based System. Under Knowledge-Based System we mean the analytical and calculation tool that: contains all relationships between all variables of the object; allows to calculate the values of one part of variables through others; allows to solve direct and inverse problems; allows to predict the characteristics of an object that have not been investigated yet; allows to predict technology parameters to obtain an object with desired characteristics. This paper presents a methodology for a creation of the Knowledge-Based System of metamaterials based on nanofilms of linear-chain carbon by means of artificial neural networks [1-3] as well as the descriptions in brief of other examples of Knowledge-Based Systems in various fields of basic and applied research such as combustion processes [4], inverse problems of optics [5], socio-economic systems [6,7], etc. All steps of the creation of the Knowledge-Based System of metamaterials as well as and illustrations are presented on the Web-site of the paper [1] as the autonomous computer modules. Anyone can download the modules and execute research on their own and obtain all graphs which depict relationship between all variables of the concrete metamaterial.

The reported study was funded by RFBR according to the research project No.16-53-48010

References

1. Web-site of the paper, "Creation of the knowledge base of nanomaterials and nanotechnologies of Chuvash Republic", http://amf21.ru/biblioteka/meroprijatija-provodimye-associaciei/proekt-rffi-sozdanie-bazy-znanii-nanom, 2015. (in Russian)
2. Victor Abrukov. Knowledge base is a future of nanomaterials world. Abstracts of the 3rd International Conference on Nanotek & Expo, December 02-04, 2013 Hampton Inn Tropicana Las Vegas, USA //http://www.omicsgroup.com/conferences/ACS/conference/download-pdf.php?file=2912-Speaker-Pdf-T.pdf
3. V.S. Abrukov, V.D. Kochakov, S.V. Abrukov, A.V. Smirnov. "A new nanotech computational experiment: data mining for modeling, creation of knowledge base, and presentation", Abstracts of XII International Conference on Nanostructured Materials (NANO 2014), 2014, p.1, Moscow, Lomonosov State University, Russia, July 13 – 18, 2014.

4. V.S. Abrukov et al., "Creation of propellant combustion models by means of data mining tools", Int. J. of Energetic Materials and Chemical Propulsion, vol. 9, No. 5, p. 385, 2010.
5. Abrukov V.S., Pavlov R.A., Ivanov P.V., Troeshestova D.A. Artificial neural networks and inverse problems of optical diagnostics. Proceedings - ISDA 2006: Sixth International Conference on Intelligent Systems Design and Applications ISDA 2006: Sixth International Conference on Intelligent Systems Design and Applications. Jinan, 2006. C. 850-855.
6. Electronic resource: Development of models of social phenomena (family relationships) using data mining techniques http://www.chuvsu.ru/2008/proekt.html (in Russian)
7. Electronic resource: Application of data mining techniques to improve management system of higher education http://mfi.chuvsu.ru/opros/ (in Russian)

# DIGITAL TECHNOLOGIES TO ENHANCE SOLAR ENERGY EFFICIENCY AND RELIABILITY

*A. Bobyl, E. Terukov*

Ioffe Institute

bobyl@theory.ioffe.ru

The economic efficiency of solar energy depends on the efficiency of solar cells and modules, attained at the stage of their development; the cost-effectiveness of solar energy depends on the efficiency provided in their manufacture; and, finally, it depends on the reliability and service life of the whole system in the stage of its operation. The cost of control, accumulation, and connection units for standalone and network-connected power plants is of importance as well. The amount of electricity generated during the life cycle of a power plant is an integral parameter. When the total cost, including relatively low operating costs of the power plant, is estimated, the additional benefits of solar energy, namely environmental and social ones (population's comfort), are to be taken into account. These benefits are to varied extent relevant to the RF territory covered and uncovered with power supply lines.
To analyze large amounts of data and the data acquisition rate at a large number of variables (so-called 3V Big Data parameters) with consideration for specific and mutually dependent requirements of these 3 stages, it is necessary to employ digital technologies. These, in fact, provide a search and ranking of relationships between the parameters in the degree of their correlation. For example, the correlations between the parameters of the technology of solar cells and their efficiency will be of use for improving their efficiency, and the correlation between the power capacity and meteorological data, for forecasting the power generation, which is important for, in particular, reducing the failure rate of power supply lines. Thus, the development goals of the digital technology are to increase the effectiveness and reliability of DPs (solar cells, modules, and power plants) and methods for their long-term tests.

# DEVELOPMENT OF COMPUTATIONAL MODELS OF SOCIAL-ECONOMIC PHENOMENA BY MEANS OF DATA MINING

*V. Abrukov*

Chuvash State University

abrukov@yandex.ru

The social systems belong to the class of very complex systems. They include various subsystems and elements, are characterized by a lot of parameters. Their development depends on interaction of the various internal and external factors. Therefore a creation of their models is accompanied always by large difficulties. From this point of view, the problem of development of methods of simulation of social systems on the basis of modern methods of the analysis of the data is very actual. A family is example of a complex social system. A formation of family and a disintegration of family (divorce) are one of the most spread social phenomena. On the well known data of sociological interrogations, a family is considering as the most significant orb of life by both young and elder, both rich and poor people. Therefore research problems of the family relations are very important. In particular, the tasks of determination of conditions of formation a long-time ("happy") family, diagnostics of the existing family relations, determination of reasons of family crisis origin, development of ways of preventing of family crisis are very important. But now there are no scientifically justified quantitative criteria of determination of perspectives of the family future and diagnostics of an existing family, there are no multifactor quantitative models of the family relations. The complexity of the family relations, in which are interlaced psycho physiological, social, economic, etc forces, is a main reason. From this point of view, methods of Data Mining (DM) could be considered as perspective methods of simulation, because they allow simultaneously analyzing of both quantitative and quality data, allow gaining of multifactor computational models [1-3]. The methodology and technologies of application of DM tools at the analysis of social phenomena on an example of the analysis of family relations in divorced families and existing families are designed in this work. The DM tools included in the analytical platform Deductor [4] were used. The structure of the database is designed. Multifactor computational models of a marriage that capable to approximate influence of the complex of the internal and external factors on duration of a marriage are constructed. The models of prediction of duration of a marriage and degree of satisfaction of marriage for various cases are developed. We think that the work can be considered as a start of a "big" work of many the "data mining investigators" in this direction that can be considered as a real-world problem. The outcomes obtained in this work depict, that DM can be considered as perspective methods at problem solving and simulation for other social phenomena, in particular, at the analysis of such problems, as search of job and selection of staff (warning of the fast "divorce" of firm and worker), management of education systems, terrorism modeling and prevention (see, please, http://www.chuvsu.ru/2008/proekt.html (in Russian), http://www.chuvsu.ru/2008/proekt_eng.html (in English), http://mfi.chuvsu.ru/opros (in Russian).

Schedule of future work

1. A collection and processing of new sociological data deal with family relations by means of new interviews (that will be obtained by means of Web-technologies).
2. Development of measures of rendering assistance to young families, measures of social protection of institute of family as a whole.
3. Development of the methodical guides for the sociologists that will help them during a carrying out of an analysis of the data by means of DM.
4. Development of ready platforms for the data analysis in the field of family relations.
5. Investigation of DM possibilities in a work of HR services and agencies.

6. Use of Data Mining for Improved Management of Education System.
7. Terrorism Modeling and Preventing by means of Data Mining.

References

1. Abrukov V.S., Ya.G. Nikolaeva. Quantitative and qualitative methods: combine and run! SOTZIS (Social Research), Moscow, 2010, N1, p. 142-145.)
2. Abrukov V.S. A happy marriage: Analysis and management of family relationships using artificial neural networks, the Community Managers Executive www.e-xecutive.ru, Moscow, 2010, 1 - 23. Direct address of the article: http://www.e-xecutive.ru/community/articles/1437975/
3. Abrukov VS Development of models of family relations with the help of artificial neural networks. Collection of popular scientific articles and photographic materials - the winners of the 2016 RFBR competition. Natural science research methods in the humanities. 2016, Issue 19, 227-241: http://www.rfbr.ru/rffi/ru/libsearch/o_1959577
4. BaseGroup Lab. Available: www.basegroup.ru

# BIG DATA IN SEPARATION PROCESS SIMULATIONS FOR MINERAL RAW MATERIALS

*L. Vaisberg, I. Ustinov*

Mekhanobr-Tekhnika Research and Engineering Corporation

ustinov_id@npk-mt.spb.ru

Mining processing plants are the most high-performing industrial enterprises in the world economy system. They process hundreds of thousands to several tens of millions of tons of raw materials annually, which consumes between 12% and 16% of the total electricity generated in the world. This scale of resource consumption requires optimization of the mining industry, its mining and processing segments, with the obvious need for adequate simulations at different stages of the industrial process. In the course of rock mass processing, initial lumps of material undergo several crushing stages with their size being reduced from $1.5 \cdot 10^3$ mm to 5-20 mm, with a simultaneous grain-size classification. Raw materials are generally further crushed to a medium size of $2 \cdot 10^{-1}$ to $5 \cdot 10^{-2}$ mm and subjected to separation into target mineral components on the basis of contrast characteristics, including by their magnetic, electrical, hydrophobic and other properties. The ground product can be separated both in water (where the water acts as a dispersing process medium) and in the air, for example, in a state of vibrational fluidization. The above several grain size reduction operations are generally jointly referred to as disintegration, and the operations for classification by grain size and physical properties are referred to as separation. Separation processes represent the most difficult task in terms of modeling. Even at an average-scale mining enterprise, the grain-size separation equipment continuously processes $10^5$ to $10^6$ dissimilar mineral particles. In separation by physical properties, their number increases by at least 10 to 50 times. Mathematical modeling of such a high flow of discrete particles is carried out using two different approaches: the discrete-particle and the phenomenological methods. As an example, consider the processes of analyzing and processing big data with respect to modeling grain-size separation technologies for bulk material. The area of application of the discrete-particle method is the study and evaluation of the elementary separation of particles of bulk material on a vibrating screen surface. In this case, probabilistic and geometric parameters of the process are used, and the physical processes of collision, friction and reflection of particles are taken into account. Calculations are based on the fundamental measured physical properties of mineral particles, which can be obtained by the methods of process (engineering) mineralogy. Four to six such properties may be considered in a simulation for particles of a different nature. The mathematical methods used for the calculations are represented by a variety of discrete element method types in combination with the particle swarm method. The use of this combination enables a reduction in computation times on an advanced PC, including the time required to compile the respective protocol, from several tens of hours to 20-30 minutes as compared to traditional calculations using the discrete element method. That is, this method is most similar to on-line operation with the receipt of important information for engineering calculations. Further development of big data processing methods for grain-size separation processes is associated with the use of phenomenological approaches. These approaches are particularly efficient when applied to the processes of grain-size classification of a fine polydisperse bulk material with an average particle size of $10^{-1}$ to $10^{-2}$ mm, used, for example, in additive manufacturing. Phenomenological approaches to such systems are based on the representation of bulk material in the form of a granular gas, the enlarged energy characteristic of which is the concept of the special granular temperature. The final level in the hierarchy of methods for the mathematical modeling of large arrays are the methods developed by us for the complete thermodynamic description of the processes of separation and mixing of granular systems, enabling identification of the thermal and entropic components of such processes. The work was carried out under the grant from the Russian Science Foundation (project No. 17-79-30056).

# SANDIMS: A NEW SOUTH AFRICAN DATA PORTAL FOR POLAR AND HIGH LATITUDE POLAR RESEARCH DATA FROM THE SOUTH AFRICAN ANTARCTIC BASE, MARION ISLAND AND GOUGH ISLAND.

*K. Niemantinga, P. Cilliers, G. Lamprecht*

South African National Space Agency

kniemantinga@sansa.org.za

SANDIMS: A new South African Data Portal for polar and high latitude polar research data from the South African Antarctic base, Marion Island and Gough Island. The archiving and dissemination of geophysical research data collected at the South African Antarctic research base (SANAE IV) and at the South African high latitude observatories on Marion Island and Gough Island, which fall under the South African National Antarctic Programme (SANAP) has until recently been fragmented and inaccessible to international researchers. The data collected from these remote locations during the past decade, is a vast resource for addressing global challenges and Data-Driven Science in the areas of Heliophysics, Geophysics and Science of the Earth's upper atmosphere and ionosphere.

The Space Science Directorate of the South African National Space Agency (SANSA) in Hermanus is in the process of implementing a scientific Data Portal called the South African National Geophysical Data and Information Management System (SANDIMS) which will for the first time make the geophysical data collected and used by SANSA available through a single Data Portal. The SANDIMS Data Portal is complemented with a Metadata Manager Application for easier creation, editing, collation and dissemination of metadata. The Metadata complies with the international DIF Standard used by the GCMD. It will therefore be easier to submit data to and get data from other data repositories of interest. SANDIMS will allow our Space Agency to control and manage our data for research and application purposes and allow us to offer the data to our international collaborators and commercial partners.

Our datasets comprise various space science instrumentation sets from Ionospheric Scintillation receivers, VLF receivers, Ionosondes, The SuperDARN network and from magnetic observatories dating back to 1841. Standardisation of the datasets has been key to ensuring ease of access, display and dissemination of data.

The process of developing the Data Portal has brought about key questions of data stewardship, data quality, validation of data by Principal Investigators, data lineage and provenance as well as life cycle planning and funding which has resulted in the need for a formal Data Management Strategy. SANSA is in the process of composing the Data Management policy. The system will meet national and international obligations and expectations, as well as raise the standard of South African Geophysical research. The system's unique database will contain high-quality data related to events in space that, potentially, could supply information for unanswered scientific questions and enhance scientific development. The paper will present the design philosophy and various aspects of the implementation of the SANDIMS Data Portal and the SANSA Data Management Policy.

# RECENT ACHIEVEMENTS IN GEOMAGNETIC DATA ANALYSIS FOR ADVANCED MONITORING OF THE EARTH'S MAGNETIC FIELD

*A. Soloviev*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

a.soloviev@gcras.ru

The continuous growth of geophysical observations requires adequate methods for their processing and analysis. On the other hand, geomagnetic studies require accurate and reliable measurements of the Earth's magnetic field (EMF) carried out from ground and space. The system analysis methods and data mining techniques are able to sustain the solution of these problems. This paper presents an innovative holistic hardware/software system developed for efficient management and intellectual analysis of geomagnetic data, registered by Russian geomagnetic observatories. The designed system provides sophisticated automatic detection and multi-criteria classification of extreme geomagnetic conditions, which may be hazardous for technological infrastructure and economic activity in Russia. The geomagnetic activity indicators include measure of anomalousness (MA), rate of change and amplitude of the geomagnetic field and near real-time K-index. An essential feature of the developed system is its ability to deal continuously with real time data streams. It enables the online access to digital geomagnetic data and its processing results along with its visualization on conventional and spherical screens. In addition to recognition of increased magnetic activity, the fuzzy-logic based MA indicator is dually used for defining magnetically quiet periods. This feature makes the MA applicable for selecting quiet days and consequently more precise and timely determination of Sq variations at specified observatories, as compared to classical IAGA approach. This method also plays important role in selection of observatory data for constructing proper geomagnetic field models of internal origin. With this respect, we give a description of the principally new approach to calculating magnetometer calibration values (baseline estimations), which is a crucial operation in producing continuous high quality measurements of the complete field vector at geomagnetic observatories. As opposed to the baseline estimation method, traditionally accepted in INTERMAGNET, this approach involves all information on the geomagnetic field variations (both, vector and scalar), available in between infrequent absolute measurements taken by observatory operators. As a result, it provides information on the observatory data errors, which is especially important in robust modeling of rapid core magnetic field variations. Secular variations (SV) of the EMF on short time intervals are detected by calculating the second time derivative (secular acceleration, SA) of the field. However, charting the SV and SA at the core-mantle boundary that might reveal sudden changes of SA polarity happening within a year or so is only possible for models derived from recent, high accuracy satellite observations. We propose a new way to SV modelling and detection of SA pulses that are direct manifestation of the dynamic processes in the liquid core, using observatory data solely. It enables retrospective historical data processing for recognition of SA pulses and geomagnetic jerks before 2000, when full-scale geomagnetic satellite observations started.

# GIS-ORIENTED DATABASE FOR SEISMIC HAZARD ASSESSMENT FOR SEISMICALLY ACTIVE REGIONS OF RUSSIA

*A. Soloviev, B. Nikolov*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

a.soloviev@gcras.ru

The increased seismic hazard zones amounts to about 20% of the overall territory of Russia. Among those 5% are extremely hazardous including Caucasus and Crimea, inhabited by about 15 million people. We've created and continue to develop a database with associated full featured user's interface for seismic hazard assessment that contributes to minimization of consequences of possible earthquakes in these regions. For the first time in a unified environment, the most complete information is accumulated with regard to both input data and their processing results, which enables multi-criteria assessment of seismic hazard. Modern GIS systems significantly facilitate preparation, management and analysis of such data as their majority has spatial reference. Extended tools for data analysis and interactive requests integrated into GIS environment provide means for individual assessment of the risk degree in the given regions according to various criteria. The database and its user's interface are implemented in ArcGIS environment and fully meet the scalability requirement in terms of both functionality and volume. When necessary, the volume and variety of the stored information increases, and the geographic coverage is expanded.

# GEOMECHANICAL MODEL OF THE PREPARATION AND OCCURRENCE OF THE GREAT EARTHQUAKES IN SUBDUCTION ZONES FOR THE EXAMPLE OF THE CATASTROPHIC SEISMIC EVENT TOHOKU 2011

*I. Garagash*[1], *L. Lobkovsky*[2]

[1]Institute of Earth Physics,
[2]P.P. Shirshov Institute of Oceanology

garagash@mail.ru

Narrow seismic belts are connected with conditions of contact on the borders of lithosphere plates. Interaction of plates in the contact zones are responsible for seismic process in island arcs and active borders of continents. The strongest earthquakes occur in the subduction zones in the vicinity of gently sloping of contact plane in the result of rupture. In this case, tsunamis caused by rapid movements of the seabed over the focal zone of the earthquake are the great danger. For determine of the coseismic displacements in the subduction zone the static solution about dislocation in elastic half-space is often used. The limitations of this approach are obvious. They are related to the fact that the real structure of the earth's crust and lithosphere, nonlinear properties of the medium and also the initial stress-strain distribution in the zone of earthquake preparation are not taken into account. Besides the static solution does not allow to study the dynamic process of formation of sea bottom displacements. Therefore for determine the seismic movement along the plane of contact of lithospheric plates it is proposed to use the model of elastoplastic medium with the Coulomb-Mohr yield condition. For modeling of plates contact use the interface with dry friction. In the basis of the subducted plate the distribution of the velocities causing the slow accumulation of deformations and elastic stresses in system is set. Earthquake occurs, when stresses in local zone of contact surface exceed the strength and movement on it begin to accelerate. As the dynamic friction less the static friction, the friction forces sharply fall and the earthquake occur. It is established that this process depend on the time of preparation and the level of the initial stresses which was achieved before the seismic motion and can proceed completely variously. The analysis of dynamic displacements of the sea-bottom shows, that the dynamic component of vertical displacement can exceed the displacement of the bottom which are established after earthquake almost twice. In March 2011 on the eastern coast of Honshu Island in Japan, there was the strong earthquake with magnitude 9.1, which caused the giant tsunami that had devastating impact on the Tohoku region. To calculate the processes in the seismic source the geomechanical model of subduction zone was developed. For this the structural lithostratigraphic model and seismic velocity distribution were used. On the basis of these data the distributions of density and mechanical properties were determined. Before the earthquake, the Earth's crust and lithospheric plate are in the stressed state caused by the forces of gravity and horizontal tectonic motions. The computation demonstrates that the accumulation of stress and strain during the plate dipping occurs heterogeneously and is accompanied by the formation of zones of inelastic strain localization. The earthquake is realized in two zones on the plates boundary with different patterns of motion along the fault localized at different depths. At first the friction coefficient in the first deep region is set and then in the second region closer to the surface. The process of rupture in the earthquake sources depends on the inhomogeneity of the stress state. It is established that for earthquake located closer to the surface the displacement is much larger than for a deeper zone of the earthquake source. It is demonstrated that as result of the earthquake the new system of zones of localized deformations is being formed. The work was supported by the Russian Sci. Foundation Project No. 14-50-00095.

# AUGMENTED POST SYSTEMS: STRING-OPERATING KNOWLEDGE REPRESENTATION FOR BIG DATA AND INTERNET OF THINGS APPLICATIONS

*I. Sheremet*

Russian Foundation for Basic Research

sheremet@rfbr.ru

Augmented Post systems (APS) were introduced in [1, 2] as an efficient and flexible tool for implementation and maintenance of network-centric systems, operating in highly volatile environment. Due to representation of database (DB) as updated finite set of facts, being strings with metadatabase-defined structure, both SQL-like and NoSQL (WWW-, twitter- and other hypertext-like) query and data manipulation languages are simply implemented in this framework. APS knowledge representation is generalization of this data model in a Prolog-like manner so that knowledge base (KB) is set of so called S-productions, establishing content intersection between various facts, and P-productions, providing procedural connection of attached "rigid"(non-modified) programs, such as various sensors and actuators drivers, big data storage controllers etc. APS logical inference axiomatics and its algorithmical implementation provide integration of deductive databases paradigm with distributed data flow online analytic processing. Due to the unified string representation of input, output and stored facts, APS provide simple description of processing logic of heterogeneous data flows, which sources are various Internet-connected devices, as well as logic of these devices control in hard real time. Processing is implemented by local soft/hardware inference engines, applying local KB/DB to input data flows through local blackboards, accepting facts entering from remote devices of distributed network. In turn, facts generated by mentioned inference engines are transported by network services to the mentioned remote devices. Generalized model of associative storage of APS KB based on the original model of information incompleteness provide fast access to facts as well as minimization of redundant steps of logical inference. Large volumes of data are stored in a minimally possible space due to original data compression techniques called grammatical coding [3]. All mentioned features may provide flexible APS-based implementation of various Big Data, Internet of Things, Cyberphysical Industry as well as most complicated Cybersecurity solutions, which are already installed at more than 5,000 facilities in Russia and abroad.

References

1. Sheremet I.A. Augmented Post Systems: The Mathematical Framework for Knowledge and Data Engineering in Network-Centric Environment. // Berlin: EANS. 2013. P. 215.
2. Sheremet I. Data and Knowledge Bases with Incomplete Information in a "Set of Strings" Framework. // International Journal of Engineering and Applied Sciences, Vol.3 (2016), No. 8, pp. 90-103.
3. Sheremet I.A. Grammatical Codings.// Hannover: EANS. 2012. P. 54.

# TECTONOSTRATIGRAPHIC ATLAS OF THE ARCTIC (EASTERN RUSSIA AND ADJACENT AREAS)

*O. Petrov, N. Sobolev, S. Kashubin, E. Petrov, D. Leontiev, T. Tolmacheva*

Russian Geological Research Institute (VSEGEI)

vsgdir@vsegei.ru

Intensive study of the Arctic region during last 15 years under a series of major international and national projects has resulted in the accumulation of enormous amount of new data on the geological structure of the region, including those obtained during geophysical sounding of the ocean floor, the examination of bottom rock material, studying geology of islands and the continental part of the Russian Arctic. One of the most successful projects is the international cooperation of eight circumpolar countries on the creation of the Atlas of Geological Maps of the Circumpolar Arctic at a scale of 1:5,000,000, which already includes geological and tectonic maps, potential fields map, mineral resources map. More detailed geological information on the Arctic region can be generalized in the form of atlases of major geological structures and regions of polar areas. The atlas "Geological History of the Barents Sea" published in 2009 by Russia and Norway was first successful experience in the implementation of this work.

Modern level of knowledge of the Arctic allows the analysis of tectonostratigraphy and litho-geodynamics of the region and the interpretation of geological complexes in terms of tectonic settings of stratigraphic sequences formation based on the sedimentary basin and the entire lithosphere. This work has already begun; in 2014, geological surveys of Denmark, UK, Norway, Iceland, and the Netherlands complied the "Tectonostratigraphic Atlas of the North-East Atlantic region".

Currently, the Russian Geological Research Institute (VSEGEI) actively works on the creation of paleogeographic reconstructions and litho-geodynamic models showing the formation of the sedimentary cover of the continental shelf of Eurasia and adjacent Central Arctic elevations and other geological structures of the Arctic Ocean. Compilation of the "Tectonostratigraphic Atlas of the Arctic (Eastern Russia and adjacent areas)", covering the eastern Russian Arctic continental part and eastern Eurasian basin has been started. The Atlas summarizes the accumulated to date geological material on this key region of the Arctic resulted from geophysical sounding of the ocean floor in recent years, examination of bottom rock material, collected by several national and international expeditions, studying the geological structure of islands and the continental part the Russian Arctic carried out, including, during several international expeditions since 2006. At present, the Atlas includes geological, gravimetric maps, anomalous magnetic field map, seismic knowledge sketch-map, and the set of composite seismic profiles crossing main geological structures of the northeastern Arctic, a set of structural maps (after the acoustic basement, after the top of Cretaceous and Eocene sediments), structural-geological map, map of tectonic zoning of the basement and sedimentary cover, and the set of paleogeographic maps (for the Early Triassic (-250 Ma), Late Jurassic (-145 Ma), Early Cretaceous (Aptian-Albian) (-112 Ma), Paleogene (Eocene ) (-35 Ma), and Neogene (Miocene) (-10 Ma). Paleogeographic maps are accompanied by stratigraphic columns and photographs of cross-sections.

Creation of the Atlas as a set of systematic modern geological information is of great importance for understanding the geological structure and history of tectonic development, formation time and the correlation of sedimentary and volcanic complexes in the region. These studies allow understanding the peculiarities of the formation of the Arctic Basin, which is transformed and deformed margins of the Paleo-Asian Ocean based on the new geological-geophysical and lithogeodynamic basis.

# DRILLING PERSPECTIVES IN LOW CARBON ENERGY SCENARIOS

*N. Nakicenovic, S. Busch*

International Institute for Applied Systems Analysis

naki@iiasa.ac.at

## Energy for sustainable development

Energy is a fundamental human need. Increases in global demand for energy services and decarbonization will require a fundamental transformation of the energy systems. Energy sources like renewables can provide both, carbon-free energy and efficient provision of needed service. Fossil energy like natural gas can achieve the same in conjunction with carbon capture and storage. More than a dozen carbon capture and storage facilities exist around the world, but they are still in the pre-commercial development phase and can remove on the order of million rather than billion tons of carbon (Benson et al. 2012; Metz et al. 2005; Pachauri et al. 2014).

## The role of natural gas

Natural gas is well-suited for carbon capture and storage as it contains half of carbon per unit energy compared to coal. Gas hydrates are the most abundant form of hydrocarbons and with estimated up to 800 ZJ (Zettajoules or 800,000 EJ) they represent an essentially unlimited potential source of natural gas which exceeds the current global energy demand per year of 500 EJ by a factor of 1,600. (Rogner et al. 2012; Krey et al. 2009). However, extraction technologies would yet need to be developed. Shale gas sources with estimated global resource base of over 15 ZJ are already commercial due to recent improvements of steering techniques allowing long stretches of horizontal drilling with fracture stimulation (fracking) to keep the gas and oil flowing. According to Cook (2014), oil and gas development wells will exceed 100,000 per year by 2020, of which about 2,500 will be offshore development drills. In comparison, about 80,000 development wells were drilled in 2013 and in addition about 10,000 exploratory well drills were made. Taking the expected 2020 number of 100,000 per year till 2100 translates into some nine million oil and gas wells to be drilled cumulatively.

## Climate change mitigation

Rapid decarbonization of the world as foreseen in the historical Paris Climate Agreement (UNFCCC, 2015) is needed to achieve stabilization of global mean temperature below 2°C (above pre-industrial levels) and if possible down to 1.5°C. This is a very stringent requirement and implies reduction of global emissions to zero by about mid-century meaning each decade emissions need to be halved (Rockström et al. 2017; Pachauri et al. 2014; Rockström et al. 2017). In contrast, current global emissions are about 40 billion tons (40 Gt $CO_2$) of carbon (Le Quéré et al. 2015).

Many scenarios in the literature can stay within this stringent limit provided immediate and vigorous reduction of emissions is achieved. Any delay in emission reductions however increases the likelihood that carbon needs to be removed from the atmosphere. Carbon storage needs may require up to 10,000 wells per year assuming cumulative storage of up to 2,500 Gt $CO_2$ during this century. Most of the carbon to be stored would be captured from fossil energy sources and smaller amount from the atmosphere, for example, by capturing carbon from sustainable biomass. Rapid diffusion of renewables and nuclear (where it is accepted and considered safe enough) in conjunction with vigorous efficiency improvements could reduce the storage requirements to almost zero. On average capacity of each well would be about one million tons of $CO_2$ (1 Mt $CO_2$) per year (Blunt 2010) and lifetime more than 40 years (Smith et al. 2010). To store up to 2,500 Gt$CO_2$ this would translate to up to more than 30,000 operational wells in each year between 2030 and 2050. In other words, during the first decades of deployment some 30,000 wells would need to be drilled and during the second half of the century at least that many.

References

Benson, Sally M., Kamel Bennaceur, Peter Cook, John Davison, Heleen de Coninck, Karim Farhat, Andrea Ramirez, et al. 2012. “Chapter 13 - Carbon Capture and Storage.” In Global Energy Assessment - Toward a Sustainable Future, 993–1068. Cambridge University Press, Cambridge, UK and New York, NY, USA and the International Institute for Applied Systems Analysis, Laxenburg, Austria. www.globalenergyassessment.org.

Fuss, Sabine, Josep G. Canadell, Glen P. Peters, Massimo Tavoni, Robbie M. Andrew, Philippe Ciais, Robert B. Jackson, Chris D. Jones, Florian Kraxner, and Nebosja Nakicenovic. 2014. “Betting on Negative Emissions.” Nature Climate Change 4 (10): 850–53.

Johansson, Thomas B., Anand Patwardhan, Nebojša Nakićenović, Luis Gomez-Echeverri, and International Institute for Applied Systems Analysis, eds. 2012. Global Energy Assessment (GEA). Cambridge : Laxenburg, Austria: Cambridge University Press ; International Institute for Applied Systems Analysis.

Le Quéré, C., R. Moriarty, R. M. Andrew, J. G. Canadell, S. Sitch, J. I. Korsbakken, P. Friedlingstein, et al. 2015. “Global Carbon Budget 2015.” Earth System Science Data 7 (2): 349–96. doi:10.5194/essd-7-349-2015.

Metz, Bert, Ogunlade Davidson, Heleen De Coninck, Manuela Loos, and Leo Meyer. 2005. “IPCC Special Report on Carbon Dioxide Capture and Storage.” Intergovernmental Panel on Climate Change, Geneva (Switzerland). Working Group III.

Pachauri, Rajendra K., Myles R. Allen, Vicente R. Barros, John Broome, Wolfgang Cramer, Renate Christ, John A. Church, Leon Clarke, Qin Dahe, and Purnamita Dasgupta. 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC.

Riahi, K., F. Dentener, D. Gielen, A. Grubler, J. Jewell, Z. Klimont, V. Krey, D. McCollum, S. Pachauri, S. Rao, B. van Ruijven, D. P. van Vuuren and C. Wilson, 2012: Chapter 17 - Energy Pathways for Sustainable Development. In Global Energy Assessment - Toward a Sustainable Future, Cambridge University Press, Cambridge, UK and New York, NY, USA and the International Institute for Applied Systems Analysis, Laxenburg, Austria, pp. 1203-1306.

Rogner, Hans-Holger, Roberto F. Aguilera, Christina Archer, Ruggero Bertani, S. C. Bhattacharya, Maurice B. Dusseault, Luc Gagnon, et al. 2012. “Chapter 7 - Energy Resources and Potentials.” In Global Energy Assessment - Toward a Sustainable Future, 423–512. Cambridge University Press, Cambridge, UK and New York, NY, USA and the International Institute for Applied Systems Analysis, Laxenburg, Austria. www.globalenergyassessment.org.

Rockström, Johan, Owen Gaffney, Joeri Rogelj, Malte Meinshausen, Nebojsa Nakicenovic, and Hans Joachim Schellnhuber. 2017. “A Roadmap for Rapid Decarbonization.” Science 355 (6331): 1269–1271.

Smith, L., Billingham, M., Lee, C. H., Milanovic, D. Z., & Lunt, G. (2010, January 1). CO2 Sequestration Wells - the Lifetime Integrity Challenge. Society of Petroleum Engineers. doi:10.2118/136160-MS

UN GA (2015). Transforming our world: the 2030 Agenda for Sustainable Development A/RES/70/1 UN. New York, United Nations General Assembly (UN GA).

UNFCCC (2015). Adoption of the Paris Agreement. Paris, United Nations Framework Convention on Climate Change (UNFCCC). FCCC/CP/2015/L.9/Rev.1.

# DATA DRIVEN MODELLING IN GEODYNAMICS: METHODS, APPLICATIONS AND CHALLENGES

*A. Ismail-Zadeh*[1], *I. Tsepelev*[2], *A. Korotkii*[3]

[1]Karlsruhe Institute of Technology
[2]Russian Academy of Sciences

alik.ismail-zadeh@kit.edu

Geodynamics, whose past and current behaviours are of great scientific interest, deals with dynamic processes in the Earth's interior and on its surface. Thermal convection in the mantle, lithosphere dynamics and subduction as well as their surface manifestation, such as volcanism, seismicity, and sedimentary basins evolution, are among principal geodynamic problems. With great advances in understanding the geodynamic processes based on geophysical and geodetic monitoring, observations and data analysis, and with a technological progress in computer simulations, data-driven numerical models in geodynamics become feasible and important in recovering the past, analyzing the present, and forecasting the future. These models, derived from the mathematical models describing dynamical processes in the Earth and on its surface, are considered in the cases of known effects of the processes – available geophysical, geological, geochemical, geodetic, and other observations/data – but (some) unknown causal characteristics of the models. The goal of data-driven numerical modelling is to determine the model characteristics using data sets related to Earth's deep, shallow and surface observations. A challenging issue in the modelling is data availability and their quality: if geophysical data are incomplete and somehow biased, the data can be improperly assimilated. Novel methodologies, such as adjoint and quasi-reversibility techniques, will be presented, discussed, and illustrated using a few model examples of mantle-lithosphere dynamics, salt diapirism, and lava flow (Ismail-Zadeh et al., 2016). These applications might present an interest to hydrocarbon exploration industry and to disaster risk reduction management. Ref: Ismail-Zadeh, A., Korotkii, A., and Tsepelev, I., 2016. Data-Driver Numerical Modelling in Geodynamics: Methods and Applications. Springer-Nature, Switzerland. http://www.springer.com/gp/book/9783319278001

# SEISMIC HAZARD MODELING USING INSTRUMENTALLY RECORDED, HISTORICAL AND SIMULATED DATA

A. Ismail-Zadeh[1], V. Sokolov[2]

[1]Karlsruhe Institute of Technology,
[2]Saudi Geological Survey, Jeddah

alik.ismail-zadeh@kit.edu

The probabilistic seismic hazard analysis (PSHA) based only on instrumentally recorded seismic observations and a few historical data, has a disadvantage because these observations cover a much shorter time interval compared to the duration of the tectonic processes responsible for earthquake generation. Numerical modeling of realistic seismogenic processes allows generating catalogues of synthetic earthquakes (simulated data) covering relatively long time intervals and, therefore, providing a basis for estimates of the parameters of the earthquake occurrences and the ground shaking. We present an approach to assessment of regional seismic hazard, which accounts for observed (instrumentally recorded and historic) earthquakes, as well as for seismic events simulated for a significantly longer period of time than that of observations. We apply this approach to PSHA for the Tibet-Himalayan region. The large magnitude synthetic events, which are consistent with the geophysical and geodetic data, together with the observed earthquakes are employed for the Monte-Carlo PSHA. Earthquake scenarios for hazard assessment are generated stochastically to sample the magnitude and spatial distribution of seismicity, as well as the distribution of ground motion for each seismic event. The obtained peak ground acceleration values, which are expected to be exceeded at least once in 50 years with probability 0.1, show that the hazard level associated with large events in the Tibet-Himalayan region significantly increases if the long record of simulated seismicity is considered in the PSHA. The magnitude and the source location of the 2008 Wenchuan M=7.9 earthquake are among the range of those described by the seismic source model accepted in our analysis. We analyze the relationship between the ground motion data obtained in the earthquake's epicentral area and the obtained PSHA estimations using a deaggregation technique. The proposed approach provides a better understanding of ground shaking due to possible large-magnitude events and could be useful for risk assessment, earthquake engineering purposes, and emergency planning.

# CODATA SPECIAL ISSUE ON HISTORICAL DATASETS FOR DISASTER MITIGATION AND RESEARCH

*G. Li*

The Institute of Remote Sensing and Digital Earth

ligq@radi.ac.cn

Data has been regarded as one of the most important resource for disaster reduction, such as 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. If natural disaster is understood as the special phenomena for health of earth system, , the multidisciplinary data to record the event of certain disaster is the historical documents for future research, as the medical examination data for illness of human beings. However, most of such historical datasets for certain natural disaster events have been lost in the past decades, even these events impacted and heat all heart around the world for their damage. In recent years, there was a surge in the volume of the historical disaster collection, preservation and data sharing. More and more such datasets can be discoverer and accessible. But few of disaster datasets are published to make them trust, copyright-clear, and cited. Supported by CODATA Task Group of Linked Open Data for Global Disaster Risk Research (LOD), the Special Issue on Historical Datasets for Disaster Mitigation and Research in journal of China Scientific Data is an attempt to apply data publish in the field of disaster reduction. This report will give introduction of this issue and the guidance to submit the datasets and data papers.

# KNOWLEDGE BASE ABOUT PAST EARTHQUAKES CONSEQUENCES AS A TOOL TO INCREASE THE RELIABILITY OF NEAR REAL TIME LOSS ESTIMATION

*N. Frolova[1,2], V. Larionov[3], J. Bonnin[4], S. Sushchev[5], A. Ugarov[6], I. Gabsatarova[7], A. Kijko[8]*

[1]Russian Academy of Sciences and Extreme Situation Research Center,
[2]Seismological Center, IGE, Rus. Acad. Sci,
[3]Seismological Center, IGE, Rus. Acad. Sci.,
[4]Institut de Physique du Globe, University of Strasbourg,
[5]Extreme Situations Research Center,
[6]Extreme Situations Research Center,
[7]Geophysical Survey, Rus. Acad. Sci.,
[8]University of Pretoria Natural Hazard Centre

frolova@esrc.ru

Information on possible damage and expected number of casualties due to strong earthquakes is very critical for taking the proper decisions about search and rescue operations, as well as rendering humanitarian assistance. The experience of earthquakes disasters in different earthquake-prone countries shows that the officials who are in charge of emergency response at national and international levels are often lacking prompt and reliable information on the disaster scope.

At present, three global systems exist that allow to provide earthquake loss estimation just after the event. They are : the Russian "EXTREMUM" System which allows to simulate the distribution of seismic intensity, damage to buildings of different types, number of casualties in damaged and destroyed buildings as well as identify effective response measures in the case of emergency; the Global Disaster Alert and Coordination System (GDACS) developed by JRC, European Commission which allows in near-real time to monitor the seismic situation and provide estimation of expected number of inhabitants in the affected area by using the information on population density; and the "Prompt Assessment of Global Earthquakes for Response" (PAGER) System of the US Geological Survey which allows to simulate expected shaking intensity and estimate expected number of inhabitants in zones of different I by using the information on population density. The paper is analyzing the reliability of loss estimation with different global systems' application. It is proposed to use the knowledge base on physical and socio-economical consequences of past earthquakes, which may be used for calibration of near real-time loss assessment systems based on simulation models for shaking intensity, damage to buildings and casualty estimates. Such calibration allows to compensate some factors which influence on reliability of expected damage and loss assessment in "emergency" mode. The knowledge base contains the description of past earthquakes' consequences for the area under study. It also includes the current distribution of built environment and population at the time of event occurrence. Computer simulation of the recorded in knowledge base events allow to determine the sets of regional calibration coefficients, including rating of seismological surveys, peculiarities of shaking intensity attenuation and changes in building stock and population distribution, in order to provide minimum error of damaging earthquakes loss estimations in "emergency" mode.

The need for coordinated efforts and research at international level is mentioned if one wants to increase the reliability of loss estimation in "emergency" mode.

References

1. Frolova, N., Larionov, V., Bonnin, J.: Data Bases Used In Worldwide Systems For Earthquake Loss Estimation In Emergency Mode: Wenchuan Earthquake. In Proc. TIEMS2010 Conference, Beijing, China, 2010.

2. Frolova N. I., Larionov V. I., Bonnin J., Sushchev S. P., Ugarov A. N., Kozlov M. A. Loss Caused by Earthquakes: Rapid Estimates. Natural Hazards Journal of the International Society for the Prevention and Mitigation of Natural Hazards, vol.84, ISSN 0921-030, Nat Hazards DOI 10.1007/s11069-016-2653

# NEW ACHIEVEMENTS IN FUNDAMENTAL AND APPLIED RESEARCHES OF THE "COLD" EARTH DEGASSING

*A. Tatarinov, L. Yalovik*

Geological Institute, SB RAS (Russia)

tatarinov@gin.bscnet.ru

In the last 15-20 years, we can see a remarkable progress in the doctrine for "cold" (non magma) branch of the Earth degassing. Currently, the most pressing of its problems are considered from the standpoint of fluid dynamics and the plume concept. Due to the wide use of modern geological, geophysical, petrological, geochemical and other research methods, the deep drilling data - primarily oil and gas sedimentary basins, the knowledge about the "cold" degassing structures, its manifestation forms (pipe, cold seeps, hydrovolcanoes, "astroblemes", thermals) are greatly expanded. At the same time, some research areas of fundamental nature being critical importance for ecology, minerals ore deposits forecasting and prospecting were not yet reflected in the fluid-dynamic concept of the "cold" Earth degassing doctrine, due to insufficient knowledge of some phenomena and processes. The authors propose the following results, received during long-term studies of the known and identified by them the "cold" degassing structures of Mesozoic-Cenozoic age rift systems in Eastern Siberia.

1. The plasmoid form phenomenon (phenomenon) of the "cold" Earth degassing is proved to exists [1]. The previous researchers discussed this phenomenon at the level of hypotheses and assumptions. Various factors, plasmogeneration conditions in lithosphere and atmosphere were estimated. We created the morphostructural classification of registered plasmoids in the Baikal rift zone, and the peculiarities of their internal structure are revealed. The decisive role of water phase transitions in the lithospheric plasmoids in near the ground atmosphere occurrence and structuring is shown.
2. The structure features, mineral composition, Genesis, first discovered bio-inert travertine and stromatolite like small - and mini structures in the Baikal rift zone depressions are established [2]. They are shown to have been formed with active participation of the bacterial communities at the sites of the actual mud micro volcanoes unloading gas-water fluids.
3. The lithocomplexes forming main features, first identified by the authors, in a major South-Siberian region of Mesozoic-Cenozoic mud volcanism inland type are set. They identified mineral associations that characterize gas-explosive, oil-gas-water-lithoclastic and oil-gas-water phase of the mud volcanic activity, the mechanisms of mineral Genesis [3]. The mud volcanoes fluid-dynamic functioning modes, which differ in dominant genetic mechanisms of mineral formation are considered: root (local) fluidogeneration structures, fluidomineral substrate transit channels, hydrotherms transit channels. The main factor of fluids generation, initiating mud volcanism in the South Siberian area is dynamometamorphic mechanism.
4. The mud volcanic genetic type of Cenozoic age gold mineralization associated with the mud volcanic depression structures formation is predicted to be new for Lena area [4]. We prove the spatial-genetic link with it of large industrial gold placers.
5. It is established that Baleyskaya gold field is a subvertical degassing pipe laid down on the Precambrian, in varying degrees granitisated in the Paleozoic Onon-Shilkinskaya Greenstone belt [5].

The concentrating processes of precious metals in the Baleyskaya field ore and host rock complexes were determined by the fluid system C-O-N-N-S evolution from restored to highly oxidized compounds.

References

1. Tatarinov A.V., Yalovik L.I. Plasmoid Phenomenon of "Cold" Earth Degassing (The Baikal Rift Zone as Illustration) // International Journal of Science and Research. 2016. V. 5, N 2. P. 981-987.
2. Yalovik L., Tatarinov A., Danilova E., Doroshkevich S. Bio-inert Dome and Columnar Structures of Mud Microvolcanism in Baikal Rift Zone // International Journal of Advanced Research in Science, Engineering and Technology. 2016. V. 3, N 9. P. 2589-2600.
3. Tatarinov A.V., Yalovik L.I., Kanakin S.V. Formation particularities and mineral associations of mud volcanoes lithocomplexes at the south of East Siberia //Volcanology and Seismology. 2016, No. 4. P.34-49
4. Tatarinov A.V., Yalovik L.I. Placer-Forming Cenozoic mud-volcano genetic type of gold mineralization in the Lena area, Patom Highland, Russia // Global Journal of Earth Science and Enigineering. 2014. N 1. P. 24-33. 5. Tatarinov A.V., Yalovik L.I., Kolesov G.M., Kanakin S.V., Prokopchuk S.I. Platinum group elements in over ore thick of Baleyskaya gold field// RAS Reports. 2011. V.436. No. 1. P. 94-98

# THE AMUR RIVER AS A CENTURY-LONG PROGNOSTIC TESTING GROUND FOR DISASTER RISK RESEARCH

*G. Sokolova*

Institute of Water and Environmental Problems of the Far-Eastern Branch
of the Russian Academy of Sciences

pozhar@ivep.as.khb.ru

The database on the regime of the Amur River, the largest waterway in Northeast Asia, and statistical analysis for research into the risk of hydrological disasters (floods, low water levels) is the basis for resolving the issue of long-term forecasting of these risks. For the indicator of the highest wave of rainwater in flooded years with floods and in years with low water we adopted the maximum annual water level relative to the "0" chart of Khabarovsk post, where the total impact of the main tributaries of the Amur is manifested. For the purposes of the forecast, the following analysis results are obtained.

1. A significant correlation was established between annual peaks and average monthly water levels for September ($R^2 = 0.70$), which characterizes exclusively rain floods without the participation of thawed runoff. Using the method of smoothing out local fluctuations by a 5-year-old slip window, "september" wave-rhythms with a regular alternation of groups of low-water and high-water years are distinguished. So, already in 2008, after the catastrophic summer low water on the Amur River near Khabarovsk, a very high flood wave was expected in the coming years (according to the established rhythmicity) with the probability of reaching an unprecedented horizon for the entire observation period. Given the rhythm and anomalies in the deviation of annual maximums from the norm, in September 2011 it was assumed that in 2013 the water level at the peak of the flood in Khabarovsk could rise by about 2008 (in absolute terms). The error of the author's long-term forecast for two years was 11 cm.
2. It has been established that the average duration of the september rhythm waves between peaks (groups of deep-water years) and depressions (groups of low-water years) is about 20 years. In the dynamics of 20-year periods for 1896-2016 years, the number of years with flooding and shallowing of the Amur floodplain was recorded. In 2016 the author's forecast of the Amur regime was made before the end of the current 20 years (2010-2029): 4-5 years of low water and 1 year with flooding of the high floodplain, in the remaining years peaks of rain floods will reach marks between 300-500 cm Relative to the "0" chart of Khabarovsk post.

# PROBLEMS OF SCIENTIFIC ANALYSIS AND INTERPRETATION OF GEOLOGICAL AND GEOPHYSICAL DATA ON THE EXAMPLE OF THE ARCTIC CONTINENTAL SHELF

*V. A. Vernikovsky*[1,2], *A. E. Vernikovskaya*[1,2], *D. V. Metelkin*[1,2], *N. Yu. Matushkin*[1,2], *N. E. Mikhaltsov*[1,2], *A. I. Chernova*[1,2]

[1] A. A. Trofimuk Institute of Petroleum Geology and Geophysics of Siberian Branch of the Russian Academy of Sciences (IPGG SB RAS, Russia),
[2] Novosibirsk State University (NSU, Russia)

vernikovskyva@ipgg.sbras.ru

The geology of one or another segment of the Earth's territory is a complex system formed during a long time (hundreds of million years) due to reorganization of lithospheric plates under the influence of deep processes in the core and the mantle. The main goal of researchers in Earth sciences is to understand when these systems were formed, which is extremely important for solving current global problems. To solve this issue and constructing a geological and geophysical model of the lithospheric segment we are interested in, a complex of data is needed. Because the geological and geophysical information for large lithospheric segments is usually very heterogeneous, and is a medium where large arrays of data combine with singular, but very important data points, it must be deeply analyzed by researchers and correctly interpreted. Based on the example of the geological structure of the Russian Arctic continental margin, we consider the problems of scientific analysis and geological and geophysical interpretation of geostructural, geochronological, paleontological, paleomagnetic and other data for solving current global issues. One of these issues if the problem of outer boundaries of the continental shelf of states with arctic shorelines.

# MODELING THE BLOCK STRUCTURE DYNAMICS AND SEISMICITY IN THE CAUCASIAN REGION

*A. A. Soloviev*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia),
Institute of Earthquake Prediction Theory
and Mathematical Geophysics, RAS (IEPT RAS, Russia)

soloviev@mitp.ru

Based on the morphostructural zoning scheme of the Caucasus, the block structure reflecting the real fault geometry and the block formation of the region has been constructed. Several dozens of numerical experiments have been conducted for simulating the dynamics of the block structure and the arising seismicity. The total size of the data obtained in these experiments is more than 5 gigabytes The modeling relies on the following principles. It is assumed that the structure is composed of perfectly rigid blocks separated by infinitely thin fault planes. On the fault planes and on the blocks' bottoms, the blocks interact viscoelastically with each other and with the underlying medium. At each time instant, the translational displacements and rotations of the blocks are calculated from the condition of quasi static equilibrium of the entire block structure. The earthquakes occur in accordance with the dry friction model when within a certain segment of the fault the stress-to-pressure ratio exceeds a given threshold. The synthetic catalog of the earthquakes in the Caucasian region obtained as a result of the numerical modeling reproduces the key features of the real seismicity. The magnitude--frequency diagram for the synthetic earthquake catalog has virtually the same slope (b-value) as its counterpart for the real seismicity of the Caucasus, whereas the spatial distribution of the synthetic earthquakes reflects the characteristic features of the pattern of the real distribution. At the same time, the study revealed the presence of model seismicity in the areas where the real seismic events were absent but which were recognized by the pattern recognition methods as earthquake prone for the magnitude $M \geq 6.0$. The distinctions between the model and real seismicity can probably be associated, inter alia, with a more complicated geodynamics of the region than specified in the model. The real seismogenesis processes also involve the lower-order faults which are disregarded in the model. The motions of side boundaries of the blocks structure which are specified in the model perhaps need to be updated in order to more adequately reflect the motion of the lithospheric blocks in the considered region.
The study was supported by the Russian Science Foundation (Project # 15-17-30020).

# SEISMOGENIC NODES (M6+) RECOGNIZED IN THE ALTAI- SAYAN-BAIKAL LAKE REGION

*A. Soloviev, A. Gorshkov*

Institute of Earthquake Prediction Theory
and Mathematical Geophysics, RAS (IEPT RAS, Russia)

gorshkov@mitp.ru

The phenomenological approach based on pattern recognition is employed to identify seismogenic nodes capable of earthquakes with M6+ in the South Siberia Mountains including Altai, Sayan, and Baikal Lake area. The study region exhibits moderate and rare strong seismicity. The highest recorded magnitude in 2003 reaches 7.3. The methodology is based on the idea that earthquakes nucleate at nodes, specific structures forming around intersections of morphostructural lineaments. Morphostructural zoning (MZ) is used to identify nodes over the entire study region without using the a priori knowledge of regional seismicity. MZ delineates a hierarchical system of blocks with their boundaries, morphostructural lineaments. Each intersection of lineaments is treated as a node. The pattern recognition is employed to identify seismogenic nodes where earthquakes with $M^3$ 6.0 may occur in the future. Nodes are divided by pattern recognition algorithm into seismogenic (D) and non-seismogenic (N) on the base of geomorphic, geological, gravity, and magnetic parameters describing nodes.
As a result, seismogenic nodes capable of earthquakes with *M*6+ have been defined in the study region. We also found characteristic geological-geophysical features that discriminate seismogenic nodes from non-seismogenic ones.
The results obtained provide information for long-term seismic hazard assessment on the potential earthquake sources south part of Siberia from the Altai Mountains up to the Baikal Lake area. The recognition performed, pinpoints a number of D nodes where target events have not been recorded to date that could provide additional information to improve seismogenic source models. Some of the recognized D nodes give more knowledge about seismic risk affecting special sites like water power plants and dams as well as cities. The result could improve the performance of seismic hazard map of the studied region. The reported study was funded by the Russian Science Foundation (RSF Project 15-17-30020)

# POSSIBLE LOCATIONS OF M7+ EARTHQUAKES IN GREECE DEFINED BY PATTERN RECOGNITION

*1. Gorshkov*[1], *Y. Gaudemer*[2], *O. Novikova*[1]

[1] Institute of Earthquake Prediction Theory
and Mathematical Geophysics, RAS (IEPT RAS, Russia),
[2] IPGP (France)

gorshkov@mitp.ru

The study aimed to identify capable nodes for M7+ and their characteristic geological-geomorphic features in Greece where a number of large events is well documented in seismic history. Nodes were delineated by morphostructural zoning method based on the formalized analysis of the present-day topography and geological data. Nodes formed at the intersections of lineaments, are most likely locations for large earthquakes. The epicenters of M7+ earthquakes in the Hellenides correlate with delineated nodes. The other potential nodes were defined with the help of the pattern recognition technique on the basis of geological, geomorphic parameters of the nodes. Nodes experienced M7+ earthquakes were treated as the samples for training the recognition algorithm. As a result, we have properly recognized all nodes hosting M7+ earthquakes and a number of potential nodes where events of such size have not happen so far. The work contributes to the assessment of the seismic hazard in Greece. The reported study was partly funded by Russian Foundation of Basic Research (RFBR) according to the research project 16-55-12033.

# REALIZATION OF DIGITAL CORE TECHNOLOGY USING LASER-ULTRASONIC STRUCTUROSCOPY

*E. Cherepetskaya, I. Sas, D. Bagryantsev, N. Morozov*

National University of Science and Technology “MISiS” (NUST “MISiS”, Russia)

echerepetskaya@mail.ru

The digital core imaging and analysis technology is widely used in geological investigations and oilfield services. As a result, data acquisition accuracy and resolution is significantly higher now - owing to modern scientific and technological tools (in particular, computer tomography). However, the use of this technology is frequently very expensive. The National University of Science and Technology MISiS (NUST "MISIS") in cooperation with Terra Service LLC is developing a technology of structural 3D visualization of rocks based on laser ultrasonic structuroscopy and analysis of physical and mechanical properties. It is supposed that expensive tomography methods can be replaced by more cost-effective solutions. Digital core analysis makes it possible to reveal structural macro-heterogeneities in dispersive soils. Specifically, the heterogeneity of core samples from offshore wells can be characterized using this approach; solid and liquid phases can be discriminated, as well as frozen and thawed phases. It is possible to carry out examination of cores in capsules, so stratification of soils remains unchanged. The results of the study can be used in numerical simulation so as to estimate porosity, permeability, and rock fracturing and more accurately predict the behavior of foundations in terms of stability and deformation. This work was carried out with financial support from the Russian Science Foundation (grant no. 16-17-10181).

# SEISMIC HAZARD ASSESSMENT AND EARTHQUAKE PREDICTION IN A BIG DATA WORLD

*V. G. Kossobokov*

Institute of Earthquake Prediction Theory
and Mathematical Geophysics, RAS (IEPT RAS, Russia)

vkossobokov@gmail.com

The digital revolution started just about 15 years ago has already surpassed the global information storage capacity of more than 5000 Exabytes (5 × 1021 bytes) per year. Open data in a Big Data World provides unprecedented opportunities for enhancing studies of the Earth System. However, it also opens wide avenues for deceptive associations in inter- and transdisciplinary data and for inflicted misleading predictions based on so-called "precursors". Earthquake prediction is not an easy task that implies a delicate application of Statistics. So far, none of the proposed short-term precursory signals showed sufficient evidence to be used as a reliable precursor of catastrophic earthquakes. Regretfully, in many cases of seismic hazard assessment (SHA), from term-less to time-dependent (probabilistic PSHA or deterministic DSHA), and short-term earthquake forecasting (StEF), the claims of a high potential of the method are based on a flawed application of Statistics and, therefore, are hardly suitable for communication to decision makers. Self-testing must be done in advance claiming prediction of hazardous areas and/or times. The necessity and possibility of applying simple tools of Earthquake Prediction Strategies, in particular, Error Diagram, introduced by G.M. Molchan in early 1990ies, and Seismic Roulette null-hypothesis as a metric of the alerted space, is evident. The set of errors, i.e. the rates of failure and of the alerted space-time volume, can be easily compared to random guessing, which comparison permits evaluating the SHA method effectiveness and determining the optimal choice of parameters in regard to a given cost-benefit function. These and other information obtained in such a simple testing may supply us with a realistic estimates of confidence and accuracy of SHA predictions and, if reliable but not necessarily perfect, with related recommendations on the level of risks for decision making in regard to engineering design, insurance, and emergency management.

# SMART PARKING

*P. Stanchev, J. Geske*

Kettering University

pstanche@kettering.edu

One of the challenge to build smartcities is the smart parking. Several solutions have been proposed: different types of sensors (magnetometers, light sensors, microphones, etc.), different communication technology (wired, wireless), and different types of cameras. Smart Parking is a system capable of extracting specific information from the captured images and different sensors. Solutions based on computer vision and big data are deployable on top of visual sensor networks. The IoT paradigm fits particularly well in urban scenarios as a key technology for the Smart City Concept.

The paper presents an efficient solution for real-time parking lot occupancy detection based on Convolutional Neural Network classifier, real time image segmentation and analysis, and streaming data. It takes in account different light conditions, parts of the day, and seasons. It has been used benchmarks collections for parking occupancy detection. Problems that we solved are: significant changes of lighting conditions - sunny, rainy and snowing days; different time of the day;partially occupant, moving cars and peoples, additional objects. In our approach we use OpenCV library (http://opencv.org/) and Python to find the frames spaces. The parking classification is done with mix of the following techniques: background subtraction, defining and analyzing moving cars, applying Gabor filters as feature extractor to train a classifier with empty spaces under different light conditions, using edge detection algorithms. Deep Learning that allow computers to learn complex perception tasks. With the help of Caffe system (http://caffe.berkeleyvision.org/) we train the neural networks.

We use HAAR CASCADE to detect moving cars. The system interface is done for mobile app to show the free space and live time images. The project is in process of realization on Raspberry Pi platform equipped with a camera module.

References

1. Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Carlo Meghini, Claudio Vairo, Deep learning for decentralized parking lot occupancy detection, Expert Systems With Applications 72 (2017) 327–334
2. Paulo R.L. de Almeida, Luiz S. Oliveira, Alceu S. Britto Jr., Eunelson J. Silva Jr. Alessandro L. Koerich, PKLot – A robust dataset for parking lot classification, Expert Systems with Applications 42 (2015) 4937–4949

*Prof. Peter Stanchev* is Chair of the Software Engineering and Information Systems Department at the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences. He has forty years of professional experience in of multimedia systems, database systems, multimedia semantics, education, open access to scientific information and data and medical systems. He is also a professor at Kettering University, Flint, Michigan, USA. He has M.Sc., Ph.D. and D.Sc. in Mathematics/Computer Science from Sofia University. He has published 2 books, more than 200 chapters in monographs, journal and conference peer-reviewed papers, more than 200 conference papers and seminars, and has had more than 1500 citations, h-index - 33, impact factor – 77.03. Serving also on many database and multimedia conference program committees, he is currently editor-in-chief and member of the editorial boards of several journals. He is the Bulgarian representative in the EU OpenAIRE 2020 project.

*Dr. John G. Geske*, department head and professor of Computer Science. Ph.D. in Computer Science from Iowa State University, with expertise in software engineering, computational complexity, theoretical foundations, discrete mathematics, logic and the philosophy of computing. Geske, who has taught at Kettering since 1994, describes Kettering students as driven, self-assured and hard-working. In his eyes, they are also more career-oriented and inquisitive than students at other schools.

# DATA PUBLICATION: MOVING BEYOND THE METAPHOR

*S. Callaghan*

Science and Technology Facilities Council (STFC)

sarah.callaghan@stfc.ac.uk

Metaphors are a quick and easy way of grasping (often complicated) concepts and ideas. However, they are a tool that should be used carefully. There are as many arguments about how datasets are like cakes (they're a structured object created of raw materials, requiring set processes to create, and are generally more palatable and usable if presented in an appropriate way) as there are about how datasets aren't like cakes (datasets aren't necessarily physical objects, consumption of a dataset doesn't mean the dataset no longer exists, datasets can be transported from one place to another via the internet).

Yet it can be easy to pigeonhole a dataset into being a special class of, for example, academic paper. On one hand, this means that the tools and services for scholarly publication can be utilised to transmit and verify datasets, improving visibility, reproducibility, and attribution for the dataset creators. On the other hand, if a dataset doesn't fit within certain strict criteria to meet the "academic publication" mould (for example because it is being continually versioned and updated, or it is still being collected and will be for decades) it can be considered to be of not as much value to the community. It is often said that "all models are wrong, but some are useful". The key thing is to determine the usefulness and limits of models and metaphors, especially when trying to develop new processes and systems. This paper further develops the metaphors for data outlined in Parsons and Fox (2013), given the real world context of data stored in the Centre for Environmental Data Analysis (CEDA) – a discipline-specific environmental data repository, and the processes that created the datasets. It tests the metaphors against specific datasets, identifying where they work, and where they aren't helpful. It also discusses the processes of data publication from the viewpoint of a repository manager, who is aiming to ensure the quality, longevity, usability and sustainability of the data in the archives, as well as ensuring appropriate attribution and credit is given to all of the actors in the dataset creation and dissemination process.

Parsons, M. A., & Fox, P. A. (2013). Is Data Publication the Right Metaphor? Data Science Journal, 12, WDS32–WDS46. doi:http://dx.doi.org/10.2481/dsj.WDS-042

# ACCESSING SEISMIC HAZARD AFTER LARGE EARTHQUAKES

*P. Shebalin*

Institute of Earthquake Prediction Theory and Mathematical Geophysics

p.n.shebalin@gmail.com

Strong aftershocks after large earthquakes often pose a hazard comparable to the main shock. Significant part of the devastating aftershocks occurs within first hours after the main shock. The hazard of those aftershocks is intrinsically bounded to the main shock, because the danger of the construction collapses right after earthquakes is generally evident. The hazard assessment of later aftershocks is however an independent task. A feature of this task is a possibility to use the information about the early aftershocks, and not only parameters of the main shock. In the framework of the Russian Science Foundation Project 16-17-00093 "Development of information system for automatic seismic hazard assessment after large earthquakes based on geophysical monitoring" we develop a WEB-based System for Aftershock Hazard Assessment (AFCAST – www.afcast.org). The system software downloads earthquake data every 2 hours from ANSS Comprehensive Earthquake Catalog (ComCat, http://earthquake.usgs.gov/data/comcat/) provided on-line by USGS. Currently the system is aimed to assess hazard of aftershocks of M5.5+ after earthquakes of M6.5+. The access to the system is open to the registered users only. At this stage of the development, the system estimates in quasi real time the area where strong aftershocks are expected. This area is modeled using epicenters of the first 12 hours aftershocks. The next step is the estimation of the maximum magnitude of future aftershocks as a function of the current time and parameters of the past aftershocks. We plan to include this function by the end of 2017. Currently, we develop the theoretical basis. We compare three basic approaches. The first is based on the Reasenberg-Jones model of the aftershock decay formed by an independent temporal decay of the number of aftershocks according to the Omori-Utsu law, and the magnitude-frequency distribution according to the Gutenberg-Richter law. The second approach is based on the "generalized" Omori law, the study of the decay of various physical characteristics of the relaxation process through aftershocks, such as total scalar seismic moment, total faults area etc. The third approach can be entitled "Time-dependent Båth law", we study the difference of the magnitude of the main shock and the magnitude of the largest delayed aftershock as a function of the delay time from the main shock.

# VALUE CREATING THROUGH DATA SCIENCE PROJECTS: INSUFFICIENCY OF THE STANDARD WORKFLOWS

*A. Nesvijevskaia*

Conservatoire National des Arts et Métiers

anna.nesvijevskaia@gmail.com

While Big Data phenomenon is touching an increasing number of commercial and scientific organizations, the lag between potential and concrete benefits from field projects is still significant. Data Science, as one of the main articulations between the data and the knowledge useful for decision-making in these projects, is pointed out as a trigger of value creation through a dedicated workflow. Although the workflow in use by Data Scientists is issued from Data Mining process models, especially CRISM-DM which still dominates in the profession, it has evolved in order to take in account not only the new technical possibilities, but also the increasing expectations and uncertainties. The aim of this research paper is to establish the categories of potential value creation observed on Data Science projects, compare the standard model with its application in the field workflow, and highlight the characteristics of the model which omit, or ever restrain the concretization of the potential value creation. Besides minor chronological adjustments of the CRISP-DM process, this work shows a dangerous lack of initial consideration for the operational purpose of the data science results, including their appropriation by the practitioners, and for the capitalization of indirect results, such as practitioner's knowledge, data science competences or data quality leveraging. Thus, the lag between potential and concrete benefits is justified by the fact that the standard workflow risks to uproot the Data Science project from the practitioner's activity. Several tracks for this risk mitigation are discussed, including Human-Data mediation.
This paper is based on 3 years of field observation and action research, which represents more than a dozen complete projects, realized with an established French Data Science Company working since 2008 in health, chemistry, perfume, insurance and other sectors. The objective of this research, anchored in Information and Communication Science, is to make a significant contribution to the interdisciplinary understanding of the subject, and focus the proposal of adaptation of the standard data science workflow.

# GENERATION AND ANALYSIS OF BIG DATA FOR ACCELERATED MATERIALS DESIGN

*I. Abricosov, M. Nezhurina, A. Khvan*

National University of Science and Technology “MISiS” (NUST “MISiS”, Russia)

abrikosov_igor@mail.ru

The time it takes to discover advanced materials and to prove their usefulness to a commercial market is far too long. There is a need to reduce it significantly, from 10-20 years at present to 5-10 years or less. Basic science has a leading role in achieving this goal. We will address the challenge along the entire materials design chain and characterization of novel materials attractive to our industrial partners. Using methods and tools of Big data we will qualitatively improve our knowledge of the behavior of matter and employ it to increase information technology contents in new materials. The exponential growth of data science has brought the information revolution. Data-driven methods are now in ubiquitous use in life sciences, economics, social networks, etc. It has been suggested that the visionary idea of integrating materials development with data-driven methods, materials informatics, is going to bring a disruptive paradigm shift in materials science. Large-scale programs have been launched, like the Materials Project Center for Functional Electronic Materials Design in the USA. In 2016, three Centers of Excellence for data-driven research in materials science start their activity in Europe: MAX, NOMAD, and ECAM. It is of crucial importance that Russia participates and takes a leading role in this international development. Hence, we propose a multidisciplinary project "Materials Informatics and Quantum Synthesis" based on opportunities offered by the revolutionary progress in quantum physics, superconducting and information technologies, computational science and computer power. We will generate, collect, curate, and explore Big Materials Data. We will create databases of relevant materials parameters. Success of our project can be measured by its main deliverable, a "Materials Informatics" demonstrator, which takes the form of licensed functioning software attractive to broad communities of materials scientists and engineers, and used by our stakeholders for the knowledge-based materials design. We note that Big Materials Data could hardly be generated in physical experiments alone. However, recent advances in condensed matter theory made it possible to carry out virtual computer experiments with accuracy comparable to a physical experiment. A material modeling has become mature enough to start generating truly Big Materials Data. With this data, we will produce qualitatively new insights, beyond the reach for conventional research techniques. Within our project, we will capture the scientific context of the defined concepts to support modeling and prediction in variable environments and applications.

Our project will be carried out in collaboration with internationally leading partners in academia, including NoMaD EU Centre of Excellence, National Institute of Standards and Technology, Harvard University, EPFL Lausanne, and Pennsylvania State University. The key Russian partners include Lomonosov MSU, Tomsk State University, JIHT RAS and IGIC RAS. Our commercial stakeholders include Scientific Group Thermodata Europe (SGTE), OMK, IBS, IBM, Yandex and Hewlett Packard. It is strategically important to improve current practices, where theoretical research is directed primarily towards an explanation of earlier experimental findings. We will introduce new dimension provided by big materials data. Within 5 years, our work will result in paradigm shift, substituting the trail-end-error approach totally dominating Russian materials science at present by a knowledge-based materials design, and making computer simulations and data analysis into a natural first step of the design process. Moreover, we will develop new educational programs at undergraduate and graduate levels, as well as training programs for specialists of R&D centers of industrial companies. This will make students and young researchers involved in the project into highly attractive employees for academia and industries.

# UPDATES TO THE GIGADB OPEN ACCESS DATA PUBLISHING PLATFORM

*S. Xiao*

GigaScience

jesse@gigasciencejournal.com

GigaScience (www.gigasciencejournal.com) aims to revolutionize publishing by promoting reproducibility of analyses and data dissemination, organization, understanding, and use. As an open access and open-data journal, we publish ALL research objects (data, software tools and workflows) from 'big data' studies across the entire spectrum of life and biomedical sciences. The journal's affiliated database, GigaDB (www.gigadb.org) serves as a repository to display the data and tools associated with GigaScience publications, all under a public domain (CC0) waiver. GigaDB aims to improve the data reproducibility and reuse maximized with open licenses. It is running on the PostgreSQL database and has the well design schema to accommodate the growing variety of data. We define the dataset as a group of files (e.g., sequencing data, imaging files, software programs, virtual machines). Over 331 datasets (>30TB in size) have been made available in GigaDB to date. Through our association with DataCite, each dataset in GigaDB is assigned a DOI that can credit future use of these research objects in other articles. We installed the JBrowse (a dynamic web platform) in GigaDB for genome data visualization and analysis. Recently GigaDB has integrated with protocols.io (protocols repository) and CodeOcean (a Cloud-based executable research platform) to browse protocols and software, and run protocols on your smartphone and execute code on AWS. All datasets in GigaDB are indexed and has searchable metadata enabling discoverability and reuse. We have updated our schema and now have a REST API to retrieve and search all metadata held in GigaDB. It is possible to have results return all metadata for each dataset with "hits" to the search term, or to specify a particular portion of the metadata, these portions are currently "dataset", "sample" and "file". The current API returns result in XML, and it is planned to have the option to specify the format to be JSON or ISA2.0-JSON in the next version. The website (www.gigadb.org/site/help#0.1_API) provides detailed instructions on how to use the GigaDB API.

# URBAN BIG DATA AND SUSTAINABLE DEVELOPMENT GOALS: CHALLENGES AND OPPORTUNITIES

*A. Kharrazi*

International Institute for Applied Systems Analysis (IIASA)

ali@pp.u-tokyo.ac.jp

Cities are perhaps one of the most challenging and yet enabling arenas for sustainable development goals. The Sustainable Development Goals (SDGs) emphasize the need to monitor each goal through objective targets and indicators based on common denominators in the ability of countries to collect and maintain relevant standardized data. While this approach is aimed at harmonizing the SDGs at the national level, it presents unique challenges and opportunities for the development of innovative urban-level metrics through big data innovations. In this article, we make the case for advancing more innovative targets and indicators relevant to the SDGs through the emergence of urban big data. We believe that urban policy-makers are faced with unique opportunities to develop, experiment, and advance big data practices relevant to sustainable development. This can be achieved by situating the application of big data innovations through developing mayoral institutions for the governance of urban big data, advancing the culture and common skill sets for applying urban big data, and investing in specialized research and education programs.
*Keywords*: urban big data; sustainable development goals; big data public policy; SDG targets and indicators; big data research and education

# INFORMATION AND DATA MANAGEMENT REPOSITORY PLATFORM FOR NANOSCIENCE IN EUROPE

*R. Aversa[1], S. Cozzini[2], T. Jejkal[3]*

[1,2]Italian National Research Council (CNR), [3]Karlsruhe Institute of Technology (KIT)

aversa@iom.cnr.it

The project Nanoscience Foundries & Fine Analysis (NFFA-EUROPE) brings together European nanoscience research laboratories that aim to provide researchers with seamless access to equipment and computation. To achieve this goal, a distributed open-access research infrastructure as a platform supporting comprehensive projects for multidisciplinary research at the nanoscale, extending from synthesis to nanocharacterization to theory and numerical simulation, is implemented.
NFFA-EUROPE provides a single entry point for submitting and managing research proposals, and an Information and Data management Repository Platform (IDRP) supporting registration, access, sharing and publication of resulting experimental data.
The main goal of our work is to establish the IDRP as a pervasive tool within the project, allowing NFFA-EUROPE users to semi automatize registration of scientific data coming from many different instruments among the NFFA-EUROPE facilities, and identify the correct metadata making them retrievable accordingly to the FAIR Data Principles.
The IDRP, together with an introduced data management policy, will favor data shareability, will validate findings, and will promote reuse of data.
In collaboration with Karlsruhe Institute of Technology (KIT), we deployed the first prototype of the NFFA-EUROPE infrastructure on the CNR-IOM OpenStack cloud platform. The infrastructure comprises an instance of the NFFA-EUROPE portal as single entry point for the user, an instance of the IDRP as central data platform, and an instance of the local Data Management Service at CNR-IOM. The seamless integration into the local instrumentation is realized by an ad-hoc plugin for the Scanning Electron Microscope (SEM) instrument with automatic metadata inclusion. We also developed, packaged, and tested a Python Command Line Interface called 'click' (Command Line Interface Cnr-iom Kit) to the local Data Management Service. 'Click' allows to ingest/download instruments' data and basic metadata produced at the facility to/from the local Data Management Service. In order to be able to find data sets located at CNR-IOM via the NFFA-EUROPE infrastructure, each data set ingested into the local Data Management Service is also registered at the IDRP.Moreover, we applied transfer learning techniques for image recognition, automatic categorization, and labeling of nanoscience images obtained by SEM. We used the statistical outcomes from testing to deploy a semi-automatic workflow able to classify and label images from the SEM, before the ingestion to the local Data Management Service. Our approach paves the way towards the implementation of new methods and tools which can be applied to a wide range of nanoscience use cases and suitably tuned to resolve specific features of nanomaterials.

# ANALYSIS OF THE TECTONIC DEFORMATIONS IN THE KURIL ARC AFTER THE 2006-2007 SIMUSHIR EARTHQUAKES BASED ON SATELLITE GEODETIC DATA

*I. Vladimirova[1], L. Lobkovsky[2], Y. Gabsatarov[1], I. Garagash[3], B. Baranov[2], G. Steblov[3]*

[1]Geophysical Survey of the Russian Academy of Sciences,
[2]P.P. Shirshov Institute of Oceanology,
[3]Schmidt Institute of Physics of the Earth of the Russian Academy of Sciences

ir.s.vladimirova@yandex.ru

The region of Kuril island arc is one of the most seismically active regions of the world due to very high plate convergence rate. In 2006-2007, a doublet of Mw>8 earthquakes occurred in the central segment of the Kuril arc, the region with no great earthquakes for a century. More than 9 years of continuous satellite geodetic observations on Kuril Islands provide us unique material for investigating surface displacements caused by postseismic transient motion and the interseismic motion at a constant speed. The difference revealed in the directions of motions of the Earth's surface observed along the Kuril island arc provides evidence that different segments of the arc are in different stages of the seismic cycle. We used the keyboard model concept of the structure of island arc margins combined with the models of the nonstationary convective system in the upper mantle and viscoelastic relaxation in the asthenosphere and the upper mantle to explain the variety of motions observed along the entire length of the Kuril Islands. The direction of motion observed on islands on the southwestern and northeastern flanks of the arc is close to the direction of plate convergence, which reflects compression of the rear block of the island arc during the long interseismic stage of the cycle. At the same time, the stations located in the central part of the island arc between the Kruzenstern and Bussol straits displace towards the ocean. Their initially high velocities decrease in time gradually turning to the stationary interseismic state. The character and duration of these displacements indicates that a transient postseismic process exists in the source region of the 2006 earthquake, which is most likely caused by the viscous response of the asthenosphere. The possibility of viscoelastic relaxation is confirmed by the results of direct modeling. We used in modeling a density and lithostratigraphic model characterizing the modern state of the transition zone and the FLAC3D programming code. We also used and an open-source programming code VISCO1D by F. Pollitz to show that viscoelastic relaxation can last about ten years after the event before the region returns to the interseismic stage of stress accumulation. Similar behavior happens with the motion of the Urup island located southwest of the source zone of 2006 event, which indicates that the initial seismic dislocation in this direction was possibly developing in the first few months after the earthquake. The character of displacement of the Kharimkotan island located northeast of the source of the 2006 earthquake is clearly different from the motions of other islands. During the entire period of observations, its velocity kept the orientation directed to the ocean and it only slightly decreased in time, which can hardly be explained from the standpoint of the keyboard model. Such a character of displacements can be explained by applying the model of nonstationary horizontally widening upper mantle cell adjacent to the subduction zone. The observed process of sufficiently intense displacement of the station to the ocean due to the passive extension of the lithosphere caused by the return sublithospheric mantle flow may appear dominating, for example, in the conditions of weak interplate coupling. On the other hand, a similar situation can arise in the case of late pre-earthquake part of interseismic stage. It is noteworthy that the zone approximately limited by the Kruzenstern Strait and the Fourth Kuril Strait is related to the probable locations of future major earthquakes and the seismic "silence" here has continued since 1915. The work was supported by the Russian Sci. Foundation Project No. 14-50-00095.

# ANALYSIS OF THE TECTONIC DEFORMATIONS IN THE CHILEAN SUBDUCTION ZONE CAUSED BY THE 2010 MAULE EARTHQUAKE ON THE BASIS OF SATELLITE GEODETIC DATA

*Y. Gabsatarov[1], I. Vladimirova[2], L. Lobkovsky[3], I. Garagash[3], B. Baranov[2], G. Steblov[3]*

[1]Geophysical Survey of the Russian Academy of Sciences,
[2]P. P. Shirshov Institute of Oceanology,
[3]Schmidt Institute of Physics of the Earth of the Russian Academy of Sciences

y.v.gabsatarov@yandex.ru

The Chilean subduction zone is one of the most seismically active regions on Earth, due to the shallow depth of the seismogenic zone in combination with the high coupling coefficients and high plate convergence rate. Focal zones of earthquakes with magnitude M>8, registered in this region for the last 200 years, cover almost the entire length of the Chilean coast. From 1835, between the outbreaks of the 1960 Great Chilean earthquake and the 1985 Valparaiso earthquake, there was a seismic calm zone, the Darwin seismic gap, which was interrupted in 2010 with the Maule earthquake Mw = 8.8. We analyzed the data of 8 years of continuous observations at 27 stations of the Chilean GPS network in order to distinguish interseismic, coseismic and postseismic deformations in the vicinity of the 2010 earthquake. The analysis of interseismic and postseismic deformations is based on displacement velocities of points on the earth's surface estimated over the 1-year interval. We used the keyboard model concept of the structure of subduction regions combined with the model of viscoelastic relaxation in the asthenosphere and the upper mantle to explain the variety of motions observed over the region of Central Chile. Interseismic deformations were estimated from data of three stations. Two of them show a displacement in the plate convergence direction agreed with the model, while the third station is practically not moving. These differences in the behavior of the stations before the 2010 earthquake are well explained by the keyboard concept, in which in the interseismic stage weak seismogenic blocks can shrink more than the more consolidated rear block. The coseismic displacements registered at the stations closest to the epicenter of the 2010 event amounted to 1 to 3 meters, which characterizes the magnitude of the displacements of seismogenic blocks at the time of the earthquake. All the coseismic displacements are directed toward the ocean, which agrees well with the expected model movements of the blocks in the seismic stage of the seismic cycle. Data from first year after the 2010 event show high rate of displacements toward the ocean over the whole Central Chile region which indicates passing of aftershock stage of seismic cycle. Two years later, the aftershock stage is still ongoing, but the displacement rates of the relaxing seismogenic blocks and the rear block is substantially reduced. The peculiarities of the displacement velocity field three years after the Maule earthquake are the starting rotation of the displacement vectors in the frontal part of the subduction zone to the direction of plate convergence vector, which indicates the transition to the interseismic stage of the seismic cycle, and the absence of a change in the direction and magnitude of the displacement vectors in the rear block. Such a behavior of displacement vectors in the rear block can be explained by the model of viscoelastic relaxation in the asthenosphere and upper mantle, the occurrence of which was confirmed after the Maule earthquake. Over the next 4 years of observations, the process of restoring the stationary state of stress accumulation for seismogenic blocks in the frontal part of the subduction zone continues, in combination with the continuing displacement of the rear block by a viscous asthenospheric flow. The duration of the viscoelastic relaxation process for the Maule earthquake is estimated to last for more than 15 years. The work was supported by the Russian Sci. Foundation Project No. 14-50-00095.

# DATA FLOWS MANAGEMENT OF MINING NATURAL/MAN-MADE SYSTEMS INTEGRATED STATE MONITORING

*V. Cheskidov*

College of Mining

vcheskidov@yandex.ru

The reported study was funded by RFBR, according to the research project No. 16-35-60116 mol_a_dk Mining intensification and complex mining operational conditions necessitate constant monitoring of the mining natural/man-made systems state to ensure environmental and industrial safety. One of the most important objective while monitoring is to ensure the adequacy of the collection process, transmitting, processing and storing geodata that have a weak structure. The natural/man-made systems parameters have highly spatial and temporal variability and strong correlation ties, however it is often almost impossible to identify and evaluate them. Thereby monitoring systems of modern mining enterprises are redundant in data collection. Correct information flows distribution will ensure maximum use of information, cost reduction, and efficiency increase of online assessment of natural/man-made systems state. In coming years the development and active implementation of robotic and automated technologies will determine the need for methods developing and collecting information tools regarding the state of the geological environment and data flows management.

While open-pit mining, different types of sloping structures (open pit mines, dumps, hydraulic structures dams, etc.) required more detailed study. The numerical features characteristics of the sloping structure state are determined by the following groups of factors: physical-geographical, natural-geological, hydro-geological, geoengineering and technological. Each group has different dynamics of change over time.

The hydrogeological conditions of the territory can change very quickly as a result of high rainfall, a dramatic snowmelt, a redistribution of the pulp output in the hydraulic structures formation. This group of factors has the greatest impact on the stability of argillo-arenaceous mountain range. The mechanical properties of soils in the sliding zone vary much more slowly. A sudden change in the angle of internal friction and specific adhesion in argillaceous rocks can occur as a result of considerable moisture. However most of the rocks with destructuring and cluster shifts are displaced in completely watered state and there are no dramatic changes in their mechanical properties.

Analysis of the modeling experience of the sloping structures behavior shows that the most acceptable method for estimating the transition of an array from a stable state is a method based on the boundary indices determination on separate factors. At the same time, for each sloping structure critical indicators (hydrogeological, geoengineering and others) must be calculated in the following sequence:

1. to determine normative margin of stability rate based on structure class responsibility and technological requirements;
2. to build an geoengineering model at a particular point in time, indicating the physico-mechanical rock properties in the identified geoengineering elements and the level of the man-made (natural) aquifer;
3. to build a hydrogeological model, taking into account the filtration properties of the rock mass and man-caused sediments ;
4. to create a geomechanical model reflecting the spatial position of the landslide block, the position of the slip curve;
5. for a short-term forecast of the array state, calculate the critical levels of the aquifer at which the safety factor is correspondingly equal to the normative and 1.0;

6. to determine the main regularities of the spatio-temporal variability of the material properties composing the slope structure and the intensity of the change in the basic physicomechanical characteristics when the hydrogeological regime changes and there is a man-made load on the array;
7. to calculate critical levels of anthropogenic aquifer taking into account changes in physical and mechanical properties of rocks.

Forecasting the sloping structure condition according to the proposed scheme ensures reliable estimates of mountain range , while additional engineering and geological surveys are necessary to confirm and refine the models built.

# KEYBOARD MODEL OF SUBDUCTION DEFORMATION CYCLES FOR GREAT EARTHQUAKES

*L. Lobkovsky*

P. P. Shirshov Institute of Oceanology

llobkovsky@ocean.ru

The largest earthquakes are generated in subduction zones and the great earthquake rupture typically extends for hundreds of kilometers along a single subducting plate. These ruptures often begin or end at structural boundaries within the overhanging plate that are associated with the subduction of prominent bathymetric features of the plunging plate. The block-like structure of the front part of the overhanging plate (segmentation) of the Kuril, Aleutian, Japan, Peru-Chile, Sumatra, Solomon subduction zones is considered. The understanding of separated frontal blocks within island-arc and active continental margins as main seismogenic elements associated with the sources of great earthquakes has led to the formulation of the mechanical keyboard model of the subduction deformation cycles for the catastrophic events (Lobkovsky, 1982; Lobkovsky et al., 1991). According to this model frontal keyboard blocks are separated from each other by transcurrent vertical faults that reach the surface of a subducting plate. A longitudinal fracture zone separates them from the main arc massif or active continental margin, being situated at the distance over hundred kilometers from the trench. The principal source of mechanical energy for the seismogenic system under investigation is supplied by the movement of the subducting plate which interacts with keyboard blocks due to the dragging mechanism. This interaction results in gradual increase of compression in a frontal block which is jammed by the main island arc or the active continental margin massif in the rear. The stage of elastic-energy accumulation within each block persists for the major part of the periods between great earthquakes which occur due to the block energy release. The block projection onto the surface during the long-lasting energy-accumulated stage is associated with a seismic gap according to the given model. Release of the seismic energy of the whole keyboard block occurs when a critical value of the tangential stress is reached along the superior part of the contact surface between the block and the subducting plate. This results in a rupture of the contact surface accompanied by co-seismic displacement and by the great earthquake of the thrust type. As a result the block abruptly shifts in direction of the ocean losing part of the accumulated compression energy, while the adjacent blocks which have not reached the critical state staying at the same locations. The ruptured block-plate contact surface is mechanically weakened just after the main shock, and post-seismic block energy release and effective block motion take place oceanward. Such displacement of the bottom of the block relative to the plate surface is accompanied by weaker aftershocks in those places where the residual stress concentration are the highest. The sources of great earthquakes usually coincide with keyboard blocks but occasionally they can spread over several adjoining blocks if triggering conditions occur, resulting in the formation of abnormally large sources as in case of the December 26, 2004 Mw=9.2 Sumatra earthquake. The work was supported by the Russian Sci. Foundation Project No. 14-50-00095.

# REVEAL OF THE PRE-SEISMIC PHASE OF THE SEISMIC CYCLE FROM SPACE GEODETIC OBSERVATIONS OVER THE AREA OF TOHOKU EARTHQUAKE 2011

*G. Steblov*[1], *I. Sdelnikova*[2]

[1]Schmidt Institute of Physics of the Earth of the Russian Academy of Sciences
[2]Geophysical Survey of the Russian Academy of Sciences

steblov@ifz.ru

Deformation processes in subduction zones, the features of their temporal evolution and spatial variations, their correlation with the seismicity are the key issues for understanding the mechanisms for the strongest earthquakes preparation. Compared to seismological observations, satellite geodetic observations (GPS) reflect not only fast, but also slow processes in the earth's crust and asthenosphere, including that portion which is not exhibited seismically, but contributes significantly to the accumulation of the deformation potential or to its release through aseismic motion (creep), viscoelastic relaxation in the asthenosphere, etc. The long-term satellite geodetic observations that have been collected globally over the past decades have made significant adjustments to the pre-existing understanding of the dynamics of many seismically active regions. Comparison of the of surface displacement velocity with the plate convergence rate in the subduction zones discovered mismatch between the observed motions and the hypotheses of complete mechanical coupling in the interplate contact zones, which is broken only during the seismic jumps. For quantitative relation between the above velocities, dislocation models of deformation of elastic or viscoelastic medium are used, which provide the basis to solve the problem of determining the geometry of the interplate coupling in the subduction zone from observations of surface displacements. Such an inverse problem has a stable physically feasible solution with discretization adequate to the observation density and with regularization by the solution norm minimization while retaining the condition of statistical agreement between the residuals of the observational equations and the errors of the raw measurements. Analysis of temporal variations of interplate coupling in the contact zone requires separation of effects from various mechanisms generating deformations in subduction zones: in addition to plunge of the oceanic plate slab and slipping along the asperities the observable surface displacements exhibit coseismic jumps, postseismic transient processes, instrumental variations of various periodicity (diurnal and seasonal). Separation of all these effects is based on the regression analysis of time series by characteristic features in the time dependence of each listed mechanism. Application of the described approaches to observations over the Japanese satellite geodetic network provided by the Geospatial Information Agency of Japan made it possible to identify the specific spatial-temporal deformations preceding the strongest subduction earthquakes. The localization of the focal zone of one of the strongest earthquakes within the Japan subduction zone in 2011 is characterized by the maximum gradient of the interplate coupling in the direction of the strike of the contact zone. In addition, this event was preceded by local temporal variations in the zone of the maximum coupling gradient, as well as variations involving the entire subduction zone, which makes it possible to distinguish the instrumentally recorded pre-seismic phase of the seismic cycle.

# NUMERICAL MODELING OF THE STRONGEST TSUNAMIS CAUSED BY GREAT EARTHQUAKES IN SUBDUCTION ZONES FOR THE EXAMPLES OF 21ST CENTURY CATASTROPHES

*R. Mazova, L. Lobkovsky*

P. P.Shirsov Institute of Oceanology RAS

raissamazova@yandex.ru

The simulation of the catastrophic tsunami generated by the great earthquakes of the magnitude M ~ 9 in subduction zones was carried out. The need for such studies is related to the recent catastrophic undersea earthquakes on December 26, 2004 in the Indian Ocean near the island of Sumatra, on November 15, 2006 in the central part of the Kuril island arc, on February 27, 2010 in northern Chile, and on March 11, 2011, near the northeastern coast of Japan and, following them, the catastrophic tsunamis that led to a huge number of casualties and immense material damage, especially in Sumatra and Japan. Conventional models were unable to explain the unusual characteristics of the earthquake process, such as the unusually long source of the earthquake in Sumatra (1200 km) and the unexpected magnitude for the earthquake in Tohoku (M = 9), which caused a strong tsunami with a height of 40 m on the coast. The keyboard model of the seismic cycles, firstly proposed by Lobkovsky in 1982 [1] and developed in subsequent works [2,3] allows us to explain not only the scale of the source and the strength of great earthquakes, but also their long duration. This model also provides broad possibilities to calculate the generation of a tsunami source and the further propagation of the tsunami in the water area. Such calculations were carried out for the Kuril-Kamchatka, Sumatra-Andaman, and Komandorsky subduction zones. Moreover, for the first time in the numerical modeling of the tsunami, the localization of tsunami source in the Kuril Islands was closely predicted six months before the event on November 15, 2006. Based on the mechanical model of interaction between the oceanic lithosphere and island-arc blocks, a good coincidence of the maximum tsunami wave heights with observation data along the coast of Japan, including in the Sanriku area, was obtained [4]. Using the keyboard model [1-3], the authors also carried out numerical modeling of a possible earthquake and tsunami in the Aleutian seismic gap.

Acknowledgements

The work was supported by the Russian Sci. Foundation Project No. 14-50-00095.

References

1. Lobkovsly L.I. (1982). The model of seismic gaps and catastrophic earthquakes in island arcs. In: Proceedings of 5th School of marine geology.1982 (ed. A.P.Lisitsin). Moscow. P.P.Shirshov Inst. of Oceanology RAS. T.2. p.41-42.
2. Lobkovsky L.I., & Baranov B.V. (1984). Keyboard model of strong earthquakes in island arc and active continental margins. Doklady of Acad. of Sci. of USSR, 275, 843-47.
3. Lobkovsky,L.I., Kerchman, V.I., Baranov, B.V., Pristavakina,E.I. Analysis of seismotectonic processes in subduction zones from the standpoint of a keyboard model of great earthquakes. (1991). // In: L.P.Zonenshain (Editor), The Achivements of PlateTectonics in the USSR. Tectonophysics, 199: 211-236.
4. L.Lobkovsky, I.Garagash, B.Baranov, R.Mazova, N.Baranova, Modeling Features of Both the Rupture Process and the Local Tsunami Wave Field from the 2011 Tohoku Earthquake // Pure Appl. Geophys. (2017). doi:10.1007/s00024-017-1539-5 , pp.1-20.

# DIGITAL GEOLOGICAL EXPLORATION – RELOADING IN GEOLOGY OF THE 21 CENTURY

*O. M. Prischepa, A. M. Karnaukhov*

All-Russia Petroleum Research Exploration Institute

omp2007_61@mail.ru

Today the petroleum complex of Russia is in a pre-crisis state. Geological exploration is not conducted. It is necessary to reload it's elements on the base of fundamentally new technologies with an emphasis on the exploitation economy, on the using of hydrocarbon systems and on increasing the competitiveness of petroleum complex.

Digital geological exploration is an active transition to digital technologies, big data technologies; the introduction of robotics and artificial intelligence elements into geological exploration is not a fantasy; it is a transition to a new scientific and technological level in the geology of the 21st century, and this is the matter of the immediate and long-term perspective covering the period 2020-2040.Effective geological exploration can be achieved by a profound transformation of the main business processes in the implementation of innovative technologies:

- Application of new sounding methods for areas with potential fields (lidars);
- Application of new information processing methods and visualization of results (on-line);
- Introduction of decision support system (with elements of artificial intelligence);
- Introduction of artificial intelligence elements in the processes of geological and geophysical modeling, interpretation of primary data and decision making;
- Introduction of intelligent drilling of search and exploration wells;
- Introduction of robotics in geological exploration processes;
- Introduction of the knowledge base with artificial intelligence;
- Introduction of big data technologies.

Big data from the oil and gas industry significantly influence on the adoption of managerial and technical decisions.

The use and analysis of big data take place both at the stages of exploration (search) works and at the stages of the exploitation of discovered oil and gas fields.

The task of detecting hydrocarbon accumulations, often occurring at great depths (up to 5-7 km) without pronounced signs, allowing it to be identified directly in any manifestations of a changing observation environment (the anomalies of geophysical fields, in the first place) makes it necessary to develop approaches both in terms of improvement of equipment for observations, and methods of collecting and processing information.

It can be attributed to the most common and applicable method to identify well locations, which use in practise by oil and gas companies – seismic exploration.

Methods of conducting seismic surveys are constantly being improved, the volume of digital information obtained during seismic exploration today is already calculated by terabytes and petabytes. On the background of increase opportunities to obtain important information through the re-processing of previous years seismic data, and the use of more sophisticated computing devices, analysis and re-processing of big data sets is one of the most important and urgent tasks.

Oil and gas companies are currently collecting data in real time, which also implies their storage and operational processing for making current decisions. For example, during wells drilling, it is important not only to obtain current information from a particular well, but also to process (compare) with drilling data in neighboring wells in real time and this is especially

important for drilling in unconventional (low permeable) shale strata with the aim of operative adaptation of drilling technology and opening of productive layers.
Often big companies make huge and costly mistakes due to the lack of systematic analysis of big arrays of geological information obtained from neighboring areas. Such errors are typical for works in areas with difficult geological conditions and in conditions of complex low-permeability strata development.
Analytics of big data can play an important role in improving the efficiency of works in unconventional shale objects (Bazhenov suite, Domanic formation, Tyumen suite, etc.). While it applies to areas of Western Siberia, Timan-Pechora and the Volga-Ural, where a big number of wells have been drilled that have penetrated the shale strata, but they have not been studied due to the dominant concept of oil and gas formation and accumulation, where these layers were associated only with the places of the oil and gas origin, but not with the places of accumulation and, especially, exploitation due to low filtration properties.
The strategy for analyzing big data is relevant for geological exploration, both at the stage of analyzing the geological structure, analyzing the selection of the most promising priority site, and analyzing the choice of drilling points for priority wells and the strategy of drilling in identified reserves of oil and gas.
Thus, digital geological exploration will achieve the following results:

- Increased efficiency and effectiveness of decision-making;
- Geological study of areas and water zones, as well as the results reliability will approach 100%;
- Geological and geophysical modeling and building of geological sections with complex analysis in real time;
- Expenses for geological exploration work and their terms will be reduced at times;
- Environmental problems in geological exploration will disappear.

# TSUNAMI IN 21ST CENTURY

*E. Kulikov*

Shirshov Institute of Oceanology RAS

kulikove@gmail.com

A series of powerful earthquakes and subsequent catastrophic tsunamis occurred in the early 21st century - Andaman 2004, Chile 2010, Tohoku 2011. Similar amplification of seismic activity was observed in the mid-20th century: Kamchatka 1952, Chilean 1960, Alaska 1964. It should be noted that for the 106-years' period of observation (1906-2011) the contribution of these 6 largest earthquake events to the total energy released by all the earthquakes is more than a half.

We will try to assess the progress which has been made over 50 years in understanding tsunami process, especially in practical aspects: 1) measurement technology and optimal tsunami monitoring networks, 2) operational forecast and warning systems 3) an inundation modelling studies and effective technology of risk assessment.

In the USSR the national tsunami warning service was established after the tragic lesson of the Kamchatka Tsunami of 1952. At that time, the decision to announce tsunami warning was based on seismic monitoring (magnitude criterion) only, which led to a large number of false warnings and evacuations of population. Today modern sea level observational systems (for example, DART) allow us to track the tsunami practically from the time it is generated, which allows us to make reliable tsunami forecast for the coastal area. Current progress of computer technologies provides for development of inundation numerical modelling studies which are applied in both the operational forecast and tsunami hazard assessment. Despite significant progress in science and technology tragic consequences of the Andaman and Tohoku tsunamis could not be prevented. If in the first case the main cause of death of more than 200 thousand people was lack of an operational tsunami warning system in the Indian Ocean, while during the Japanese tsunami of 2011 the main factor which led to more than 15 thousand people lost, as well as resulted in the Fukushima nuclear power plant disaster was underestimation of the tsunami hazard level at the coast of the Honshu Island. Protective structures had not been designed for such huge tsunami waves.

We will consider examples of new technologies successful use for studying the generation and propagation of tsunami waves, as well as for tsunami risk assessment at the coast, using several strongest events of the 21st century – the Andaman tsunami (2004), Simushir tsunami 2006 and 2007 and Tohoku 2011 as examples.

# THE WORLD'S LARGEST OIL AND GAS INDUSTRIALLY EXPLORED DEPOSITS: ROSA DATABASE AND GIS PROJECT

*A. Odintsova, A. Gvishiani, A. Rybkina*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

a.odintsova@gcras.ru

The target of this paper is the ROSA database and the GIS project as an information basis for analytical study of hydrocarbon extraction methods development in the world. The period from 1900 to 2000 is considered.

To achieve the target the authors have implemented a number of tasks: data collection and processing; development and filling the database with geospatial data on largest hydrocarbon deposits. The ROSA is based on systems analysis and integration of specialized domestic and international open sources, data compilation and completion of the attribute fields and its further assembling. Accumulated data could be divided into two types: static and dynamic. Static data included the deposit parameters that do not change over time. On the other hand, dynamic data is constantly changing. To increase the visibility of the ROSA database the web service with advanced functionality was created based on the Esri Geoportal Server software platform: search by parameter presets; viewing and filtering of selected data layers using online mapping application; sorting of metadata, including bibliographic information for each field. The collected and processed information will allow to systematically analyze the methods and volumes of hydrocarbon exploration in the world and assess the place of oil and gas industry development in different countries. Such approach enables the analysis of the geospatial data in the perspective of time and the assessment of the oil and gas industry development worldwide in 20 century.

Expanding and filling the database with large amount of smaller oil and gas deposits of Russia in the 21 century is essential to comparatively assess the Russia's role on global arena and to study the hydrocarbon exploration development in the country. This project is being developed in the framework of the program No. I.28п of Russian Academy of Sciences "Study of historical process of science and technology development in Russia: its place in the world scientific community, social and structural transformation".

# HOW DO UNCERTAINTIES IN BIG DATA SETS DIFFER FROM DOMAIN SCIENCE DATASETS? THE CASE OF CLIMATE SCIENCE

*M. Wälchli, B. Knüsel*

Swiss Federal Institute of Technology Zurich

marius.waelchli@usys.ethz.ch

To understand the epistemic implications of uncertainties in datasets is of major importance. In climate science, theory, domain knowledge, plausible assumptions and models are used to filter, correct, and extent measurements, and finally craft them into datasets used for scientific inquiry (Frigg, Thompson, & Werndl, 2015a). Uncertainties can be of different kind, but the literature on classification of uncertainties is still in its "pre-paradigmatic" phase (Frigg, Thompson, & Werndl, 2015b, p. 971). This paper distinguishes between uncertainties in datasets concerning the representation of the phenomena of interest, and uncertainties due to reuse and recombination of available datasets, which touches upon the complexity and transparency of the dataset generation.

Incautious usage of datasets and data curation can for instance lead to false or distorted conclusions. A recent example is the global warming hiatus, an assumed pause or slowdown in the global temperature increase, which kept the scientific community occupied for over a decade. A study showed that the usage of different datasets, representing the same phenomenon - beside applying different definitions - was leading to differing views among scientists (Medhaug, Stolpe, Fischer, & Knutti, 2017). Although the importance and usage of Big Data methods is increasing and has already penetrated many components of climate sciences (Knüsel & Wälchli, 2017), the role of uncertainties in (exhaust) Big Data for scientific inquiry, is poorly understood. Big Data allows for novel, mainly statistical modelling approaches. The coupling from impacts to damages often cannot be directly modelled by principle-based relations, which is mainly due to lack of understanding of the underlying processes. Additionally, Big Data allows for higher resolution and more ubiquitous measurement of phenomena. In a recent example, two studies employed Twitter data to predict damage caused by Hurricane Sandy (Kryvasheyeu et al., 2016; Shelton, Poorthuis, Graham, & Zook, 2014). In contrary to traditional climate science, which has already built up competence in the collection, analytics and standardization of large amounts of data, in climate impact studies, there is an increased use of unstructured datasets which lack standardized collection and curation procedures.

Nonetheless, in order to understand how the uncertainties differ across different datasets – for instance between traditional and Big Data – a classification system for datasets, relevant to uncertainties needs to be established first.

Building on a relational concept of data (Leonelli, 2015), we propose a six dimensional classification, which is relevant in order to describe the discussed uncertainties in datasets. The dimensions are the phenomena-data distance, the aggregation level, the velocity, the processing network, the curation and the sampling. We apply the dimensions to two datasets and their processing history, namely a dataset describing global temperature extremes (HadEx2) (Donat et al., 2013), a well curated dataset, and to a social media dataset used in disaster damage assessment (Kryvasheyeu et al., 2016), representative of what people typically perceive as Big Data.

By providing a first proposition for a systematic classification of datasets relevant for the discussed uncertainties, we show that, (1) the dimensions normally used to distinguish traditional versus Big Data are neither sufficient nor all relevant in order to understand the different kinds of uncertainties within datasets. (2) Both investigated datasets differ concerning both discussed uncertainties. This depends on the manifestation of the proposed dimensions and

the phenomena of the target system under investigation. These dependencies have direct implications for establishing curation practices for datasets that could be labelled "Big Data". The proposed dimensions provide a first approach to better understand uncertainties and their propagation in datasets. Based on an increased understanding of such uncertainties in datasets, guidelines for data curation in the age of Big Data and can be established.

References

1. Donat, M. G., Alexander, L. V, Yang, H., Durre, I., Vose, R., Dunn, R. J. H., … Villarroel, C. (2013). Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century : The HadEX2 dataset, 118, 2098–2118. https://doi.org/10.1002/jgrd.50150
2. Frigg, R., Thompson, E., & Werndl, C. (2015a). Philosophy of Climate Science Part I : Observing Climate Change, 12, 953–964.
3. Frigg, R., Thompson, E., & Werndl, C. (2015b). Philosophy of Climate Science Part II : Modelling Climate, 12, 965–977.
4. Knüsel, B., & Wälchli, Mg. (2017). How Big Data changes the role of domain theory – The case of climate science. Manuscript in Preparation.
5. Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Hentenryck, P. Van, Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. Science Advances, 2(March), 1–12. https://doi.org/10.1126/sciadv.1500779
6. Leonelli, S. (2015). What Counts as Scientific Data ? A Relational Framework, 82(December), 810–821.
7. Medhaug, I., Stolpe, M. B., Fischer, E. M., & Knutti, R. (2017). Reconciling controversies about the "global warming hiatus." Nature, 545(7652), 41–47. https://doi.org/10.1038/nature22315
8. Shelton, T., Poorthuis, A., Graham, M., & Zook, M. (2014). Geoforum Mapping the data shadows of Hurricane Sandy : Uncovering the sociospatial dimensions of " big data ." GEOFORUM, 52, 167–179. https://doi.org/10.1016/j.geoforum.2014.01.006

# CAPTURING THE COMPLEXITY IN RESILIENCE USING A FEDERATED DATA PLATFORM

*J. R. Stevenson[1], V. Ivory[2], D. Johnston[3]*

[1]Resilient Organisations Ltd.,
[2]OPUS,
[3]Massey University

joanne.stevenson@resorgs.org.nz

The human brain processes information visually. So, how can we help people see resilience? Researchers associated with the Resilience to Nature's Challenges National Science Challenge in New Zealand have embarked on a cross-institutional, transdisciplinary research programme to identify new ways to assess the disaster resilience of built and human systems, and to identify actions that will reduce casualties, enhance post-disaster productivity, and mitigate long-term social dysfunction caused by disasters.

We have amassed a bank of over 400 indicators that have been used to capture some aspect of resilience to disruption in the complex systems that make up life in New Zealand. These indicators are being matched to available secondary data and geocoded in a way that will allow us to see the differences across space and time. Researchers, communities, and resilience practitioners will be able to mine existing public data in a resilience context to assess their strengths, weaknesses, assets, and capabilities, and assess progress. It is clear, however, that this type of assessment does not capture sufficient nuance for people on the ground.

In the early stages of this programme, the research team socialised prototypes of maps displaying a composite resilience index. The response to early versions of the maps were polarized. The maps stimulated conversations but in some cases shut-down meaningful discourse, in part because there was not adequate context around the numbers and colours in the display.

There are significant challenges and pitfalls to displaying and sharing visualisations of resilience measurements, such as conveying measurement quality issues to users with different degrees of data literacy, clearly linking information about risk and vulnerability to solutions and ongoing efforts to make resilience improvements, and others using your information as a weapon rather than a tool.

It is important to find ways to assess and track the progress of complex social phenomena like resilience to disasters, but different audiences speak different measurement 'languages'. This paper explores the way different audiences relate to resilience measurement information and discusses how data can be filtered and displayed in ways that are meaningful and stimulating to different users. We also present preliminary efforts to integrate different types of resilience measurement information through metadata and spatial data literacy in a way that allows data from quantitative composite indices to narratives and photos to be hosted and displayed in a federated data system.

# BIG DATA IN MINING AND METALLURGICAL TECHNOLOGIES: APPLICATIONS AND PROSPECTS

*M. Kruglov, M. Nezhurina*

National University of Science and Technology “MISiS” (NUST “MISiS”, Russia)

michkruglov@mail.ru

Using Big Data in Mining and Metallurgical Technologies passes 3 main stages.
Stage 1 (2010 – 2014). On this stage, Big Data technologies are used as an "analytical head" of information system. Implementation of ERP and MES-systems in metallurgy had increased operational efficiency and reduced production costs, however, the possibilities of data platforms do not allow online analysis of large information flow, related to product quality and equipment reliability. Big Data solutions for Statistical Process control and equipment reliability analysis incorporates a machine learning algorithm and big data analysis. This functionality can provide such systems as SAP HANA. The input information for Big Data solution are data received from the ERP and MES systems.
Stage 2 (2012 – 2016). On this stage, Big Data technologies are used as a "basic platform" of corporate information system. The input information for Big Data systems are not only the data from ERP and MES, but mostly the data received from equipment directly using Internet of Things. Mathematical models using big data combined with the dynamic development of the Internet of Things will make it possible for industrial companies to cut costs by 5-10%. Even though machine learning plays an important role in the discovery of potential non-obvious trends lurking within their data, the main users combine these automated tools with some kind of human supervision.
Stage 3 (2015 – ). On this stage, Big Data technologies revolutionary change the industries across the product lifecycle, from R&D up to delivery. The Materials Genome Initiative (MGI) aims to greatly reduce time and cost to commercialize new materials technologies. New methods that predict fundamental materials properties have made it possible, to design materials properties in a computer. Through MGI program new experimental databases such as the Structural Materials Data Demonstration Project, Materials Data Facility, and Materials Atlas are being funded. Three databases that have received MGI support are the Materials Project (MP), AFLOWlib, and Open Quantum Materials Database (OQMD).
Selection, manufacturing, and qualification of advanced materials using Big Data technologies can increase the efficiency in each of these areas to accelerate the development of physical products. This combination only increases the need for Big Data and analytics systems to be integrated with manufacturing processes.

# SPECIFIC FEATURES OF CORRECTIONS FOR METEOROLOGICAL EFFECTS IN MATRIX DATA OF MUON HODOSCOPE URAGAN

*A. Dmitrieva[1], I. Astapov[1], V. Getmanov[2], A. Gvishiani[2], A. Kovylyaeva[1], R. Sidorov[2], A. Smirnov[2], I. Yashin[1]*

[1]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
[2]Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

andmitriyeva@mephi.ru

Muons are generated as a result of interactions of primary cosmic ray particles with air nuclei in the upper atmosphere. Muons reach the Earth surface and are capable to bring information about processes in the interplanetary magnetic field. Muon hodoscope URAGAN (MEPhI, Moscow) detects muons with high spatial and angular accuracies (1 cm and 1º, respectively) over a wide range of zenith angles (0–80º). Every minute, angular distribution of muons is recorded in a two-dimensional angular matrix M(Θ,ϕ) (Θ and ϕ are zenith and azimuth angles for matrix cell centers, ΔΘ = 1º, Δϕ = 4º), which represents a snapshot of the upper hemisphere with 1-minute exposition. One-minute matrix contains ~ 70-80 thousand events. For the analysis of muon flux variations caused by extra-atmospheric processes it is necessary to introduce corrections for meteorological effects: barometric effect (BE) and temperature effect (TE). Barometric effect is the anticorrelation of cosmic ray intensity with the pressure at the observation level. For barometric effect correction it is necessary to know the pressure at registration level. Temperature effect is caused by changes of the temperature at all altitudes of the atmosphere. For temperature effect correction it is necessary to know the temperature profile of the atmosphere. In this work different sources of this information are discussed: data from direct measurements of air temperature with help of meteorological balloon flights, vertical temperature profiles obtained from meteorological satellites and data obtained from models of atmosphere used for weather forecasting. Specific features of calculation of barometric and differential temperature coefficients are discussed. The report presents methods and results of correction for atmospheric effects in matrix data of muon hodoscope URAGAN.

# DATA PROCESSING AND IDENTIFICATION IN PROCESS MINERALOGY

*V. A. Arsentyev*[1], *A. M. Gerasimov*[1], *E.L. Kotova*[2]

[1]Research and Engineering Corporation "Mechanobr-Tekhnika"
[2]Saint-Petersburg Mining University

gerasimov_am@npk-mt.spb.ru

Process or engineering mineralogy is an independent and relatively new research area in geology and mining science, closely related to mineral processing. It analyzes big data on the composition and physical properties of minerals that determine the rational use of certain processing methods for raw minerals mined. In a number of cases, for example, in the enrichment of coal or in the production of natural building materials, machining is the final processing stage. In these cases, the rational application of process mineralogy methods and of the data obtained using these methods determine the quality of final commodity products. Studies in the field of process mineralogy imply the wide use of effective instrumental analysis methods both for the surfaces of mineral aggregates and individual minerals and for their crystalline and bulk structures. Traditional methods of optical examination, microprobe and X-ray phase analysis of the surface of minerals, electron microscopy and other research methods are supplemented by new physical methods for studying the porous structure and the microrelief of mineral samples.

Let us consider the two new areas in process mineralogy related to the study of the volumetric structure of rocks and the structure of the surface of mineral aggregates, which opened up new possibilities for the further design of mineral processing technologies and layouts. Both these areas are essentially digital technologies, requiring big data processing.

The first area is the x-ray microtomography of mineral samples, enabling quantification of pores and microcracks in samples and evaluation of their distribution by size and shape. The method is based on a layer-by-layer evaluation of the transmission of X-ray radiation through lump material with the determination of the cross-sectional area of the solid phase of microparticles. The resulting arrays of two-dimensional data are processed on a computer with the reconstruction of a three-dimensional model of the structure of the sample. Standard software for calculating the experimental data requires approximately 1 Gb of space, and the processing time on a PC ranges from 10-15 minutes to several tens of hours, depending on the structure of the sample. Data on the microporous structure of a material are most valuable for the subsequent development of size reduction (disintegration) technologies for mined rock mass and combined thermochemical technologies for the enrichment of such minerals as high-ash coal or potassium ores with a high content of clay minerals. These technologies are based on intentional modifications of the porous structure and transformation of mineral forms of raw materials processed.

The second new area in process mineralogy is the study of the fine structure of the surface relief (roughness) of mineral samples. Digital information on the relief of samples is obtained by laser microscopic scanning of the surface, rendering protocols containing information on the frequency response of peaks and depressions on the surface of minerals and their maximum and medium height distribution. Standard software for calculating the experimental data requires approximately 1.6 Gb of space, and the processing time on a PC ranges from 5-10 minutes to several hours, depending on complexity of the sample structure. Based on the digital data obtained and the image of the mineral surface provided by reflected laser beams, a digital 3D model of the object is developed. Our studies have shown that the information obtained duly explains such properties of loose mineral raw materials (important for the enrichment technology) as the coefficients of internal friction and slip, as well as the peculiarities of the electrical state of the surface of mineral particles of various shapes.

The work was carried out under the grant from the Russian Science Foundation (project No. 15-17-30015).

# THE MULTIPLE USES OF ARCHIVED DATA FROM THE GLOBAL NAVIGATIONAL SATELLITE SYSTEM REFERENCE RECEIVERS

*P. J. Cilliers*

South African National Space Agency

pjcilliers@sansa.org.za

Archived data from the international array of multi-frequency receivers of the Global Navigational Satellite System (GNSS) has been applied to a wide range of scientific fields. Whereas mobile single-frequency GNSS receivers are used mostly for navigation and asset tracking, the archived data from multi-frequency GNSS receivers located at fixed base stations permit applications that include land surveying, geodesy, monitoring of movements of the Earth's crust, time transfer, local real-time ionospheric mapping, estimation of ionospheric drift velocity, space weather prediction based on long-term trends in ionospheric total electron content, and tropospheric weather prediction based on the derivation of the total precipitable water content in the atmosphere. The data from the GNSS network of multi-frequency receivers is archived in the Receiver Independent Exchange (RINEX) file format. RINEX stores data in ASCII and binary format. The standard format facilitates data exchange from a wide range of receiver manufacturers. Most of the reference receivers log and archive the data sampled at 1s or 30 s intervals. Archived GNSS data are available on many international databases including IGS, UNAVCO, POLARNET, TRIGNET, SANDIMS and many others. A subset of GNSS receivers are designed to sample the amplitude and phase data from the GNSS satellites at high sampling rates of 20 to 50 samples per second, which permits the assessment of ionospheric scintillation, a phenomenon which degrades the accuracy of many of the other applications of GNSS data. Software tools for reading and manipulating GNSS data are freely available. The TEQC software from UNAVCO is one example of such tools that is well maintained and regularly updated to provide for increasing diversity in the signals transmitted by GNSS satellites and receivers from different manufacturers. The excellent spatial distribution of GNSS reference receivers and the high sampling rate of GNSS data, and the dissemination of the data via publically accessible databases provides a significant improvement in the accuracy, spatial distribution and frequency with which the derived geodetic, ionospheric atmospheric data can be obtained, compared to conventional means. The data collected by GNSS reference receivers is a vast resource for addressing global challenges and Data-Driven Science in the areas of Geodesy, Geophysics, Weather, Climate and Science of the Earth's upper atmosphere and ionosphere.
The paper will present an overview of the current status of the GNSS data archiving, and provide examples of the many applications of this valuable data set.

# ON RECOGNITION OF "PRIMARY DATA" PRODUCERS THROUGH DOI MINTED TO "SECONDARY DATA" DERIVED FROM THE PRIMARY DATA

*Y. Murayama*

National Institute of Information Communications Technology

murayama@nict.go.jp

Adding permanent identifiers (such as DOIs) to datasets is an important practice for data citation providing attribution to researchers and contributors who create, process, manage data (hereinafter "data producers"). This data-DOI practice is becoming increasingly popular and is gradually being extended to many more scientific communities.
In specific scientific disciplines (e.g., Earth and Planetary Sciences), studies are conducted based on data directly retrieved from experiments, observations, and/or simulation, etc. (hereinafter "primary" data or PD), while important are other studies based on data generated from compilation and/or further data processing of multiple PD products. We call this kind of data a "secondary" data (or SD). SD products are derived from processing multiple PD products, e.g., at different geographical locations, at different time, at different observational techniques/conditions and so on. In a research field where SD is more popularly used in published original papers, the current data-DOI mechanism may not be sufficient in supporting data producers to keep recognition of them and sustainability of their data creation/curation/management works of PDs. It is often difficult to ensure they receive proper recognition (and with it, enhanced reputation) even when SD products that are generated from PD products are used frequently and correctly cited in a number of research papers. In this paper, we attempt to focus on this difficulty for PD producers and we propose a revision in a metadata schema that enables the PD producers' work is appropriately recognized.
In studies of geomagnetism or terrestrial magnetism, the Dst index is typically used and analyzed as a measure of global-scale activity of geomagnetic disturbances (fluctuations). The Dst index is a dataset reduced from PD products or observed geomagnetic datasets of four worldwide observatories: Hermanus (South Africa) of the SANS Institute, Kakioka (Japan) of the Japan Meteorological Agency, Honolulu (Hawaii) and San Juan (Puerto Rico) both of the United States Geological Survey [Sugiura and Kamei, 1991]. Observation and initial data processing of the PD requires highly scientific/engineering expertise. Currently, some of the papers that cite the Dst index refer to the World Data Centre for Geomagnetism, Kyoto (the creator and provider of the index data) in their acknowledgements, or they refer to Sugiura (1964), Sugiura and Kamei (1991), or other such major contributions. However, while the number of peer-reviewed research papers using the Dst index counts 657 between 2002 and 2009, it is almost impossible for the four observatories to know the number of papers to which they have directly/indirectly contributed to.
This difficulty in calculating the observatories' citations, and thus estimating how much their observations contribute to international scientific achievement, may ultimately lead to questions about the their continued funding and operation.
In the present poster, a new method using DOIs is proposed for measuring indirect use of PD when SD is cited, whereby a DOI-RA (DOI Registration Agency) defines a metadata format (schema) for attribution information. The central idea is that the metadata of the SD-DOI should include all PD-DOIs. Then, once the SD-DOI is published, provenance of the SD can be tracked via the PD-DOIs embedded in their metadata. Search and statistical software tools for this purpose are also another indispensable key in this idea for make benefit for PD producers.

# INVESTIGATION OF GEOEFFECTIVE CMES IN 2014-2016 ACCORDING TO THE DATA OF MUON HODOSCOPE URAGAN

*[1]I. Astapov, [1]N. Osetrova, [1]A. Dmitrieva, [1]A. Kovylyaeva, [1]I. Yashin, [2]S. Bogoutdinov, [2]V. Getmanov, [2]R. Sidorov, [2]A. Soloviev*

[1]National Research Nuclear University (MEPhI)
[2]Geophysical Center of the Russian Academy of Sciences

iiastapov@mephi.ru

One of the manifestations of solar activity are the coronal mass ejections (CME). Every day in the period of increased solar activity, dozens of CMEs occur, which has a strong effect on cosmic ray fluxes in interplanetary space. Such events can have a direct impact on the Earth and near-Earth space. The report presents the results of studies of geo-efficient ejections that occurred in 2014-2016 according to the data of the ground-based muon hodoscope URAGAN, which, in addition to the integral characteristics, allows recording spatial-angular variations of the muon flux on the Earth's surface.

# INTELLIGENT GIS-BASED REMOTE SENSING DATA ANALYSIS

*[1]M. Tsvetkov, [2]F. Galiano*

[1]St. Petersburg Institute for Informatics and Automation
of the Russian Academy of Sciences (SPIIRAS),
[2]St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
Hi Tech Research and Development Office Ltd (SPIIRAS HTR&DO Ltd.)

acmm@rambler.ru

Among all the data sources used by a modern GIS, remote sensing data are of particular importance. Such data is represented in the form of a photo and, less commonly, video received by spacecraft or aircraft's sensors. Remote sensing data are characterized by a significant volume and wide variety of properties, which impose hard restrictions on the algorithms for their automatic processing.
In St. Petersburg Institute for Informatics and Automation (SPIIRAS) a universal scalable solution for GIS data processing based on the process approach was developed. This software allows to perform distributed processing of remote sensing data, integration of external algorithms in the form of plug-ins, and provides an interface for expert in a subject domain, which simplify automation of remote sensing data processing.
However, the developed solution is intended to work using communication channels with high throughput between nodes and a significant amount of computing resources available to each node, which prevents preliminary selection of the received data directly on-board. The relevance of such a filtration arises from the constraints on communication channels from the aircraft to the data receiving point.
This paper focuses on the development of a computationally effective subprocess for preliminary filtering of graphical data for Intelligent GIS. Different classes of filtering algorithms are considered. An example of processing data on aircraft, obtained by a visible-range camera in the Gulf of Finland waters is discussed.
The algorithm is based on automatic division of the analyzed remote sensing data into overlapped or not-overlapped regions, and following selection of regions to be stored and transmitted. A wide range of features can be used as the filtering basis.
Software realization of the algorithm was performed on the basis of the imagemagick package, which provides computationally effective implementation of image processing algorithms. The estimates of the algorithm accuracy are given. Dependence on speed and volume of the required RAM from the parameters of the analyzed data is investigated. It is shown that the proposed algorithm can be used for processing large volumes of data. Directions for further development of the proposed subprocess are considered.

# PORTABLE SYSTEM FOR MAGNETIC SENSORS CALIBRATION

*V. G. Petrov, A. S. Amiantov, V. A. Gorbatsevich*

Institute of Terrestrial Magnetism, Ionosphere and Radio Wave Propagation named after Nikolay Pushkov of the Russian Academy of Sciences (IZMIRAN)

vpetrov@izmiran.ru

Calibration of digital magnetometers is an important factor for obtaining qualitative magnetic data. The key metrological parameter of the magnetic sensor is the digital code to the value of the magnetic field conversion factor. In case of long-time operation of stations, metrological parameters can change, due to, first of all, to aging of electronic elements. The special calibration systems used in production and metrology centers with stable high-precision sources of current are cumbersome and expensive and are not applicable in observatory conditions. A traditional calibration system requires high-quality programmable current sources and is rather expensive.

The proposed system is based on the use of modern high-quality analog- digital converters (ADC), which have a high resolution, good linearity and are much cheaper. The system consists of a personal computer (PC), a current amplifier (CA), a calibration coil and a high-stability precision resistor (PR) connected in series with coil. The digital code set by the PC through DAC and the CA goes to the calibration coil, and the voltage at the PR is measured by an ADC converter. By measuring the voltage drop across the reference resister, we can measure the current in the coil and determine the field actually created. The accuracy of the measurement of the field turns out to be much higher than the accuracy of the field setting. If to set the coil axis parallel to the magnetic field vector and to carry out calibration of such system using an absolute magnetometer (proton or quantum), we can determine the conversion factor with an accuracy determined by the quality of the absolute instrument used (actually up to 0.01-0.001%) without knowing values of the coil constant and resistances.

The described system is realized on the basis of the Z220 ADC/DAC module, which has 2 16-bit DACs and 8 differential 24-bit ADCs. Thus, in the case of a three-coils calibration system, we can programmatically control the current in 2 coils and manually (by potentiometer) in the third and to measure the magnetic field in 3 channels with high accuracy. In the case of a mobile system, a three-component station can be calibrated by setting single coil in three positions.

The developed software provides feeding the coils of a current of a given shape and amplitude: a constant field of a given value, a meander with variable amplitude, a step-changing field, a sinusoidal field. During operation, the given and measured currents are recorded in the file with a frequency of 10 measurements per second.

On the sensor, in addition to the field specified by the coil system, an external variable magnetic field also acts. To compensate these variations, it is necessary to have an additional manetic sensor installed outside the zone of scattering of the field of the coils themselves. The simplest way to account for variations of the external field is simply to subtract the additional magnetometer data from the measured data of the calibrated sensor. In the conditions of a magnetic observatory, where the field of variations is fairly uniform, this method works quite satisfactorily. If the reference station is not available, external variations can be significantly reduced by calibrating sensor repeatedly many times and averaging the results. The software allows performing this procedure automatically.

# SEISMIC CONDITIONS OF THE ARAB WORLD AND THEIR REFLECTION IN MUSLIM ARCHITECTURE

*E. Elmanova*

National Research Moscow State University Of Civil Engineering (NRU MGSU)

eleosu@yandex.ru

One of geoecological factors that have a significant effect on the appearance of monuments of the Islamic countries, is the increased seismicity of the region. Comparing the proportions of monuments with national and European regulations on earthquake-resistant construction, we demonstrated the compliance of the seismic activity of the area. Therefore, traditional architecture was not from a random search, and under the influence of centuries of experience protecting the building from adverse natural influences.

The style of architectural monuments of Muslim peoples was formed depending on geoecological factors, in particular, under the influence of the requirement to protect the building from seismic influences. Throughout the territory where the style of Muslim architecture was formed, from Central Asia to Gibraltar, increased seismic activity is noted. Consider the famous monuments of architecture - the Syrian Umayyad Mosque (built in 708) in Damascus, Ulugbek madrasah (15th century) in Samarkand and Kalyan Mosque (XVI century). We will show that they are built in architectural proportions to ensure their seismic stability. Thanks to this, they have been preserved for centuries, withstanding many earthquakes.

These and similar ancient structures became models for later construction. They asked the main features of the Muslim architectural style.

Seismic stability of monuments is determined by comparing their architectural proportions and sizes to the requirements of modern regulatory documents [SP 14.13330.2014 and EN 1998-1]. The analysis allows us to see that the architects of the Muslim world knew the secrets of earthquake-proof construction. For example, that pointed outlines of arches, high domes, conicity of minarets, contribute to increased stability. Wooden ties perceive the expansion and reduce the stresses in the arch during the earthquake.

Balconies and upper floors, the most prone to vibration during an earthquake, should be made of light and elastic material - wood. Thus, the whole appearance of the building and the architectural style are not accidental, they are not only the creative decision of the architect. Forms, proportions and sizes of buildings are dictated by the requirements of seismic resistance, observing which the architect created his original structure.

The framework in which he was placed became a common style for Muslim countries that eventually acquired the right of a religious canon, was established and is now applied not only in seismically active, but also in aseismic areas.

# APPLICATION OF WIRELESS TECHNOLOGIES FOR DATA COLLECTION AND ANALYSIS FOR DISASTER RISK GOVERNANCE FOR SUSTAINABLE URBAN DEVELOPMENT

*K. Chaudhari1*[1], *P. J. Philip*[2]

[1]Institute For Sustainable Development and Research, ISDR, India
[2]National Institute of Technology, Kurukshetra

isdrklc@hotmail.com

As per the United Nations estimates, the world population reached 7.3 billion as of mid- 2015, implying that world has added approximately one billion people in the span of the last 12 years. Sixty per cent of the global population lives in Asia (4.4 billion),16%in Africa (1.2 billion), 10 % in Europe (738 million), 9 % in Latin America and the Caribbean (634 million), and the remaining 5 % in Northern America (358 million) & Oceania (39 million). The total population on earth is predicted to increase by more than one billion people within the next 15 years, reaching 8.5 billion in 2030, and to increase further to 9.7 billion in 2050 & 11.2 billion by 2100. Looking at the ever increasing urbanization, In 2016, an estimated 54.5 % of the world's populations inhabited in urban region. By 2030, urban areas are projected to shelter 60 % of people worldwide & one in every three people will live in cities with at least half a million inhabitants. On the basis of these figures and other global trends, it would appear that Africa and Asia will have the highest share of world's urban growth in next 25 years, resulting consideration rise of several metropolitan cities and towns along coastal region of Asia- Pacific. Therefore the task of transformation through environmental sustainability and building disaster resilient societies can be achieved through scientific data collection, management and integration for solving organizational, operational and financial management problems in urban environmental system as well as coastal ecosystem will be vital. This presentation deals with issue involved in conclave of issues related to data collection, analysis and integration for transformation towards disaster resilient societies through sustainable use of marine and coastal ecosystems in urbanized world for better urban governance in globalised world. The presentation focuses on policy formulation for collection and application of data for resources planning ,coastal risk and vulnerability, social-ecological vulnerability and disaster resilience in coastal communities, Human Pressures on Coastal Environments as well as the cost efficient technology for data collection and integration . The presentation aims to provide the platform to enable the Future Earth Knowledge-Action Network communities to work together to create new perspectives on data creation and storage emerging global environmental change science for disaster risk reduction that are relevant to Sustainable development Goals using new tools of information and communication technologies and wireless technologies. The presentation also focuses on different methods of data collection and analysis for achieving the sustainable development goals in developing and industrialized world.. The presentation examine the current status of data collection , management and recovery for disaster governance in different regions involved in e-governance process for the assessment of impacts of climate change and natural disasters along coastal regions and response of coastal communities towards transformation in the globalized world through networking with universities, research institutes and data agencies.

# IMPROVEMENT OF ACCESS TO THE GEOMAGNETIC MEASUREMENT DATA

*L. P. Zabarinskaya, N. Sergeyeva, M. Nisilevich, M. Gumeniuk, T. Krilova*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

mila@wdcb.ru

The Database named "Geomagnetism" is developing in the World Data Center for Solar-Terrestrial Physics, Moscow, Russia (WDC for STP). The goal of this product is the improvement of the big data repository system and access to it by scientists or institutions in all countries without charge. A simple, user-friendly interface for the work with online databases will be provided to all users. WDC for STP has a very extensive collection of geomagnetic data in electronic form received from different geomagnetic observations all over the world. This collection of geomagnetic data consists of digital images of magnetograms, 1-min values, hour mean values, annual mean values of the geomagnetic field components, geomagnetic indices, catalogues of magnetic storms and pulses. Data refer to the time interval from 1977 to the present In recent years, the Center is in the active process of converting historical data from paper to electronic media and then digitizing them to preserve this information for future use. Data in electronic form with the format adopted at the WDC are published on the web site of the WDC for STP.As a result, it is expected that this online Database will contain 1-min values, hour mean values, annual mean values of the geomagnetic field components and the unique long-time series of K-indices recorded at the observations of Russia and CIS countries since 1957 up till now.

Administrative interface and user's site was made for user query and extraction of data. Data type, Combinations of the three elements frequently used in geomagnetism (HDZ, XYZ and FDI), Observatory name (one or more), period of observation can be considered as parameters for user query. It is also possible to choose the Observatory name through user friendly map-based interface. The developed programme will place the information on the page of search engine to reply the request of the user in one of three possible formats: WDC, CSV or JAGA-2002. The user has the option to save the available data sets to his computer.

Written in PHP the given software uses a MySQL database server. This geomagnetic database will make collected relevant data more readily available to users.

# A SYSTEM PROTOTYPE FOR REAL TIME AUTOMATIC FRAUD DETECTION IN TEXT DATA

*N. Voinov, I. Nikiforov, P. Drobintsev, V. Kotlyarov*

Peter the Great St.Petersburg Polytechnic University

igor.nikiforovv@gmail.com

One of the most important issues which is dealt with in the modern digital world is security assurance. Usually security is understood as miscellaneous measures to provide data safety, to avoid data access by unauthorized users, to ensure reliable data storage. New technologies in software and hardware domains primarily aimed to process big data make it possible to introduce new aspects of security, one of which is detection and prevention of suspicious or unauthorized actions before they would actually perform. For example, revealing abnormal transactions in bank sector, checking adequacy of prescribed medicines and treatment to the diagnosis in hospitals or adequacy of a title of a paper or essay to its contents and so on. To implement described functionality it is required to develop a software system for automatic fraud detection based on data correctness checking. As long as data volume which shall be analyzed for the mentioned purposes is measured in petabytes in modern world and the time spent on the analysis is a very critical criteria in most cases, the system shall perform in real time mode. A prototype of such system was developed and described in this paper.
The paper considers a novel method and implementing algorithm within a prototype of a software system for automatic fraud detection of anomalies in non-formalized text data in natural language in real time. It also contains a rationale why the issue of automatic detection of outlier and suspicious data is so important. Considered are the main and most popular tools of automatic fraud detection. Conclusion is made that they all deal with data of numerical types. Proposed is a new concept of a system for automatic fraud detection in text data in real time. Mathematical model of the method of searching for correlation between texts is described. Several examples of the method usage are shown.

# NEXT GENERATION MASTER DATA MANAGEMENT – INTELLIGENT DATA MANAGEMENT IN THE ERA OF BIG DATA

*E. Okorafor*

African University of Science & Technology

ekpe.okorafor@gmail.com

There is a great deal of enthusiasm about the prospects for the Big Data held in many data platforms. We are generating data at an enormous rate and with it comes the promise of greater insights but also the daunting task of managing this data.
Managing the data is crucial to the success of the many applications, implementations, industrial and scientific data projects in the Big Data era. These projects leverage separate datasets from many sources that can be integrated then intelligently combined in new ways that permit the identification of deeper relationships between them. It is this "connectedness" that creates enormous value and insight from the data. This changing landscape that aims to promote a more complete dynamic view of data rather than a single version of truth has given rise to context and analytical data management models This paper describes the design and implementation principles for a 'new' Master Data Management in the era of Big Data. By leveraging big data and related technologies, institutions and organizations can provide an enhanced 360-degree view of critical master data entities; meanwhile, master data can help convert information gleaned from big data sources into actionable insight.

*Keywords:* Data, Master Data, Contextual Models, Analytical Models, Artificial Intelligence, Machine Learning

# THE GOETTINGEN ERESEARCH ALLIANCE – JOINT DATA SERVICES ON CAMPUS

*J. Brase*

Göttingen State and University Library

brase@sub.uni-goettingen.de

All over the world the establishing of research data services is starting to become a new task for universities and other Research Performing Organisations. Ideally this could result in a complete service portfolio for researchers through the complete research lifecycle: Support in writing proposals and data management plans, repository infrastructures for the storage of data, support in publishing data, assignment of persistent identifiers, lecturing in data management, etc.

The Göttingen eResearch Alliance (eRA) is a university initiative by two local research-oriented information-infrastructure providers, the Göttingen State and University Library and the Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen mbH (computing and IT competence center) to establish an institutional support structure for research data management and eResearch. The eRA offers a new integrated set of services for digital research infrastructures. The focus of the four-year build-up phase was on services supporting professional research data management (RDM) along the research data lifecycle, such as generation of data management plans, sustainable storage and accessibility of data, as well as publication of data. Underpinning these offers is a university-wide research data policy, which has been adopted by the university Senate and presidential board, as well as the faculty council and the management board of the University Medical Center Göttingen (UMG). The offered services address all faculties and research projects on the Göttingen Campus (GC). The services will furthermore be adapted to discipline or project specific needs, in close collaboration between the researchers and the eRA. Beyond the area of RDM, the eRA will successively add digital tools and services to its portfolio.

The eRA was established in 2014 as part of the university strategy to advance the (digital) research infrastructure of the GC. It is funded initially through the presidential board of the university for a build-up phase of four years. A multidisciplinary team of 5 full-time equivalents is engaged - in this cross-cutting, horizontal approach - in coordinating campus-led activities to establish sustainable and innovative RDM services for the support of researchers in all phases of the research lifecycle. Part of this endeavor is fostering the formation of a community on data management and data science across the campus and across the disciplines. An emphasis is on the cooperation of the information infrastructure providers and the Research Department (central administration) of the university, allowing the eRA to become the central contact for all researchers.

In particular, the eRA's working program contains four areas of activities in the context of the goals mentioned above, i.e., consulting, training, connecting, and developing. In this talk we will give an overview on the structure and the services of the eRa and discuss our expertise and next steps to take

# A TECHNOLOGY FOR FILTERING GEOMAGNETIC FIELD OBSERVATIONS USING LOCAL APPROXIMATION MODELS

*V. G. Getmanov,A. D. Gvishiani, R. V. Sidorov*

Geophysical Center of Russian Academy of Science

v.getmanov@gcras.ru

A filtering technique based on the use of local approximation models for the geomagnetic field observation data, complicated by such features as significant nonstationarity, the presence of many frequency components, possible small amplitudes and noise, discontinuities of derivatives, short observation intervals, is developed. Due to these features, the use of standard filtering methods is not effective enough.
A basic statement is proposed for this filtering technology, formulated for one-component observations, consisting in constructing local approximation models on local intervals as nonlinear functions with a small number of parameters, specifying local functionals, and solving the sequence of unrelated problems of conditional minimization of these functionals. A generalization of the basic formulation of filtering technology is proposed, which takes into account the connections of approximation models at the docking points of local intervals. The problem of conditional minimization for filtering observations in the form of a spline model is solved. Further generalization of the basic formulation for multicomponent geomagnetic observations was made taking into account the relationships between the components. A set of conditions was created-the model links for the various channels and the corresponding functionals. The problem of conditional minimization for obtaining the result of filtration is solved. The examples of implementation of the filtering technology on model and experimental data are presented.

# RESEARCH DATA MANAGEMENT IN NATIONAL AGRICULTURAL RESEARCH SYSTEMS OF BANGLADESH AND INDIA

*S. Das*[1], *S. Gutam*[2]

[1]Bangladesh Agricultural Research Council
[2]ICAR-Indian Institute of Horticultural Research

susmitabarc@gmail.com

Joining the global efforts for open data, the national governments of Bangladesh and India are building necessary infrastructure and policies for opening up of the government data (data.gov.bd and data.gov.in). And following the global advocacy efforts of Global Open Data for Agriculture and Nutrition (GODAN), regionally, the National Agricultural Research Systems (NARS) of Bangladesh and India are making efforts for the establishment of data repositories for agriculture. While the India's NARS had established krishi.icar.gov.in, there is no dedicated repository for the Bangladesh's NARS. As the evidence based science is gaining importance along with reproducible research, opening up of research data would become necessary in all sciences. And with respect to agriculture would be more necessary as many of the Sustainable Development Goals (SDGs) are related to agriculture. The data published by the NARS in both the countries (Bangladesh and India) in terms of articles is growing exceptionally and the research managers should put in place the workflows to extract, curate and make it publicly available and accessible. Our paper presents the overview of the Research Data Management (RDM) in both the NARS of Bangladesh and India and the perceptions of the researchers towards openly sharing the data (published as well as unpublished) and the policies RDM policies for data cycle, rewarding system etc. The paper also talks about the way forward for both the systems to reach SDGs through open research data in agriculture.

# INVESTIGATION OF EFFECTS OF CORONAL MASS EJECTION ON IONOSPHERIC TOTAL ELECTRON CONTENT OVER SOUTH EASTERN, NIGERIA

*E. Hanson-Eduok, K. Okpala, F. N. Okeke*

University Of Nigeria Nsukka

hansonestty@yahoo.com

In this work, we attempt to investigate the contributions and effects of Coronal Mass Ejections (CMEs) on Total Electron Content (TEC) in the ionosphere of an equatorial station, Nsukka (Lat. 6.860N; Long 7.380E). Using TEC data estimated by the GPS database of the United States Air Force Research Laboratory, and CME data from the Solar and Heliospheric Observatory (SOHO) satellite, we calculated the variation of TEC in the solar maximum year 2012. From the analyses, regular Gaussian distribution of TEC was observed during geomagnetic solar quiet (Sq) days with TEC profiles characterized by normal regular variations with a peak of 42 TECU at 15:00 hours and average minimal value of 17 TEC at both pre-noon and nighttime hours resulting from photo-ionization and recombination. Under solar magnetically disturbed periods associated with CME events, TEC variations assumed very sporadic patterns; maximizing quite early Under disturbed conditions, TEC amplitudes maximized to 68 TECU at 14:00 hours – a significant forcing of CME on TEC; an indication of longitudinal drift of charged particles towards the ring current. A peculiar scenario was observed for days described as neither Sq nor disturbed, where TEC signatures assumed the pattern observed under Sq condition. Statistical analysis carried out revealed that CME speed and TEC enhancements are weakly correlated with a coefficient of 0.2. The weak correlation coefficient implies that the speed of CME might not necessarily account for the enhancement of TEC.

# DATA ARTICLES AS A MEANS TO IMPLEMENT FAIR – PROS AND CONS!

*A. Narayana, V. Dantuluri*

Profeza

aadi@profeza.com

Nature's Scientific Data aims to promote wider data sharing and reuse, as well as credit those that share their data and is open to submissions from a wide range of areas in the natural, clinical and social sciences – including descriptions and analysis of big and small data, from major consortiums, single labs and individuals. Scientific Data enables researcher's to receive credit for your research data, whether or not it is associated with a research article. As a means of promoting openness and open-data 'Scientific data' aims in implementing FAIR guidelines by means of encouraging authors to submit data description articles which expects to enhance reusability of data apart from the conventional research articles. But, this approach or initiative surely have a limitation in terms of Economics, which might have been less thought of. This can be well understood with the help of a Hypothetical Case Study;1. Every Research article which has data in it - have a scope that the author can publish a data article with data description on how this data can be analysed. So instead of a single publication, Author can have two publications - one being a conventional research article, other being a data article. 2. Assuming for once, that out of 2.5mn research articles published every year, there comes a mere interest of those authors interested in publishing Data articles, thus 2.5mn additional data articles. This might definitely happen in the due course if data articles become more prevalent, as authors for sure will receive more number of publications instead of one(from the same research study), and for publish or perish still being prevalent there is no roadblock for this case to happen in the near future. 3. Considering that this happens(Point 2) there is an outburst in the research information i.e being disseminated.a. who is going to support the publication costs of those outburst event. Doesn't this going to disturb the Economics of research, and it's allies that support the grants for conducting and communicating research?b. Does that mean funding bodies and governments needed to increase the Budget for research grants to sustain this dissemination or is this really advisable or approachable or is there a way that this can be taken care off? This session compiles these Economic pitfalls and other's which include meta-data pollution in the repositories etc because of the outburst that might happen. All of these were less talked about, and which are a MUST to re-think and reevaluate on implementing new strategies which can successfully implement and foster FAIR DATA guidelines into the business as usual workflows of different stakeholders involved in the process and prime a sustainable open-data momentum.

# EXPLORING NEXT GENERATION RESEARCH DATA SERVICES AND DATA ROLES BEYOND 2017

*L. Lyon*

University of Pittsburgh

elyon@pitt.edu

This futures paper seeks to identify and examine selected 'next generation data zones' and the research data services and data roles associated with their implementation within universities, research institutes and data centers. Three data zones will be critically reviewed and framed in five contextual dimensions: i) current scholarly dilemmas, debates and data exemplars, ii) assessing the pragmatic risks for the scientific community, iii) current activity and prospective research data service development, iv) emergent and novel data roles v) future challenges and issues to be addressed. The critical commentary will draw upon and synthesize evidence from published research (including ongoing research by the author), recent research workshop findings, practitioner studies, and media / news coverage. The scope of the three next gen data zones is briefly described below:

1) *Rescue, Re-Use and Return-on-Investment:* scope will include multi-disciplinary examples of data loss and data at risk including the successful recovery of Arctic and Antarctic ice-sheet data by the National Snow and Ice Data Center, US government climate data at risk and unseen Andy Warhol images recovered by forensic retro-computing at Carnegie Mellon University, data rescue (aka 'guerrilla archivist') efforts in Pittsburgh, proposed new roles such as Data Triage Scientist, the development of data tools and services such as forensic data recovery using BitCurator, and perceived issues and challenges, such as developing accepted global standards for data triage workflows.

2) *Transparency, Truth and Trust*: scope will include high-profile examples of data falsification and fabrication from scholars in the biosciences and medicine, subsequent data publication retractions and institutional responses, multidisciplinary researcher perceptions of the key concept of transparency and its application throughout the research data lifecycle (based on current research by the author), emergent new roles such as Reproducibility Librarian and associated data validation / verification services, plus wider implications for assuring research integrity, continuing public trust and the long-term value and sustainability of open science practices.

3) *Beyond Data Analytics*: scope will include current practice around data science in libraries (based on the findings from a recent IMLS-funded workshop held in Pittsburgh), the exciting prospective application of cognitive computing and machine learning to a broader disciplinary range of data-driven research, new sub-disciplines and emerging institutional data roles such as Decision Engineer, associated challenges such as the coordinated training of information and data professionals plus the development of an educational curriculum which encompasses the very rich blend of skills required for these new roles, and which will contribute to bridging the well-documented workforce data talent gap.

The paper will close by summarizing some of the broader management and leadership challenges for the research community and the opportunities associated with implementing these types of innovative research data service, including re-engineering the delivery of research data service models in academic libraries from remotely based or hybrid operations to immersive data informatics team science, and addressing current workforce capacity and capability issues, such as the effective recruitment and retention of data professionals and the agile re-positioning of the iSchool graduate curriculum to catalyze translational data science.

Note: This new paper/abstract is based on the author's invited Opening Keynote presentation at JCDL2017, Toronto, June 2017.

# DATA2PAPER: GIVING RESEARCHERS CREDIT FOR THEIR DATA

*F. Murphy*

MMC Ltd

fionalm27@gmail.com

Part of the Jisc Data Spring Initiative, a team of stakeholders (publishers, data repository managers, coders) has developed a simple 'one-click' process for submitting data papers related to material in a DataCite/ORCID compliant repository. Data papers cover methodological detail that is not otherwise captured and published in traditional journal articles and/or dataset metadata. As such, it can improve the findability and reusability of the underlying dataset. DataCite and ORCID information is transferred from a data repository via a SWORD-based API to a cloud-based helper app based on the Fedora/Samvera platform. In the app, the text of the data paper is combined with existing metadata drawn from DataCite and ORCID to generate a package suitable for submission into a journal submission platform without further user interaction. By reusing metadata from ORCID and DataCite that has already been previously entered/curated, the process is both simplified and made less error prone. Funders are becoming more interested in good data management practice, and institutions are developing repositories to hold the data outputs of their researchers, reducing the individual burden of data archiving. However, to date, only a subset of the data produced is associated with publications and thus reliably archived and consequently made available for sharing and reuse. This represents a loss of knowledge, leading to the repetition of research (especially in the case of negative observations) and wastes resources. Without data papers as a valid outcome, it is laborious and hard to justify for time-poor researchers to archive and fully describe their data and thereby maximize its utility to others. Equally, filling out diverse submission forms which require repetition of already entered data for journal(s), makes things even lengthier. The app makes the process of associating and publishing data with a detailed description easier, with corresponding citation potential and credit benefits.
The presentation will discuss the history of the project, including the results of an initial feasibility study, along with a demonstration of the current prototype. We will outline the current work being done to turn the prototype into an operating service with a sustainable business model. We will also consider how the service might develop in the future in conjunction with activities such as Scholix, various RDA areas of activity, such as Data Journals Publishing Policy, Credit and Attribution, and Exposing Data Management Plans, and other ongoing expressions of interest and consultations on issues of impact, reproducibility, FAIR Data, persistent identifiers and new metrics by various national and international bodies.

# CASE STUDY ON DISASTER RISK MANAGEMENT IN THE LAO PEOPLE'S DEMOCRATIC REPUBLIC

*S. Boupha*

Ministry of Science and Technology

silapboupha@yahoo.com

The country is prone to large scale natural hazards such as floods, droughts, and storms. These often trigger secondary hazards such as landslides, fires, infestations and outbreaks of animal diseases. Each year, it can lead to massive damages and losses of lives, livelihoods and infrastructure. These hazards are likely to increase in frequency and intensity. This creates an additional challenge for reaching the economic and social development goals.
The country has recognized the need for reducing the underlying risks to counterbalance hazard and disaster impacts by integrating the science and technology and space base and geospatial information in response to the National Disaster Risk Strategic Plan 2003-2020, Vision 2030 and National Strategy for Social Economic Development 10 years (2016-2025) and 7th NSEDP 2011-2015 and 8th National Social Economic Development Plan 2016-2020. These plans and vision are in line with Sustainable Development Goals (SDGs).
Based on the high commitment to promote the shift towards proactive disaster risk reduction and management (DRRM), this paper describes the needs through synergy in space with a strong focus on the poverty reduction, food, drought and nutrition security for sustainable agricultural development. Therefore, there is a need to encourage the stakeholders to offer regular capacity building and programmes focused on space technology applications in disaster management and climate change and to develop technological capacity.
This case study on Disaster Risk Management in Lao PDR describes the impact affected by major hazards/disaster causing the impact on housing, agriculture, people life, economic losses and disaster frequency. Also, it provides the statistic on affected population and economic losses and the short description on National Disaster Risk Strategic plan 2003 up to 2020 and Disaster Risk Management Activities undertaken and the regional cooperation with ASEAN. It highlights inter-contingency plan with all stakeholders. Keyword: Disaster Risk Management, Impact and Economic Losses, Contingency Plan and National Disaster Risk Strategic Plan up to 2020.

# BACKGROUND SEISMICITY DATA PROCESSING AIMED AT STRONG EARTHQUAKE-PRONE AREAS DETERMINATION

*A. Gvishiani,B. Dzeboev, N. Sergeeva, I. Belov, A. Rybkina, O. Samokhina*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

a.gvishiani@gcras.ru

Recognition of earthquake-prone areas has matured as a branch of mathematical geophysics in the works by I. M. Gel'fand, V.I. Keilis_Borok, Sh. A. Guberman, E. Ya. Rantsman, et al. at the beginning of the 1970s. In these works, the morphostructural nodes, within which the epicenters of the earthquakes with $M \geq 6.5$ may occur, were identified for the first time in the Pamir and Tien Shan by the EPA (Earthquake-Prone Area recognition) method. In subsequent years, EPA approach for recognition of potentially seismically active zones in mountainous regions has been further developed. The EPA method provided a significant contribution to the seismic hazard assessment and seismic zoning of the regions of high and moderate seismicity. This determines the practical value of EPA for estimating seismic risk. At the same time, specifying the target objects for recognizing and determining their geological and geophysical parameters in the EPA method is a big independent and challenging problem. Its solution requires the data that are not always available for the region of study. The works on constructing the scheme of morphostructural zoning and measuring the relevant parameters are conducted by highly skilled experts, largely manually, which reduces the applicability of the method.

The FCAZ (Formalized Clustering and Zoning) algorithmic system, which is an alternative to the EPA, has been developed in the Geophysical Center of the Russian Academy of Sciences. It is a superposition of applications of the DPS and $E^2XT$ algorithms, and intended for identifying the areas where the strong earthquakes may occur. In contrast to EPA, FCAZ uses neither morphostructural zoning nor the dichotomy algorithms with learning, which compose the algorithmic core of the EPA method. The algorithmic core of the FCAZ system, suggested here, is the DPS algorithm of objective classification. This system uses solely the information on the epicenters of the earthquakes in the region.

The Authors performed the FCAZ-recognition of strongest, strong, and significant earthquake-prone areas for the Andes mountain belt ($M \geq 7¾$), California ($M \geq 6½$), Caucasus ($M \geq 5.0$), Crimean Peninsula and North-West Caucasus ($M \geq 4½$), and the regions of Altai-Saiyans ($M \geq 5½$) and Pribaikalia-Transbaikalia ($M \geq 5½$).

In the course of recognition of potentially highly-seismic territories in the region of Pribaikalia-Transbaikalia, the authors discovered a unique capability of the FCAZ-system. It allows conducting successive recognition of earthquake-prone areas for sequentially increasing magnitude thresholds $M_0^1 < M_0^2 < \ldots < M_0^S$ within the studied region. The classical EPA-approaches allow to perform recognition solely for a determined threshold $M_0$. In this case, the increase of $M_0$ requires to start the recognition from the beginning. Moreover, it is not obvious that the recognized zones would be inside the ones, recognized for lower magnitudes.

The reliability of the recognition results is confirmed by the consistency of the recognized FCAZ-zones with epicenters of instrumental and historical earthquakes, the results of control experiments "complete seismic history" and "individual seismic history", and comparison with the results of recognition using randomly generated earthquake catalogs.

FCAZ is a new efficient systems analysis tool for recognition of the zones prone to occurrence of moderate, strong and strongest earthquake epicenters in the study regions under consideration. The FCAZ possesses artificial intelligence elements. This algorithmic system is under development by integration of systems analysis and pattern recognition techniques.

This research is supported by the Russian Science Foundation (project No. 15-17-30020).

# ALGORITHM BARRIER WITH SINGLE LEARNING CLASS FOR STRONG EARTHQUAKE-PRONE AREAS RECOGNITION

*A. Gvishiani, S. Agayan, B. Dzeboev, I. Belov*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

i.belov@gcras.ru

In order to partition the territory into two non-crossing zones, i.e., where epicenters of strong ($M \geq M_0$) earthquakes are possible and where they are not, the dichotomous algorithms with learning are used. As early as 1972 the "Kora" recognition algorithm with two classes learning was successfully applied for highly seismic zones in the regions of the Pamir and Tian Shan. The approach, later called EPA (Earthquake-Prone Area recognition), was used successfully to identify the locations of possible epicenters of strongest, strong, and significant earthquakes in various mountain regions.

Resulting from recognition by the mentioned algorithms, the study area (in fact, the finite set of recognition objects) is subdivided algorithmically into two disjoint parts: B, where epicenters of strong earthquakes are possible, and H, where these earthquakes are impossible. The learning set consists of two disjoint subsets: $B_0$ corresponding to learning objects of high seismic class B and $H_0$ corresponding to learning objects of low-seismic class H. It is easily seen that objects from $B_0$ are all recognition objects, in the vicinities of which epicenters of strong ($M \geq M_0$) earthquakes were reported. Subset $H_0$ is filled with either all the remaining recognition objects or objects, in the vicinities of which only epicenters of earthquakes with $M < M_0-\delta$, $\delta > 0$ were reported. Obviously, that learning sets $B_0$ and $H_0$ are not equivalent, because class $H_0$ is not a "pure" learning material.

Since the development of the EPA method, the problem regarding the reliability of the results obtained by means of mentioned algorithms, implying that learning is performed on potentially overlapping learning sets of different qualities, has been debated. In our communication, a new Barrier dichotomy algorithm is designed; learning in this algorithm is performed on one "pure" class $B_0$.In the Barrier algorithm, learning is performed on the sole highly seismic "pure" learning class $B_0$, whereas learning of the low seismic class is absent. The algorithm solves the problem of constructing a subset close to the only learning class in the $B_0$ initial finite set of alternatives on the basis of the set of scalar features. For this purpose, we construct the measure of difference between arbitrary alternatives built on each attribute. The idea of constructing a measure implies that we reveal and quantitatively estimate the "barrier" impeding the closeness of alternatives in the given attribute. The measures of the barriers play the role of metrics on the initial set. This allows us to impart an exact meaning to the term of proximity to the set subset on the basis of a number of attributes.

The Barrier algorithm is used for recognition of earthquake-prone areas with $M \geq 6.0$ in the Caucasus region and in the Crimean Peninsula. Comparative analysis of the Crust and Barrier algorithms justifies their productive coherence. This research is supported by the Russian Science Foundation (project No. 15-17-30020).

# DISCRETE MATHEMATICAL ANALYSIS AND ITS APPLICATION FOR MONITORING OF SEISMIC ACTIVITY

*B. Dzeboev, S. Agayan, I. Belov, R. Krasnoperov*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

b.dzeboev@gcras.ru

Seismic hazard assessment requires continuous geophysical monitoring in general and, above of all, seismic monitoring. In this respect, development of methods of seismic activity monitoring on the basis of algorithms of Discrete Mathematical Analysis (DMA) is undoubtedly a topical problem. DMA is a new approach to data analysis developed at the Geophysical Center of the Russian Academy of Sciences.

The developed DMA monitoring of seismic activity of a given territory implies the use of DMA methods, which earlier proved to be satisfying, for analyzing earthquake catalogs. The seismic process is analyzed by investigating its behavior at the nodes of a coordinate mesh with a set step (in reference points) and by constructing the measures of activity. As a measure of activity, we will use the parameter μ, value of which is found by a topological filtering (clustering) of the Discrete Perfect Sets (DPS) algorithm. Measure changes within the limits of [–1, 1]. In contrast to the classical measure of seismic activity, measure μ reflects the time-variant relative density of epicenters compared to the surrounding space. A local increase of measure μ with time can reflect the increase in activity of weak earthquakes that often accompany the final stage of preparation of a strong earthquake. Thus, temporal variations in measure can be useful for detecting the periods of higher seismic hazard and for estimation of the location of an earthquake under preparation. In terms of the present work, seismic activity monitoring means analysis and investigation of which behavior is demonstrated by the set of measure μ time series at the reference points. At each step, time interval ti of duration is considered.

Since the seismic process in a certain spatial neighborhood is irregular in time, whereas changes in the level of its activity, which often indicates an increase or decrease in the potential seismic risk, take place over certain time interval, the assessment of seismic activity at time moment it should take the memory of seismic activity on time into account. For this purpose, at each moment of time ti, the value of measure μ is recalculated for each node of the mesh, with the memory being taken into account in the form of power averaging with a weighting factor.

We tested the approach developed to seismic activity monitoring for the territory of California, Caucasus and Kamchatka. The non-randomness of the results is shown by means of an error chart.

This work was performed as a part of the grant of the President of Russian Federation for a state support of young Russian scientists – PhD (project number MK-4555.2016.5).

# ALGORITHMIC SYSTEM FCAZ AND STRONG EARTHQUAKE-PRONE AREAS IN THE RUSSIAN FAR EAST

*B. Dzeboev, R. Krasnoperov, S. Agayan, I. Belov, E. Vavilin*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

r.krasnoperov@gcras.ru

Recognition of potential seismically dangerous zones using morphostructural zoning (method EPA) appears to become a classical problem already in early seventies. Well-known geophysicists and mathematicians in USA, USSR, Italy and France provided contribution to these studies. However, to encircle potentially highly seismically dangerous zones remains one of the most difficult and, by far, not completely solved seismic hazard problems.

Let's outline the problem for a given seismic region. $M_0$ indicates the magnitude starting from which the earthquakes are declared to be strong and the places of possible origin of their epicenters should be recognized. Let W be the set of the recognition objects $W = \{w\}$. Formalized Clustering And Zoning (FCAZ) system provides decomposition $W = S \cup N$, where strong earthquakes epicenters coincide with and only with $w \in S$ in the past, nowadays and future. Consequently, FCAZ recognition $W = S \cup N$ should be in good coincidence with known locations ($M \geq M_0$) of strong earthquake epicenters.

Original algorithmic system FCAZ (Fuzzy Clustering And Zoning) was created by the authors in 2013. It allows to carry out recognition of possible strong earthquake epicenters occurrence based on the whole set of the earthquake epicenters used as recognition objects. FCAZ system consists of two major parts: DPS algorithm and algorithm $E^2XT$. DPS (Discrete Perfect Sets) is an original clustering algorithm. This fuzzy-logic based algorithm carries out recognition of the objects topological filtration for depriving the so called lower d-density objects. Using the algorithm $E^2XT$, FCAZ proceeds with 2D continuous generalization of the DPS zero-measure clusters, obtained by DPS.

Recognition of strong earthquake-prone 2D and 3D areas (recognized objects – the earthquake epicenters and earthquake hypocenters, respectively) was conducted for the territory of regions of the Russian Far East. The reliability of the recognition results is confirmed by the consistency of the recognized FCAZ-zones with epicenters of instrumental and historical earthquakes, the results of control experiments "complete seismic history" and "individual seismic history". The recognized 2D and 3D FCAZ-zones were compared with each other. They were compared with the recognition results of the classical method EPA (Earthquake-Prone Areas recognition). The work was conducted within the framework of the Russian Foundation for Basic Research (Project No. 16-35-00603 mol_a).

# THE NOAA GOES-16 SATELLITE SPACE WEATHER DATA AND PRODUCTS

*P. Lotoaniu*

University of Colorado

paul.lotoaniu@noaa.gov

Since their inception in the 1970s, the NOAA GOES satellites have monitored the sources of space weather on the sun and the effects of space weather at Earth. The GOES-16 spacecraft, the first of four satellites as part of the GOES-R spacecraft series mission, was launched in November 2016. The space weather instruments on GOES-16 will image the sun's atmosphere in extreme-ultraviolet and monitor solar irradiance in X-rays and UV, solar energetic particles, magnetospheric energetic particles, galactic cosmic rays, and the Earth's magnetic field. These measurements are important for providing alerts and warnings to many worldwide customers, including the NOAA National Weather Service, satellite operators, the power utilities, and NASA's human activities in space. This presentation reviews the GOES-16 space weather instrument data including initial post launch results along with a discussion of data calibration/validation activities. We also describe the space weather Level 2+ products that are being developed for the GOES-R series including solar thematic maps, automated magnetopause crossing detection and spacecraft charging estimates. Finally, we discuss data access both in terms of real-time and archiving plans for the space weather data products. These new and continuing data products will be an integral part of NOAA space weather operations in the GOES-R era.

# A DATA CITATION SYSTEM FRAMEWORK FOR IDENTIFICATION OF EVOLVING DATA

*K. Zettsu, Y. Murayama*
National Institute of Information and Communications Technology
zettsu@nict.go.jp

Recent advances of IoT encourage data scientists to analyze temporally-evolving data from a variety of sensor networks, smart services or social networks in order to solve complex problems in urban environments. It is an emerging issue for open data communities to provide access to more dynamic data (like data from IoT) together with conventional archives of static data. However, identification and citability of ever-growing IoT data such as sensing data is a crucial problem. We try to solve this problem by enhancing conventional data citation framework to handle the evolving data.

In our framework, a sensing data is continuously collected and stored in a relational-database table. Similar as conventional data citation, a DOI-like unique identifier is assigned to an arbitrary subset of the continuous data specified by a query to the table. Instead of embedding a query directly into the identifier as in many existing approaches, our framework creates a "view" of the table and assigns an identifier to the view. The target dataset is dynamically generated from the original table by executing the view, thus a view is regarded as a virtual dataset. Versioning the view definition using a version control system allows to access to an appropriate subset of evolving data with a corresponding version number. The system internally manages data identification using the name and version number of a view. For example, an identifier is in a form similar to the existing DOI like "10.234/evwh/rain_panda_000001". It shows that "10.243 " is a prefix (like prefix of DOI; namespace and registrant ID). "evwh" and "rain_panda_00001" are a DB name and a view name with version number respectively. Externally this string can be used as a unique identification of a view, then it can be used as a standard DOI to specify the dynamic data.

Based on this simple but robust identification mechanism for such evolving data, we have implemented a prototype of our framework by extending conventional data citation system. A resolver of the identifier (handle server) is extended to extract and forward a request for data (i.e., DB name, view name and view version) to the target repository. The request is processed on the DB server at the target repository then generate a landing page containing a hyperlink to the target data as well as DataCite metadata. (Figure 1). Web APIs for direct access from an application program are also provided.

The prototype is applied to our NICT Event Data Warehouse system for a feasibility test based on practical use cases. Heterogeneous sensing data from remote sensing radars, mobile sensors, Web sites, SNS etc. are continuously gathered into the Event Data Warehouse for natural disaster risk analysis. The proposed framework is subject to management of datasets based on disaster cases for facilitating multiple accesses to a specific part of sensing data from different analysis processes. The data citation mechanism contributes to improvement of accessibility, reproducibility and traceability of ever-growing data. Our framework realizes large part of recommendations of RDA Working Group on Data Citation[1] including data versioning (R1) , query store (R3), query uniqueness (R4) query timestamping (R7) , query PID (R8), store query (R8), landing page (R11), machine actionability (R12). In our presentation, more detail about the implementation is explained and conformance to the RDA recommendations is discussed based on the practical use cases.

(See the image at "https://www.dropbox.com/s/qlq5f4xdefnrgxa/ldpg.png?dl=0", if it is not shown here correctly)Figure 1: An example of landing page for data citation of weather sensing data.

References

1. Rauber, A., et. al.: Data Citation of Evolving Data- Recommendations of the Working Group on Data Citation, Research Data Alliance (October 2015). Available: https://www.rd-alliance.org/system/files/RDA-DC-Recommendations_151020.pdf [Accessed 25 July 2017]

# HANDWRITING-BASED ANALYSIS FOR ALZHEIMER'S DISEASE

*G. Pirlo*

University of Bari

giuseppe.pirlo@uniba.it

Alzheimer's Disease (AD) is one of the major neurodegenerative pathologies affecting millions of people in the world and whose growth rate is likely to increase in the developed countries, along with the rising of living expectation. AD is a pathology with enormous social and economic implications, therefore research is even more committed to finding innovative, low-cost and non-invasive solutions for early diagnosis, monitoring and treatment. A recent direction of research starts from the large body of knowledge about the neural processes occurring in the brain areas related to fine motor abilities. In particular, significant changes of the handwriting performances are a prominent feature of AD, as handwriting is the result of a complex process that concerns cognitive, kinesthetic, and perceptual-motor abilities.
This paper provides a brief overview on the use of handwriting for early AD diagnosis and monitoring, considering both signal processing and writing generation models approaches. New methodologies for the analysis of handwriting will be discussed, based on innovative signal processing approaches for the analysis of local stability and complexity of handwriting. The approaches use Dynamic Time Warping to perform a multiple matching procedure among handwriting-generated signals to detect a local degree of stability and complexity.
The effectiveness of specific writing generation models based on the kinematics theory of rapid human movements, will be discussed for AD assessment. Specifically, the Delta-Lognormal and Sigma-Lognormal models will be used to investigate on the handwriting generation processes, in order to quantify differences in fine hand motor function in AD patients and healthy people.
Finally, some considerations about further directions of research will be provided, also with respect to the need to define standard methodologies and tools for supporting the research community working in the field.

# SPATIAL DATA INFRASTRUCTURE FOR MULTIDISCIPLINARY GEOGRAPHICAL SCIENTIFIC RESEARCHES.

*T. Khromova*

Institute of Geography of Russian Academy of Science

tkhromova@gmail.com

The diversity and complexity of scientific data pose challenges in how it is managed and shared with a diverse community, visualized intuitively, and fused for answering scientific questions. Such challenges are very actual for multidisciplinary scientific researches in Institute of Geography RAS including cryosphere studies. Spatial data infrastructure is created to build a framework that will support satellite imagery, airborne data, climate-weather observations, simulations, and forecasts. Such frameworks allow us to manage data, generate science products, and develop applications that support the end users need. The concept of the data system identifies solutions aimed at assessing the completeness and qualitative characteristics of existing spatial data, exploring the territories in various ways, planning studies to collect additional data and integrating them to work with scientific hypotheses, setting up new research projects and fixing their results, engaging researchers and organizations of Russia and foreign countries. The level of metadata describing the layers of spatial data, combined GIS-projects, as well as cartographic images is allocated as the upper level. Data can be in different projections, coordinate systems, sometimes generally in the form of schemes, but nevertheless have a value for scientific work. Based on them, one can draw conclusions about the study of the territory, justify the execution of works on the collection of new data, generate and refine scientific hypotheses, work with data as evidence, see unexplored places in some respects, i.e. Work with "bad data".

# SEISMIC SAFETY AND APPROACHES FOR ITS IMPLEMENTATION

*J. Karapetyan*

Institute of Geophysics and Engineering Seismology after A. Nazarov

jon_iges@mail.ru

The study of the seismic regime and assessment of the seismic hazard and risk of large cities, responsible energy and hydraulic structures, monitoring and forecasting of earthquakes, is an extremely important scientific and applied task. It has both socio-economic and strategic importance, since it is associated with the preservation of human lives and large amount of material values.

One of the main directions for solving this problem is the "Provision of complex safety and security of the territory and infrastructure objects of critically Important objects (CIO) from natural and anthropogenic influences."

Based on the capabilities of information technology and microelectronics, the Institute of Geophysics and Engineering Seismology named after A. Nazarov of the National Academy of Sciences of the Republic of Armenia (IGES NAS RA) has developed fundamentally new series of instruments that allow geomonitoring of the territory and infrastructure of the CIO, to monitor the perimeter security and the boundaries of the CIO, digital quantitative data on the level of possible seismic impacts and their application in developing measures to reduce seismic risk and ensure their seismic protection.

Together with the Geophysical Center of the Russian Academy of Sciences in the field of software and hardware, it is planned to develop a software package that collects, accumulates, processes, identifies information from various natural and man-made sources, also elaborations for various hierarchical observing systems, elaboration of the most acceptable types of information transfer to the central Point of complex processing of the CIO (communication line, radiotelemetry, wireless communication, space communication capabilities) and recognizing objects of Natural hazard.

In the future, together with the Geophysical Center of the Russian Academy of Sciences, it is planned to create a magnetometric observatory INTERMAGNET in order to geomonitor the territory of Northern Armenia in the observational experimental-methodical seismic-prognostic base Gyulagarak IGES NAS RA.

# INTELLECTUAL INFORMATION PLATFORM FOR CREATION OF NEW APPLICATIONS AND SERVICES BASED ON INTEGRATED USE OF ERS AND ANCILLARY DATA

*V. Zelentsov, R. Yusupov, S. Potryasaev*

St. Petersburg Institute for Informatics and Automation
of the Russian Academy of Sciences (SPIIRAS)

v.a.zelentsov@gmail.com

Earth remote sensing (ERS) data are becoming more and more called-up for monitoring of natural and natural-technological objects and systems (NNTO's). Further prospects of their active use are associated with shifting from pretty simple monitoring tasks to forecasting of NNTO's status, proactive management of them and decision making support in usual conditions and in case of emergencies.

By today, there are multiple available sources of remote sensing data with different levels of processing. At the same time, there is a lack of applications, services and products to decision making support for NNTO's management, which would be adapted for a user who does not have special skills in information technologies and ERS data processing. This gap is a significant obstacle to increased use of accumulated and steadily growing volumes of ERS data. To handle this problem new level information systems are needed. These systems are to be aimed at integration in automatic mode ERS and other ancillary data (both spatial and non spatial) with models of NNTO's status forecasting and decision making support, as well as selection of necessary data sources and models should be performed. Elaboration of the simplest user interface is no less important task.

A theoretical fundament for creation of such a system is given by models qualimetry methods which have been developing in the SPIIRAS for recent years, and now they allow embodying the intellectual multi-criteria choice of models and setting their parameters without operator intervention. With that, contemporary information technologies provide new opportunities for designing of distributed systems with heterogeneous information sources combining.

Use of these two directions achievements led to creation of an intellectual information platform based on service-oriented architecture (SOA). The main components of this platform are: service bus as a "backbone" of the platform framework; software modules for decision making support that are performed as web-services; business process execution language (BPEL) tools to form components interaction scenarios; intellectual interface for choosing necessary models and adjustment their parameters; program interfaces to interconnect of the diverse ERS and ancillary data sources.

Due to advantages of SOA and also thanks to models qualimetry methods application, diverse ground-based measurements (including crowdsourcing data) and ERS data are widely used for modeling, selection and setting of NNTOs models parameters during platform operation, as well as for verification of results. Platform enable using and integration data from different Earth observation satellites, including Sentinel series, Russian Resurs-P satellites, COPERNICUS capabilities, and others, while producing and provision of new information services in different modes: from interactive, through automated, to fully automatic. By today, successful case studies have been accomplished for creating the new services of flood forecasting, environmental protection, forest management, and so on.

In general, the platform developed is a universal constructor or software suite to produce services and downstream applications for decision making support in different applied areas.

The research described in this paper is partially supported by the Russian Foundation for Basic Research (grants 15-07-08391, 15-08-08459, 16-07-00779, 16-08-00510, 16-08-01277, 16-29-09482-ofi-i, 16-07-00925, 17-08-00797, 17-06-00108, 17-01-00139, 17-20-01214), by the Russian Research Foundation (grants 16-19-0019, 17-11-01254), by ITMO University's grant 074-U01, project 6.1.1 (Peter the Great St.Petersburg Politechnic University) supported by Government of Russian Federation, Program STC of Union State "Monitoring-SG" (project 1.4.1-1), state order of the Ministry of Education and Science of the Russian Federation No.2.3135.2017/K, state research 0073–2014–0009, 0073–2015–0007.

# METHODOLOGICAL AND METHODICAL FUNDAMENTALS OF THE INFORMATION FUSION MODELS' QUALITY ESTIMATION AND MODELS' QUALITY CONTROL

*V. Zelentsov, B. Sokolov, R. Yusupov*

SPIIRAS

v.a.zelentsov@gmail.com

The proposed theory includes two main research tasks. They are: elaboration of the methodological and methodical fundamentals of information fusion models and multiple-model complexes' qualimetry; development of the computer software prototypes implementing meta-models, methods, and algorithms of multi-criteria quality estimation and control in information fusion models and multiple-model complexes used for information fusion.

From our opinion, the proposed research theory includes the following main directions of the work: elaboration of the basic definitions, principles, and approaches used in the information fusion models' and multiple-model complexes' qualimetry; development of the hierarchy of conceptual models of developing situations, when the participants are the objects, the subjects, and the models being elaborated (used); classification and systematization of information fusion models and multiple-model complexes, determination of the interconnections and mutual associations of different types and kinds of information fusion models; classification and selection of the system of the parameters estimate the quality of the information fusion models and multiple-model complexes; elaboration of the combined methods for estimation of parameters of quality that are presented by digital and non-digital scales in the information fusion models and multiple-model complexes; elaboration of the methods and algorithms for solving the problem of multi-criteria analysis, arrangement and selection of information fusion models and multiple-model complexes, and quality control of information fusion models; development of object-oriented specification of the software tool according to the developed model of developing situations.

The primary goal of our investigation is the elaboration of the theory of information fusion models' quality estimation and quality control, qualimetry of information fusion models and multiple-model complexes of CS. The main objects of the investigation are the characteristics of information fusion models and multiple-model complexes of CS. The information fusion models are here considered as models, developed on the basis of natural and artificial languages. The latter ones include all formal languages. The elaboration of the theory should be started from setting up the appropriate terminology and system of definitions that will be the basis for the further argumentations and conclusions. The basic definitions used in the theory are as follows: developing situations, gnoseological and ontological models' adequacy, certainty, completeness, accuracy, essential correctness, model's utility.

Methodological basic of the research consists of following items: concepts of system analysis and complex modeling; principles of program-aimed and situating control; principles of requisite variety; principles of exterior supplementation and embedding; subjective and objective, and integral approaches to the modeling of complex objects and processes.

The analysis of information fusion models' quality estimation and models' quality control theory have shown that first focused item of the developing theory is the problem of multi-criteria estimation and quality control of information fusion models and multiple-model complexes. For this the classification of such models' parameters as adequateness, simplicity, accuracy, efficiency of the computer implementation, universality and scalability, multi-functionality and specificity, openness, cost, adaptability, flexibility, and intellectuality should be elaborated. The second item is the development of combined methods of multi-criteria estimation, analysis, arrangement, and selection of information fusion models. In the paper software prototype for information fusion quality estimation and models' quality control in the sphere of operational river flood forecasting is proposed to display show advantages and constructive of develop theory.

# INTEGRATION OF MODERN METHODS OF INTELLECTUAL DATA ANALYSIS IN GIS ENVIRONMENT

*R. Krasnoperov, A. Soloviev, J. Zharkikh, B. Nikolov, S. Agayan, A. Grudnev*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

r.krasnoperov@gcras.ru

Modern geoinformation systems (GIS) are one of the principal instruments for representation and analysis of spatial data on Earth sciences. Existing solutions, widely presented in the web, are basically interactive tools and online catalogs that provide access to spatial data, hosted on servers of scientific institutions and governmental agencies. From the other hand, there are commercial online services with a certain level of functionality that can be regarded as full-scale web-based GIS that allow to create and share geodata and provide sufficient analytical capabilities. The presented research is focused on the results of development of a web-oriented intellectual GIS, based on the client-server approach, that provides online access to geospatial data on Earth sciences, published as map services, and a set of geoprocessing tools. Raster and vector geospatial data, stored on a geodatabase server, are compiled into data-layers and maps and published on the GIS-server as map services. The published services are freely accessed via standard geospatial data transfer protocols (WMS, WFS, KML, etc.). The main feature of the developed system is integration within the GIS environment of data analysis algorithms as geoprocessing services that allow processing of vector and raster geospatial data layers. Currently the presented GIS provides standard clustering tools (e.g. Cluster and Outlier Analysis – Anselin Local Moran's I) and original data processing algorithms based on the Discrete Mathematical Analysis (DMA). DMA is a series of algorithms aimed on solving the tasks of clustering and tracing in multidimensional arrays, time series analysis, detecting trends and anomalies. DMA-algorithms are versatile and unified by the formal basis of fuzzy mathematics. It is an efficient tool for studying the finite metric domains and their mapping, in particular – multidimensional arrays and time series. These algorithms showed their efficiency in various applications to geological and geophysical data. Integration of geodata and analytical tools in unified web-oriented GIS makes it a versatile tool for interaction with spatial data on Earth sciences. It has a flexible structure and can be easily modified and reproduced.

This work was carried out in the framework of the project of the Fundamental Research Program of the Department of Earth Sciences of RAS No. 7 "Intellectual analysis of geophysical data, geoinformatics and mathematical geophysics" (project No. 0145-2016-0007).

# CREATION OF A MODERN SYSTEM OF REGISTRATION, PUBLICATION AND CITATION OF GEOPHYSICAL DATA

*N. Sergeyeva, E. Kedrov, L. Zabarinskaya, M. Nisilevich*

Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

n.sergeyeva@gcras.ru

A digital object identifier (DOI) is used for the unique identification of a digital object. Currently, the assignment of DOI to articles published in electronic journals has become generally accepted. The use of the DOI allows readers to easily access the necessary articles. Now the DOI system is expanded to a wide range (spectrum) of observational data. The data are considered as a valuable result of scientific research to be cited as the other scientific sources of information. The Data sets with assigned DOIs in repositories are subject to citation, they should be referenced in the articles. Data citation facilitates publishing and reuse of science data by linking between articles and datasets across networked data repositories.
The ICSU World Data System, CODATA, Force 11, DataCite, and CrossRef joined forces (are developing cooperation) to (develop a system) implement the practice of data citation and facilitate the access to high-quality scientific data for scientists.
Recognizing the importance of data citation, the Geophysical Center of the Russian Academy of Sciences (initiated) develops the project “Earth Science Database” (ESDB). The goal of the project is the creation of a modern system for registration and publication of geophysical data with assignment of DOI registered in the CrossRef system, which provides the ability to cite data in scientific publications.
The Central Repository and landing pages of data are formed during the registration and publication of databases and data sets. A web site of the Central Repository was created http://esdb.gcras.ru/. Landing pages contain data descriptions and information on producers and publishers, a sample of how to cite the data and the URL to access the data. Metadata about the registered object are stored in association with the DOI name, they include a location, such as a URL, where the object can be found.
On the first stage of the Project DOIs are assigned to geomagnetic data accumulated in the World Data Center for Solar-Terrestrial Physics (regular member of the World Data System) since its creation in 1957, and new observational data obtained under the current international projects, e.g. INTERMAGNET.
Since 2014 in the ESDB system the following objects have been registered: Database, including 6 catalogues of solar proton events over the years 1970-2008, Database of one minute magnetic variation measurements of the elements of the Earth's magnetic field recorded on 22 Russian observatories for the period 1983-2009, Databases of calculated values of the geomagnetic field elements derived at the observatories "Klimovskaya" and "Saint-Petersburg" of the Russian-Ukrainian INTERMAGNET segment, map of seismotectonics of Eastern Siberia, Database of 4 catalogues of magnetic pulsations for the period 1957-1992. DOIs are assigned to each data set within these databases.
We believe that the implementation of the ESDB project is a modest but useful contribution to the development of the World Data System and will help increase the availability of information and observational data on the Earth Sciences for scientific community. The project contributes to the introduction of a culture of citing data and more intensive use of data.

# UTILIZING DEEP CONVOLUTIONAL NEURAL NETWORKS FOR LANDSCAPE MONITORING

*N. Nikiforov, N. Bilev*

Lomonosov Moscow State University (MSU, Russia)

nikiforovnv@gmail.com

Nowadays remote sensing and high-end computing provide a unique opportunity to extract comprehensive information about landscape and terrain objects. This knowledge is of crucial importance since our planet meets a lot of threats every day, some of which cause severe damage and put human lives in danger. Landscape change is an essential source of information for detection of many harmful activities and events such as illegal logging and natural hazardous processes. This means that development of an efficient tool for landscape monitoring would be highly useful and valuable for governments, enabling them to respond quickly and efficiently to potential risks, and researchers providing them with state-of-the-art methods for performing complex analysis of landscape change patterns over time. Our study is focused on satellite imagery processing with the aim of automated recognition and classification of various regions of interest on the global map. Different terrain types couldn't be separated precisely by color properties only. Moreover, diverse atmospheric conditions make the challenge even more complicated. Fortunately, shapes, edges and textures provide significant information that allows us to solve all kinds of image recognition problems very accurately. Deep Convolutional Neural Networks have an ability to extract such important features from images and use them to create decision rules. We use neural networks training from scratch as well as utilize Transfer Learning with pre-trained models which achieved outstanding results on famous ImageNet Dataset. Transfer learning lets our new networks take advantage of features extracted from about a million images containing hundreds of different classes. Another issue to be addressed is that on some images important parts of studied regions could be covered with clouds. Detecting clouds helps us make our image dataset clear and ready for comprehensive analysis. Neural Networks with Convolutional architecture are widely reusable in application to image processing problems and it is essential to continue building research foundation for understanding our planet from space employing modern technology powered by advanced Computer Science.

# INTELLIGENT INTEGRATION OF HETEROGENEOUS DATA AND KNOWLEDGE SOURCES FOR DECISION SUPPORT IN PERSONALIZED MEDICINE AND HEALTH CARE

*M. Balakhontceva, I. Kisliakovskii*

ITMO University, St. Petersburg, Russian Federation

mbalakhontceva@corp.ifmo.ru

In recent years, the development of IT in medicine tends to a healthcare paradigm shift, a transition to a value-oriented approach and P4 Medicine aspects. The trend is partly due to the fact that the world population is rapidly aging (the number of people aged over 60 is expected to reach 22% by 2050 (WHO 2012)), and therefore, the number of people who will need care is growing. Because of increased costs, it is necessary to optimize processes at all stages of care delivery. Thereby, the paradigm shift entails conceptual internal changes of the whole system. The processes become even more complex and accordingly generate more unstructured data. There are various data and knowledge sources, which describe the process of healthcare delivery. Such sources are characterized by significant heterogeneity, complexity, high level of uncertainty, incompleteness and even inconsistency of data.

Existing medical information systems (MIS) cannot cover all aspects of diagnosis and treatments occur throughout the patient's "life cycle". It is impossible to model the whole treatment process of a specific patient only with the help of data from MIS. This implies the need for using existing or developing new methods and technologies for the intellectual integration of heterogeneous data and knowledge sources. In our case, as an example of such methods, we proposed the conceptual decision support framework for healthcare quality assessment. This framework allows decision-makers to have a set of alternative scenarios in real-time. It consists of several models linked by the data flow. This framework aggregate several heterogeneous data sources for simulation processes at medical departments. At the first stage, we need to model the workload of medical center by using results of EHR data analysis. The next step is a discrete-event simulation of department's activity to create schedules for each group of staff. Combining this data with access control data and expert time estimation of each event from MIS we obtain the input data for the next stage. Analysis of the workload for different staff groups is a very important aspect in case of decision-makers' support. At this step, as a result, it is necessary to obtain schedules for each employee to determine peak hours. It allows moving from discrete-event to agent-based model and receiving updated data for emergency scenarios. The agent-based model gives an opportunity to present medical clinic internal processes in continuous space. Simulation results of departments' dynamics are individual tracks in geometric space for each employee.

For the test scenario, we used the data from HIS of the Federal Almazov North-West Medical Research Centre. First of all, we chose one group of specialists. We chose daily nurses because this staff group has already the highest workload at the cardiological department at the Federal Almazov North-West Medical Research Centre. Obtained tracks of nurses do not give us the right to conclude and should be specified by additional data of events at the department. The framework can be adapted to use additional data sets, e.g., feedbacks from patients, results of surveys. After data analysis step, there are two scenarios: first, we have the set of refined individual schedules and can rerun agent-based model; second, we have some statistics for quality assessment and can obtain alternative scenarios for decision-makers.

# BEYOND OPEN DATA: TOWARDS STANDARDISED DATA ACCESS AND INTEROPERABILITY OF DECENTRALISED DATA REPOSITORIES

*J. Wagemann, S. Siemen, S. Lamy-Thepaut*

European Centre for Medium-Range Weather Forecasts

julia.wagemann@ecmwf.int

The European Centre for Medium-Range Weather Forecasts (ECMWF) has been providing an increasing amount of data to the public. ECMWF's data archive holds around 250 PB of data and is world's largest archive of meteorological data. One of the most widely used datasets include the global climate reanalyses (e.g. ERA-interim) and atmospheric composition data, which are available to the public free of charge. The centre is further operating, on behalf of the European Commission, two Copernicus Services, the Copernicus Atmosphere Monitoring Service (CAMS) and Climate Change Service (C3S), which are making up-to-date environmental information freely available for scientists, policy makers and businesses. However, to fully benefit from open data, large environmental datasets also have to be easily accessible in a standardised, machine-readable format. Traditional data centres, such as ECMWF, currently face challenges in providing interoperable standardised access to increasingly large and complex datasets for scientists and industry. ECMWF's current WebAPI is a very efficient and flexible system for expert users to access and retrieve meteorological data, though challenging for users outside the meteorological community. The growing volume of the data makes it challenging for users to fully exploit long time-series of climate reanalysis data (more than 35 years), as generally more data than actually required has to be downloaded and then extensively processed on local machines.

ECMWF is part of the EU-funded project EarthServer-2 and explores the feasibility to provide on-demand access to archived meteorological and climate data via the OGC standard interface Web Coverage Service (WCS). This allows for the efficient retrieval and processing of geographical subsets and individual point data information at the same time. ECMWF's ERA-interim data does not have to be downloaded anymore, but can be accessed on demand. Data access can further directly be integrated into custom processing routines.

Despite the potential a WCS for climate and meteorological data offers, the standards-based modelling of meteorological and climate data entails challenges and reveals the boundaries of the current Web Coverage Service 2.0 standard. Challenges range from valid semantic data models for meteorological data to optimal and efficient data structures for a scalable web service.

Geospatial web services reveal new opportunities for large data organisation to disseminate data. However, to make open environmental data a true success, large data centres have to become more progressive towards the adoption of geo-data standard interfaces, while data users have to be trained how to benefit from geospatial web services most.

Based on a practical example from the climate science community, the presentation will review requirements geospatial web services in general and Web Coverage Services in specific have to meet in order for data users and large data organisations to benefit from them most. A specific focus will be set on the challenges that are necessary to overcome in order to achieve real interoperability of and the conformable access to environmental open data.

# PREPARING OBSERVATION DATA FOR EUROPEAN REANALYSES IN ERA CLIM AND ERA CLIM2 PROJECTS

*S. Brönnimann[1], R. Allan[2], R. Buizza[3], O. Bulygina[4], P. Dahlgren[3], D. Dee[3], P. Gomes[3], S. Jourdain[5], L. Haimberger[6], H. Hersbach[3], P. Poli[5], J. Pulliainen[7], N. Rayner[8], J. Schulze[9], A. Sterin[4], A. Stickler[1], M. Antonia[10], M. C. Ventura[10]*

[1] University of Bern,
[2] NOAA,
[3] ECMWF,
[4] RIHMI-WDC,
[5] Meteo France,
[6] University Vienna,
[7] FMI,
[8] MetOffice,
[9] EUMETSAT,
[10] University Lissabon

stefan.broennimann@giub.unibe.ch

Reanalyses efforts ultimately depend on the observations of the atmosphere, the ocean, and the land surface used in the assimilation process. Projects that focus on different aspects of a reanalysis production (data rescuing, assimilation methods, products, diagnostic and evaluation) are ongoing in several world leading centers. The outcomes of these projects provide essential advances in the science and the understanding of the Earth System (ocean, land and atmosphere) evolution. Observations provide information on the state of the atmosphere and ocean that can be assimilated into a numerical weather prediction model in order to produce the reanalysis. Furthermore, observations are also used in several other steps along the processing chain.

In this contribution we describe the observation activities that have been performed within the European reanalysis projects ERA-CLIM and ERA-CLIM2. These activities include upper-air data rescue, satellite data recalibration, the generation of snow cover products, and the development of a global station metadatabase. The efforts performed by different research groups to discover, rescue, digitise and process various available sources of observational data, and to provide these data for the reanalyses in ECMWF are discussed. These efforts included data rescue (for historical observations), data management (including metadatabases), compilation and quality control, and error assessment. ERA-CLIM2 efforts in observation activities provided substantial advances in the number and quality of the data available for reanalyses' production and evaluation.

It is worth pointing out that the work on observations should ideally be completed one cycle ahead of the a reanalysis' production, so that the observations can be used in the production. For this to happen, it is necessary that there is a continued effort in data rescue, including a series of targeted research activities aimed to identify and exploit possible newly-identified data records. A large part of the data digitized in ERA-CLIM2 are scheduled to be assimilated in the new Copernicus Climate Change Service reanalysis ERA5 (https://climate.copernicus.eu/products/climate-reanalysis) which will go back to 1950. Thanks to projects such as ERA-CLIM and ERA-CLIM2, the set of European reanalyses (ERA-Interim, CERA-20C, and the forthcoming ERA5 and CERA-SAT) now provides detailed fields of main climate parameters that help researchers to assess climate change processes and be valid for climate information support and services.

The study was supported by EU ERA CLIM and ERA CLIM2 Projects. The continuation of this effort is first contingent upon organization of data rescue, but also upon series of targeted research activities to address newly-identified records.

# WDCS IN OBNINSK, RUSSIA: ON A WAY TO WDS RESOURCE INTEGRATION

*V. Kosykh, A. Sterin, E. Vjazilov, O. Bulygina*

RIHMI-WDC

sterin@meteo.ru

RIHMI-WDC (Obninsk, Kaluga region, Russia), the research institution of Roshydromet (Federal Service for Hydrometeorology and Environmental Monitoring), hosts two World data centers (WDC for Meteorology and WDC for Oceanography), both of them are integrated to WDS.

Historically, the system of WDCs in the USSR was established in the late 1950's – early 1960's, that is prior to the establishment of RIHMI-WDC (RIHMI-WDC was established as an independent institution in 1971). The functions of WDCs were delegated to RIHMI-WDC from institutions – predecessors. In previous years RIHMI-WDC used to operate four WDCs, namely WDC for Rockets and Satellites, WDC for Earth Rotation, WDC for Meteorology and WDC for Oceanography. Currently, only WDC for Meteorology and WDC for Oceanography are operated. Both are integrated to the WDS. WDC for Rockets and Satellites and WDC for Earth Rotation are no longer operated on the basis of RIHMI-WDC.

The paper describes both strengths and weaknesses of the above WDCs with an emphasis on the problems hampering the effective customer information support and sustainable operation. The range of data, products, and services offered by WDCs is presented. A brief overview of their technical, technological and institutional capabilities is also provided. Some of the problems being considered are typical BIG DATA problems.

Two most essential streams of data coming to WDCs are described. The first data stream is from observation networks. Data from observation networks are managed at WDCs according to the established procedures. . The second stream comes from various national and international centers, where data are collected and processed. So the role of RIHMI data centers is to monitor and integrate data from various sources. The paper provides some success stories in the field of customer support, based both on data from RIHMI-WDC's data sets and data from outside sources that became available by integration technologies. Examples of such technologies are provided and ways of extending their functionality for a wider set of problems are outlined.

# RIHMI-WDC: DATA ACCUMULATION, PRESERVATION, PROCESSING AND CUSTOMER SERVICES PROBLEMS

*V. S. Kosykh*

RIHMI-WDC

kosykh@meteo.ru

RIHMI-WDC (the Russian Research Institute for Hydrometeorological Information – World Data Center), located in Obninsk, Kaluga region, 100 km to south-west from Moscow, is one of the largest data institutions in Russia responsible for collecting, archiving and processing data on hydrology, meteorology and oceanography, as well as for providing high quality data and information products to a wide range of customers. RIHMI-WDC also hosts two World data centers (WDC for Meteorology and WDC for Oceanography), both of them are integrated into the WDS. In this paper the main organizational and technological approaches to data operations in RIHMI-WDC, as well as data policies approaches are analyzed. All these approaches are in strict compliance with the WMO data principles and with national data approaches. The parameters of RIHMI's data Volumes, data updating Velocity and data Variety (the "Three V" typical for "BIG DATA" definition) are provided, and that makes it possible to consider RIHMI as a "BIG DATA ENTERPRISE". Data sets available at of RIHMI are of great value for the study of natural disasters related to the atmosphere, hydrosphere, weather and climate change. Therefore, research activity based on the data sets accumulated at RIHMI-WDC and available to research community, is also one of the main goals of the Institute. The paper contains characteristics of some of the main "Nodes" of this RIHMI-WDC "BIG DATA ENETERPRISE", such as telecommunication infrastructure, reliable high-capacity archiving infrastructure, data processing and analysis instrumentation, data support to customers, as well as several samples of solutions that enable bridges between the RIHMI-WDC's own infrastructural units as well as integration of information from various sources.

# USING INTERPLANETARY SATELLITE DATA TO IDENTIFY THE PASSAGE OF THE SOLAR WIND STRUCTURES THROUGH THE EARTH'S MAGNETOSPHERE

*A. Potapov*[1], *B. Tsegmed*[2]

[1] Institute of Solar-Terrestrial Physics SB RAS,
[2] Institute for Astronomy and Geophysics MAS

potapov@iszf.irk.ru

Magnetic and plasma data on a high-speed solar wind stream measured by two spacecraft are compared. Data were taken from ACE located in the L1 libration point and WIND situated within the remote geomagnetic tail. It is shown that the stream structure mainly persists. Its main elements are observed both on the satellite ACE, and on WIND, but with a delay of 1 hour 52 minutes. Repeated is both magnetic and plasma structure essentials. Some fine structure details such as a so called "magnetic hole" are transferred as well. Measured in the distant geotail, the magnetic hole duration is doubled, but its depth has remained the same as in the L1 libration point. However the trailing part of the observed stream shear zone is strongly deformed. The consequences of the passage of high-speed stream structure through the magnetosphere by comparing the measurements of this structure before and after its interaction with the magnetospheric cavity are analyzed. Practical promises of data using from two spacecraft orbiting around the L1 and L2 libration points are discussed. Work was partly supported by RFBR grants 16-05-00631, 16-05-00056, and 17-52-44004.

6. Data Analysis, Event Recognition + Applications

# Multifractal analysis of different-scale electronic images of fossil coal surface

*V. Malinnikov*[1], *V. Zakharov*[2], *O. Malinnikova*[2], *D. Uchaev*[1]

[1] Moscow State University of Geodesy and Cartography,
[2] Institute of Integrated Mineral Development—IPKON

malinnikov@mail.ru

Currently the description of the surface structure of coal and the estimation of outburst hazard uses various methods of computerized processing of microimages obtained with the help of scanning electron microscopy (SEM) of surface of coal specimens. The most promising method is the multifractal analysis the objective of which is plotting a scaling spectrum (spectrum of singularities of image heterogeneity).

The undertaken research shows that spectra of multifractal dimensions in the images of coal specimens taken from outburst-hazardous beds are wider than these spectra in the images of outburst-nonhazardous coal specimens. In other words, the width of the spectrum of singularities (multifractal dimensions) can be an indication of coal susceptibility to gas dynamic events. It has experimentally been found that coal beds are non hazardous in terms of gas dynamics when the spectrum width $\Delta < 1.5$ and are susceptible to gas dynamic hazards when $\Delta > 1.5$.

The multifractal analysis of coal specimen surface microimages consists of a number of stages, including:

1. SEM to obtain different-scale digital images of the surface of coal specimens under analysis;
2. identification of characteristic areas in the SEM images of coal specimens;
3. denoising of images using the multiscale filter based on the Chebyshev moments;
4. calculation of multifractal characteristics for each identified area;
5. analysis of the obtained multifractal characteristics.

The estimation and analysis of the multifractal characteristics of the coal surface images used GeoPK software.

The multifractal analysis of the microimages at the magnification by 35, 85, 370, 1500, 7000, 15000, 40000 and 130000 times has shown that the values of the parameter $\Delta$ first grow with the image magnification, reach the maximum in the image magnified 1500 times and then decrease gradually. It has been inferred that the images magnified 1500 times demonstrate the maximum scatter of the fractal properties and are, accordingly, the most informative. In the images magnified 1500 times, it is seen that the structure of outburst-hazardous coal specimens is in a greater or lesser degree composed of particles round 1 μm in size, while the same magnification images of the outburst-nonhazardous coal specimens display a more uniform structure with cleats spaced at 25–30 μm.

The images at specified magnification inform on the size of separate coal grains from which methane liberates during desorption and diffusion, permeates cleats and participates in the initiation of coal and gas outbursts. Smaller structural elements in coal are not so interesting from the viewpoint of outburst hazard as they can contain cleats (pores) having size of the order of 0.1 μm, i.e. commensurable with the length of free travel of methane. Sufficiently free flow of methane in the mode of permeation is impossible in such cleats.

The accomplished research findings illustrate the applicability of the proposed multifractal approach to classification of coal with respect to coal susceptibility to destruction in the form of a gas dynamic event.

# MODULAR DISTRIBUTED SEMANTICS-BASED STORAGE FOR BIG DATA PROCESSING

*A. Visheratin, S. Rakitin, D. Nasonov*

ITMO University

alexvish91@gmail.com

It is a known fact that 21-st century is a Big Data era, when almost any area of humans' life produces great amounts of data, and this data can be used to produce a valuable information. One of the largest and the most valuable producers of data is science. For example, Large Hadron Collider alone produces more than 15 petabytes of data every year, NASA stores more than 37 petabytes of climate change data, and the volume of data produced by social networks, which are widely analyzed nowadays, is hard to overstate.

With all the variety of data types and sources comes the diversity in approaches for storing and processing of the data. There are SQL and noSQL databases (e.g. PostgreSQL and Cassandra), general purpose data storages (e.g. HDFS) and storages for specific types of data (e.g. SciDB, designed for scientific arrays). All these solutions, due to the need of flexibility and maximal users outreach, provide the same interfaces and methods for all stored information, hence not taking into account specifics of the data. Large companies and institutes, on the other hand, sometimes develop custom solutions for their needs, which consider peculiarities of the data. However, in many cases such systems remain unavailable for people outside the company and, even if they become widely available, they still provide high performance only for limited types of data.

We strongly believe that appliance of data semantics not only during the processing, but also in the way, how the data is stored, can greatly improve the speed of data operations as well as reduce the volume of data storage. Semantics includes composition of data, connections between entities (e.g. graph structure of social networks data) and special features of file format, in which the data is stored (e.g. metadata and binary offsets of NetCDF format). One of the most viable parts, where data semantics can be used, is indexing. There are numerous methods, which take into account type of data (text, image, spatial, etc.), but techniques considering more specific features, such as inner variables and relations between them, are not very common.

Taking all aforementioned into account, in our research we aimed to develop a storage, which could provide fast access to data, effectively distribute data across the storage nodes and be easily extended for any data format. The result of this research is Exarch - modular distributed storage, which allows users to obtain high-speed data upload and retrieval along with minor integration overheads. The main features of Exarch include:

- Flexible modular architecture that allows to utilize semantics of the data to perform efficient storage and querying. Three key modules of Exarch include unpacker, which transforms input files to the inner representation of the storage; partitioner, which is responsible for placement of transformed data onto storage nodes; and packer, which transforms inner representation of the data into meaningful files, which are then transferred to the user.
- Two-level indexing schema based on meta-information extracted from the data. Proposed solution provides custom partitioning tree for indexing qualitative data and indexing using space-filling curves for quantitative data, which allows to process any data type.
- Ease of storage adaptation for new data formats through implementation of modules interfaces, which was demonstrated on the example of the widespread format for geospatial data - NetCDF.

In order to validate applicability of the developed storage for real-world scenarios, we compared its performance for typical operations over NetCDF files with two widely adopted enterprise solutions - HDFS and Cassandra. Results show that in all scenarios Exarch outperforms both HDFS and Cassandra.

# EVALUATION OF RISK REDUCTION AND ADAPTATION MEASURES IN THE PACIFIC

*P. H. Havea[1], S. L. Hemstock[2], H. J. Combes[1], K. Maitava[2]*

[1] The University of the South Pacific,
[2] SPC

ilaisiaimoana@yahoo.com

According to the Global Risk Report 2016 produced by the United Nations University and Bündnis Entwicklungs Hilft organisation, the people of the Pacific island countries are in the most at risk region in the world to anthropogenic climate change and disaster caused by natural hazards. Most importantly, the report also identified that, since the Pacific island states have limited access to the funding and technologies needed to allow them to be more resilient and recover from the damages caused natural hazards and extreme events, they also have the lowest defence mechanism to protect themselves in a long-term and sustainable manner. This is mainly due to failures in risk reduction strategies and climate change mitigation and adaptation measures (AR5, IPCC). In response to this, using mixed methods research, this paper evaluates risk reduction and adaptation measures using the sendai framework. This evaluation is then used to contribute to the building of resilience in the Pacific by 2030 and beyond.

# BIG DATA IN THE MONITORING SYSTEMS OF AEROLOGICAL AND GEOMECHANICAL PROCESSES AT UNDERGROUND COAL MINES

*O. I. Kazanin*

Saint-Petersburg Mining University

kazanin@spmi.ru

Implementation of IT and IoT technologies is presented as one of the main directions of development of underground coal mining technology. The modern mines are shown as the hi-tech enterprises with automated technological processes, equipped with monitoring systems to monitor the mine subsystems, mine atmosphere, rock mass conditions, environment and staff. To provide safety of mining operations in coal mines in the conditions of growth of longwall productivity, the intensity of technogenic impact on the rock mass and complication of geological conditions (depth of works, natural methane content in coal seams, hazard of the rock bursts and gas sudden outbursts and so forth) the need of the qualitative forecast and monitoring in real time of aerological and geomechanical processes in rock mass and mine workings is marked. Besides collecting and storage of information getting from the monitoring systems the control, analysis and forecasting of processes and states, automatic formation of reports and recommendations are necessary. Requirements to coal mine multifunction safety system which components are the monitoring systems of aerological and geomechanical processes are stated. The structure of the applied systems of automatic gas monitoring, technical characteristics of sensors, periodicity of measurements, data volumes and transmission rate are given. The monitoring systems of the stress-strain statement of rock mass based on use of measurements of deformations and also acoustic and seismic parameters of the rock mass at operating coal mine are considered. Design of monitoring systems is made on the basis of potential threats (risk) assessment from which it is necessary to protect the mine. And, for example, if the system of aero-gas control (AGC) is mounted at all coal mines of Russia, then the system of geophysical and seismic observations and also a subsystem of deformation control of the rock mass in the majority of mines are absent. Besides, even where such systems are mounted, documentary procedures of regular actions of personnel in case of detection of the threatened situations formation (for example, the collapse of the main roof) aren't defined. Also there aren't enough specialists capable to interpret correctly data of geophysical and seismic observations. Need of standardization of requirements to monitoring systems, processing and interpretation of the obtained data, automatic formation of an alarm and preventive signaling is marked. Results of aerological and geomechanical processes researches in coal mines of JSC Vorkutaugol and JSC SUEK Kuzbass are given. The interrelation of aerological and geomechanical processes with intensity of mining operations, parameters of the applied technological schemes is shown. The directions of further researches for the purpose of improvement of monitoring systems, optimum use of Big Data flows for forecasting of dangerous conditions of the rock mass (the mine atmosphere) and complete use of the modern equipment potential are defined. Effective use of Big Data on coal mines is shown on the example of uniform dispatcher analytical center of the SUEK Kuzbass company.

# CAN WE EXPOSE DATA WITHOUT HIDING KNOWLEDGE? UNDERSTANDING THE CULTURAL COMPLEXITY OF RESEARCH DATA

*J. Edmond, M. Doran, G. N. Folan*

Trinity College Dublin

nugentfg@tcd.ie

One of the major challenges facing ICT integration in university research environments today is that of how to capture provenance and facilitate the use and reuse of "data". In particular, the growing pressure to find ways to enable both technological and cultural compliance with the proposed European Open Science Cloud (EOSC) are intensifying an already difficult proposition. But the urgency felt by university-based and independent infrastructural players may be obscuring an underlying, fundamental problem, which no technological development can adequately address on its own. The research infrastructures and computational approaches that facilitate "data" sharing may appear comprehensive, inclusive, and integrative; but this is an artifact of their having over time acquired a perceived objectivity that belies their curated, malleable, reactive, and performative nature. In addition, these approaches are highly selective, excluding input that cannot be effectively structured, represented, or digitised. Data of this complex sort is precisely the kind that human activity produces, but the technological imperative to enhance signal through the reduction of noise does not accommodate this richness. In an input environment where anything can "be data" once it is entered into the system as "data," data cleaning and processing, together with the metadata and information architectures that structure and facilitate our cultural archives acquire a capacity to delimit what data are. This engenders a process of simplification that has major implications for the potential for future innovation within research environments that depend on rich material yet are increasingly mediated by digital technologies. This is particularly the case when we speak, as proponents of the EOSC do, of open research data as a primary contributor to the enhancement of innovation capacity in Europe through the facilitation of increased levels and efficacy of inter- and transdisciplinary research.

This paper presents the preliminary findings of the European-funded KPLEX project (https://kplex-project.com/) which investigates the delimiting effect digital mediation and datafication has on rich, complex cultural data. The paper presents a review of extant criteria and guidelines for the recognition and classification of data in universities and humanities research environments. The paper reviews the classification systems, taxonomies, metadata, ontologies, controlled vocabularies, folksonomies, crosswalks & skeuomorphisms that mediate, curate, and influence our perceptions of and access to data within a research environment. It examines the crossover between analogue or augmented digital practices and fully computational ones, and in particular the epistemological influence of skeuomorphic metadata on digital research environments, shedding light on the strategies that have been developed to deal with this gap. The paper highlights the need to coordinate data standards and interoperability when it comes to the classification of rich and complex data and proposes a reconceptualisation of the infrastructures that curate data as it is functionally employed within digitally-mediated research so as to acknowledge and facilitate access to the richness and full complexity of the data at hand.

# OPEN GEODATA TO SUPPORT AGRICULTURAL RESEARCH

*D. P. Drucker[1], D. M. Pinto[2], E. C. Cardoso[3], D. O. Custodio[2], D. C. Victoria[1], B. T. Almeida[2], I. Pierozzi Jr.[1], C. L. Machado[3], V. V. S. Brandão[2], G. Bayma-Silva[4], M. R. C. Laforet[3], C. M. Takemura[2], L. H. Oliveira[5]*

[1] Embrapa Informática Agropecuária,
[2] Embrapa Monitoramento por Satélite,
[3] Embrapa Solos,
[4] Embrapa Meio Ambiente
[5] Embrapa

debora.drucker@embrapa.br

Access to trustable geodata sources is crucial to understanding agricultural production expansion, transition, intensification and diversification, allowing to map land use and land cover dynamics (LULCC). On the other hand, there are challenges associated to cultural barriers related to data sharing. Researchers are often concerned with potential misuse and lack of credit for their effort in data gathering (Costello 2009) or argue they have insufficient funding or time to dedicate to data curation. Even when the sharing culture is established and the data are accessible, they may be poorly documented and their usefulness compromised (Science Staff 2011).

We describe our effort on building a Spatial Data Infrastructure (SDI) at Embrapa, the Brazilian Agricultural Research Corporation, a networked public company composed by 46 research centers distributed throughout the country. A strategy was developed within an internal process that included the participation of actors with relevant assignments to the topic (Drucker et al. 2015). Qualified teams from 12 research centers, composed of IT specialists, geographers, agricultural sciences professionals, information managers, among others, were designated to frame the SDI in five components: People, Technology, Data, Institutional Framework and Standards.

A free and open source software platform, entitled GeoInfo, was established in order to promote efficient geodata management and facilitate access to open agricultural research data, in accordance with current Brazilian policy implementation guidelines and aligned with internationally widely adopted standards recommended by OGC – Open Geospatial Consortium. A process to organize, preserve, qualify and share geodata generated by Embrapa was collectively proposed, encompassing data quality issues, semantics, metadata and webservice standards, as well as capacity building. Within the process, which comprises the preparation, cataloging and the publication stage of the data set, there are roles for professionals with different backgrounds in order to assure documentation accurateness and to prevent overloading researchers with tasks they were not trained to perform.

During the endeavour, resistance and distrust regarding open access were unwinded as researchers recognized the value of making geodata they produce findable, accessible, interoperable and reusable (FAIR) and are using the SDI to spread their scientific products and to collaborate. GeoInfo SDI enables the interoperability of heterogeneous geodata from different sources, including semantic aspects. The redundancy of efforts and investments in obtaining and producing this information is avoided and the possibilities for its integration and application in diverse areas are expanded. Next challenges include expanding GeoInfo SDI usage with geodata products deposition from all 46 Embrapa Research Centers and increasing data products exposure through the Brazilian SDI. The steps taken to build the GeoInfo SDI facilitated awareness and ripening on the subject within the institution and has the potential to provide elements that can be adapted at other organizations.

References

1. Costello, M. J. 2009. Motivating Online Publication of Data. - Bioscience 59: 418–427.
2. Drucker, D. P.; Custodio, D. de O.; Fidalgo, E. C. C.; Daltio, J.; Visoli, M. C. 2015. Preservação e organização da geoinformação em instituições: o caso da construção da infraestrutura de dados espaciais da Embrapa. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 17., 2015, João Pessoa. Anais. São José dos Campos: INPE, 2015. p. 3671-3678. SBSR 2015.
3. Science Staff 2011. Challenges and Opportunities. - Science (80-. ). 331: 692–693.

# RAPID JUST-IN-TIME AND TASK-ADEQUATE GEODATA PROVISION FOR DISASTER MANAGEMENT

*F.-J. Behr, S. Fleischhauer*

Stuttgart University of Applied Science

franz-josef.behr@hft-stuttgart.de

Statistics prove that the number of disasters, especially weather-related, increased during the last decades. This leads to high demands on participants in disaster management processes where the information base is an essential part in these processes. Geodata, as a part of this information base, play a key role.
This report shows an approach using geodata to support decision makers and helpers in disaster situations. In a preliminary theoretical consideration the potential of geodata in disaster management is explained by the results of the VALID study. The economic value is illustrated by the estimated cost savings within a case study. Evaluations by participants in disaster management illustrate the strategic and operational benefits for the use of significant geodata products. This background research delivers as well the requirement specification for geodata provision and related products. These requirements are based on the use of general information structures in disaster scenarios. Specific aspects for the application of geospatial information are accentuated. To illustrate the state of the art this work gives an overview of existing concepts using geodata. The rapid mapping approach with remote sensing data illustrates the coordinated mapping process in consequence of a disaster event. The focus lies on the collaborative data acquisition based on the International Charter "Space and Major Disasters". Flood risk maps are an exemplary approach that shows the possible area-wide realization of a precautionary measure. The attention is focused on the mapping process and the position of politics as a driving force behind this realization. The geoinformation system SARONTAR is an approach of an operations control system for alpine regions. This concept combines real time communication features with satellite-based real time positioning. The system architecture and functionality are the main points of the explanation.
As an alternative, this work introduces a concept of immediate geodata provision for disaster management based on the usage of free data and free software components. The data storage is partitioned into a file-based system and a database system. The functionality of every single system component and the process of implementation are explained. The two data sources, remote sensing data from the United States Geological Survey (USGS) and vector data from the OpenStreetMap (OSM) project, are examined in detail regarding data structure and the form of data provision. In the examination of the remote sensing data there is a comparison between the United States Geological Survey and the Copernicus Programme of the European Union in relation to the data provision.
In the implementation of the system takes batch script controls every software component and data flow. During this workflow the user specifies interactively the OpenStreetMap data for this area are downloaded using regional data extracts or automatically using the Extended OpenStreetMap Application Programming Interface. Osm2pgsql stores this data as geospatial objects in a PostGIS database. By processing incremental update information the vector data of the disaster area are kept up-to-date. The satellite images are converted by the Sentinel Applications Platform to a RGB raster image. A sub-programme of the Geospatial Data Abstraction Library (GDAL) reprojects this raster image to the consistent target spatial reference system World Geodetic System 84/Pseudo Mercator.
The render software Mapnik uses both data sources to create mapping tiles of the affected area. The style sheet is optimized for visualization of disaster related information. The image tiles are suited for immediate use by end-users in disaster management and can be provided on a web-server The presented system makes it possible to provide geodata for disaster management participants rapidly within minutes.

# TOWARDS AN INTEROPERABILITY MANIFESTO

*F.-J. Behr*

Stuttgart University of Applied Sciences

franz-josef.behr@hft-stuttgart.de

Interoperability as the ability to collaborate and to exchange information seeminglessly and without barriers is an key issue in scientific and societal development. Different levels of interoperability are explained, i.e., interoperability on technical, personal, semantical, institutional (between communities), and political level.
Inspired by "The object-oriented Database System Manifesto" (Atkinson et al. 1989) typical features of a manifesto are described in order to initiate a collaborative effort to develop an Interoperability Manifesto which should build bridges across academic disciplines and national borders in order to foster technical, institutional and political level.
The overall goal is to have an impact on future academic and technical collaboration by having an impact on political and economical stakeholders.

# GLOBAL CHALLENGES AND DATA-DRIVEN SCIENCE: AN APPROACH FOR RAISING AWARENESS ABOUT INDICATORS FOR THE SUSTAINABLE DEVELOPMENT GOALS

*F.-J. Behr, M. Yahya, N. Abzar*

Stuttgart University of Applied Science

franz-josef.behr@hft-stuttgart.de

The 2030 Agenda for Sustainable Development, launched in 2015, is a shared plan that has motto of 'to leave no one behind' and it is deliberately ambitious and transformational consisting of 17 integrated and indivisible goals. The agenda is not geographically restricted or bound to a specific country, rather than this it's a universal agenda, applying to all countries. One main intention is to collect and maintain "quality, accessible, timely and reliable data" (UN 2015) about the indicators across the globe on regular basis to monitor the progress.

For each sustainable development goal, there is a need to collect data about objective targets and significant indicators in a standardised way so that it can be compared at regional, national and international level. There are several technical hitches associated with the collection of data as conditions varies according to geographical region. There is a need to design standard procedure and strategies to collect accurate, reliable and real-time data regarding SDGs. UN statistical department maintains a statistical global database of SDG indicators that offers data in CSV and MS Excel format to download. Availability and format issues regarding this dataset are evaluated and reported.

Geospatial technologies like Geographic Information Systems (GIS) provide powerful tools for management and dissemination of data, especially spatial data across the globe as in the case of SDG indicators. It has capacity to integrate locational information with ordinary statistical data. As everything is happening somewhere on the earth, Geospatial Information Systems (GIS) make it possible to present the information exactly where it belongs like in the form of maps and charts and allows to perform spatio-temporal analysis, an approach which might be essential for the awareness of the progress towards reaching these goals.

The availability of open source and data driven document libraries makes it possible to present the information in dynamic and interactive way that enables the effective use of data and efficient decision making. United Nations is looking for robust mechanism for the collection of high quality data and aimed to apply a follow up and review mechanism to keep a track of sustainable development. Numerous efforts have been made in this regard, however, some are still undergoing for the presentation of data in a meaningful way.

Data driven science is a new paradigm that has transformed the whole concept of data collection by providing additional functionality of visualization and analysis. Data driven document (D3), as an emerging an innovative technology can play a major role in presenting the information in a number of visualization elements and even it is possible to create custom drawing functions depending upon the nature, type and scope of data. Interactivity added by D3 plays very important role from the analysis point of view as it empowers effective visualization and efficient decision making. The visualisation of UN's statistical data will be exemplarily explained and analysed.

The framework of SDGs implementation can be boosted up by blending together geospatial technologies and data driven science facilitating the user to have an insight and more interaction with data to extract the meaningful results. SDGI data is publicly available, but this can be more effective if users are able to conceive meaningful information from it. For that purpose, the use of API can be beneficial that may fill data or services gaps and facilitate the process of collection, management and presentation of data efficiently.

# A SET OF FISHERY DATA FOR STOCK MANAGEMENT USING AN ECOSYSTEM APPROACH

R. Houssa

National Institute of fisheries resources

houssa@inrh.ma

The coast of Morocco, with a length of 3500 km, is known for its important fisheries biodiversity: demersal resources, pelagic, highly migratory and coastal resources. These resources of major economic interest are exploited by a national fleet and another foreign fleet under the fisheries agreements. The management of these resources is particularly challenging. Indeed, the variability of their stocks is largely conditioned by the variability of the marine environment and the fishing effort deployed.
In order to understand the evolution of stocks in the context of an ecosystem approach and to ensure proper management of these self-renewable resources, it is essential to promote the collection of information on:1. Total commercial fishing statistics: catches (total, by resource, type of fleet, etc.), revenue, fishing effort (number of vessels, number of trips, days, fishing time, etc.) etc.2. Activities and mode of operation of fishing vessels: use of VMS (Vessel Monitoring System) data.3. Biological sampling in the landing ports.4. Biological sampling aboard commercial vessels.5. Biological study at the laboratory level.6. Biological monitoring and sampling aboard research vessels.7. Monitoring of oceanographic parameters (Sea temperature, salinity, dissolved oxygen, Ph4, etc.) on research vessels.8. Continuous monitoring of oceanographic parameters by means of oceanographic buoys.9. Environmental scanning using satellite imagery.
The integration of these multi-source and multi-dimensional data and the use of appropriate analysis methods allows to:- Define long-term biological reference points;- Carry out short and long-term forecasts of catches and stocks;- Estimate the effect of different fishing strategies in the short and long term; - Estimate the effect of short- and long-term climate change on the resource and its renewal.

# USING SCIENTIFIC DATA FOR STATE LEVEL DISASTER RISK REDUCTION (DRR) AND CLIMATE CHANGE ADAPTATION PLANS: SIGNIFICANCE OF DOMAIN SPECIFIC DATABASE AND DATA MANAGEMENT

*U. M. Munshi, V. K. Sharma*

Indian Institute of Public Administration

umunshi@gmail.com

India is one of highly vulnerable country in Asia for natural disasters and faces floods, droughts, cyclones, earthquakes, landslides and forest fire very frequently. About 2 percent of country's GDP economic loss is occurring annually because of natural disasters. As such to address issues of DRR holistically, multilayer model of governance has been adopted by India. Thus in order to manage natural disasters, National Disaster Management Authority (NDMA) has been set up as the apex body headed by the Prime Minister of the country. While at the state level State Disaster Management Authorities (SDMAs), headed by respective Chief Ministers of the state and at District level the District Disaster Management Authorities (DDMAs) are headed by the District Magistrates and local elected representatives of respective districts. There are number of research institutions assisting these authorities and providing scientific inputs in disaster preparedness, planning and mitigation.

The central Water Commission is the nodal institution to provide flood forecasting and warning system. They have enormous data for each of the 160 perennial rivers of the country and over 150 flood forecasting stations throughout the country. Similarly, Indian Meteorological Department (IMD) is collecting data for climate (rainfall, humidity, temperature, wind and atmospheric pressure etc) and for seismological information through their network. Several scientific organizations are maintaining robust databases and supporting national, state and district level agencies in making disaster management plans. The Sendai Framework for Disaster Risk Reduction (2015-2030) has clearly emphasized the role of science and technology and academia in risk reduction. The impact of climate change on natural disasters can also be seen in different parts of the country. There is a need of interoperable good databases and data management system at various levels so that the re-use of such data can generate new knowledge for inferring value (using data analytics) for policy makers. This will give more scientific backup for strengthening disaster preparedness, particularly warning system and also to develop response strategies. All research institutions and academia needs correct data for research; hence, data management becomes most crucial for any research work in DRR. India needs to develop a policy to centrally place all required climatic/environment related data so that it can be accessed by all stakeholders and can be properly used in risk reduction. In case multiple agencies are responsible for data collection, the format and their coordination becomes important. The present Government is supporting science and technology and its use in DRR and encouraging scientific institutions to share the data and have better coordination. This would certainly help policy makers and planners in getting better scientific support in all aspects of disaster management.

The paper deals with the case study of data analysis using the existing database in the domain of environment with emphasis on DRR. The data analytics based on the data in the given database has been used for developing state level plans for DRR and climate change adaptation. The present empirical study is endeavouring to show-case preparation of such plans using the case of one of the North-Eastern States of India – Sikkim with mountains terrain wherein elevation ranges from 28,169 feet (the world's 3rd highest peak is the state's highest point).

# DATA SECURITY ISSUES IN CLOUD COMPUTING ENVIRONMENT: PROPOSED ALGORITHM

*U. M. Munshi*[1], *S. Pandey*[1], *P. N. Gopal*[2]

[1] Indian Institute of Public Administration,
[2] Banasthali University

umunshi@gmail.com

In today's information era, access to the internet has become necessity for the holistic development of micro and macro economies of any nation. The evolutionary changes on the technology frontiers have resulted in new delivery model of IT Services, which is now popularly known as "Cloud computing", that refers to the general perception of allowing people to access technology-enabled services using the Internet . Cloud computing is actually revolutionized the IT industry in terms of alternative mechanism of service delivery using internet on subscription based pricing model. Several benefits of cloud model includes faster deployment, scalability, lower total cost of ownership and elasticity. The imperative of such delivery model fosters business communities to focus more on their core business competencies rather than emphasizing on capital expenditure and skills development for IT. Despite several potential benefits associated with cloud adoption, there are some key prevailing issues, such as data Security, privacy, authentication and trust issues in cloud based applications.
With above in the backdrop, the present paper while deliberating on the challenges and issues, motivates the authors to further investigate the current research gap on cloud based business application and desire to face the challenges in solving the unsolved techno- functional problems related to security . Thus this research paper primarily emphasizes on various perspectives of cloud computing environment and data security issues and subsequently proposes the technical solution to protect organization's critical data. While attempting this, the authors are proposing an algorithm for data security in cloud.

# EFFECTIVENESS OF SOCIAL MEDIA: CHALLENGES AND OPPORTUNITIES

*U. M. Munshi, S. Pandey, A. Devi*

[1] Indian Institute of Public Administration,
[2] Benzara Inc

umunshi@gmail.com

In this ever changing and evolving environment, social media has changed the people to people interaction and communication style. Everyone is using social media for personal and professional use. This new media has changed the outlook of business and their marketing strategy. There is no doubt that social media not only enhance the brand visibility, customer engagement, but also created new opportunities for business. Social media tools like Facebook, twitter and LinkedIn have become a standard networking, communication platform and radically changed the way brands and customers interact with each other. In terms of business benefit, social media marketing helps in improving sales, increase exposure, lead generation, reducing marketing expenses, improving search engine ranking, increasing traffic, and creating brand awareness. In order to get success in online Social Media marketing, the MSMEs and Enterprise must have a strong social media smart marketing strategy.
This research paper carried out descriptive study to find out the effectiveness of social media as a marketing communication tool to identify new business opportunities and challenges. Social media can be great asset to organization, if systemically measured and quantified with relation to all the efforts and marketing strategy in a holistic manner. There are tools like Google analytics, kiss metrics, MOZ through which company can measure digital trails in digital spaces such as advertising source, impression, and traffic monitoring, reach and conversion rate to enhance brand awareness. The present empirical research work has endeavoured to analyse the primary data to evolve metrics that can focus on the trends and derive inferences to portray the pathways for brand awareness enhancement. These methods could also be applied to other domains for similar purposes.

# GENERALIZED RADON–NIKODYM SPECTRAL APPROACH APPLICATION TO EXPERIMENTAL DATA ANALYSIS

*A. Bobyl, V. Malyshkin, A. Zabrodskii*

Ioffe Institute

mal-codata@gromco.com

Radon–Nikodym approach to relaxation dynamics, where probability density is built first and then used to calculate observable dynamic characteristic is developed and applied to analysis of degradation type experimental data.
In contrast with L2 norm approaches, such as Fourier or least squares, this new approach does not use a norm, the problem is reduced to finding the spectrum of an operator, which is built in a way that eigenvalues represent the dynamic characteristic of interest and eigenvectors represent probability density.
The problems of interpolation and obtaining the distribution of degradation rates from sampled degradation data are considered.
Application of the theory is demonstrated on a number of model and experimentally measured signals of Solar panel degradation and Li-Ion relaxation processes. The ideology is similar to the one of random matrix theory [1], but now the matrix is built not from a theoretical model, but directly from experimental data. Software product, implementing the theory is developed [2].

References

1. T. Guhr, A. Müller-Groeling, and H. A. Weidenmüller, "Random-matrix theories in quantum physics: common concepts," Physics Reports, vol. 299, no. 4, pp. 189–425, 1998.
A. V. Bobyl et al. "Generalized Radon–Nikodym Spectral Approach. Application to Relaxation Dynamics Study." (2016), https://arxiv.org/abs/1611.07386, arXiv:1611.07386

# VRE FOR REGIONAL CLIMATIC PROCESSES ANALYSIS: ONTOLOGY APPROACH TO SPATIAL DATA SYSTEMATIZATION

*A. Z. Fazliev*[1], *A. A. Bart*[2], *E. P. Gordov*[1,2,3], *I. G. Okladnikov*[1,3], *A. I. Privezentsev*[1], *A. G. Titov*[1,3]

[1] Institute of Atmospheric Optics SB RAS, Sq. Academician Zuev, 1, Tomsk 634055, Russia
[2] Tomsk State University, Lenina av., Tomsk 634010, Russia
[3] Institute of Monitoring of Climatic and Ecological Systems SB RAS, Academichesky av., 10/3, Tomsk 634055, Russia

fazliev@yandex.ru

Analysis of environment response to ongoing and projected climate change shows necessity of adaptation measures elaboration. Users interested in such analysis belong to the three following groups: researchers dealing with basic studies of climatic processes and environment response, researchers solving applied relevant problems, like influence of climatic situation on ecosystems productivity, and decision-makers, which should take into account results produced by researchers.

To provide the first group of users with thematic Virtual Research Environment (VRE) for studies of climatic processes and caused by those ecological response we developed information-computational prototype for analysis of big data archives of spatial climatic data presented in netCDF, GRIB and PostGIS formats. Virtual Research Environment is presented as a set of interconnected standalone Internet accessible nodes. Each node is based on the Boundless / OpenGeo architecture [1], and provided with such software components as geoportal, cartographical web services based on Geoserver (http://geoserver.org/), Web GIS client, metadata database containing descriptions of available spatial datasets, computational core backend and its processing and visualization modules.

Computational backend is developed using GNU Data Language (GDL) and Python and implements functionality of complex statistical processing of environmental and climatic spatial datasets. Web GIS client is realized according to the "single page application" approach based on GeoExt JavaScript library, complies to general INSPIRE requirements [2] and provides launching of computational services aimed at solving of climate change monitoring tasks as well as presenting of obtained results in the form of netCDF files and cartographical WMS/WFS layers in raster (PNG, JPG, GeoTIFF) and vector (KML, GML, Shape) formats.

To assist the second group of researches a dedicated expert system is developed. It is based on ontology knowledge base (KB) describing stored in IMCES climatic data collections. Developed KB allows one to perform detailed semantic search climatic and meteorological data required to solve particular applied problems of regional ecosystem response.

To assist the third group of users a applied problems knowledge base is developed, which describes those for specific Siberia regional systems. It is assumed that this knowledge base will be inherent part of a thematic DSS under development.

In this report we focus mainly on the design of expert system KB assisting researchers solving regional applied problems. From technical point of view ontology KB design is related with implementation (automatic in future) OWL-ontologies with Semantic Web framework. Two major tasks of the ontology KB design are specially discussed. Those are the reduction problem and automatic implementation of taxonomy classes presenting sets of required to users individuals characterizing properties of ecosystems under analysis.

References

1. Becirspahic and A. Karabegovic. Web portals for visualizing and searching spatial data // Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on, Opatija, 2015, pp. 305-311
2. Katleen Janssen. The Availability of Spatial and Environmental Data in the European Union: At the Crossroads Between Public and Economic Interests / Kluwer Law International, 2010, ISBN 9041132872, 9789041132871, 617 p.

# CODATA RELATIONS IN PARTICLE MASSES AS A BASE OF NEW NUCLEAR SPECTROSCOPY

*S. I. Sukhoruchkin*

Petersburg Nuclear Physics Institute

sukhoruchkin_si@pnpi.nrcki.ru

Frank Wilczek responding on successful QCD/QED calculations of nucleon mass difference $\delta m_N=m_n-m_p=1.293$ MeV noticed that it open the way for future production of modern nuclear spectroscopy (he called it nuclear chemistry) which will have high accuracy of results comparable with the accuracy of atomic physics [1]. It means a hope for a further development of the Standard Model with a representation: $SU(3)_{col}\times SU(2)_L\times U(1)_Y$. Quantum Chromodynamics and Quantum Electrodynamics – the first and the last of three SM-components should be based on empirical relations interconnected with well-known value δ mN, Such integer relation in particle masses were found in 70-ties [2] in masses of the muon, the pion, nucleons (a common period δ equal to $16m_e=8.176$ MeV). According to recent CODATA evaluation [2], the shift $\delta m_n$ of neutron mass relative to $115\delta-m_e$ is equal to δmn=161.65 derived from a ratio between the neutron and the electron masses mn/me=1838.6836605(11). It accounts an integer ratio with nucleon mass splitting δmN=1293.3321(5) keV equal to δmN/δmn=8(1.0001(1)) which resulted in the ratio between me, mp and mn: $\{m_n = 115\times16m_e - m_e - \delta m_N/8\}$ $\{m_p = 115\times16m_e - m_e - 9\delta\, m_N/8\}$. Ratios $\delta m_N$: $\delta m_n=8.0$ and $\delta:m_e=16.0$ are considered as a manifestation of the fine structure with parameters 161 keV and me/3=170 keV. The exact CODATA relation could be established from the recent CODATA version (see Table).

Table

| **Compilation** | **(year)** | **$m_p/m_e$** | **$\delta m_n$, keV** | **$\delta m_N/\delta m_n$** |
|---|---|---|---|---|
| Fund.~Const. | 1969 | 1836.1518(8) | 162.1(4) | 7.986(12) |
| CODATA | 1998 | 1836.152668(4) | 161.652(2) | 8.00072(8) |
| CODATA | 2014 [3] | 1836.1526739(2) | 161.6489(5) | 8.00087 = 8(1.0001(1)) |

QCD is responsible not only for masses of hadrons but also for strong interactions between hadrons. We consider appearance of stable excitations in near-magic nuclei close to parameters of CODATA relations described in [4] as an indirect confirmation of the fine structure (namely $\delta m_N$, $m_N/8=161$ keV, $m_e$, $m_e/3=170$ keV etc.). Another confirmation of the period $\delta=16m_e$ was obtained in analysis of distribution of differences between particle mass data in the recent compilation Particle Data Group-2016 [5]. Combined analysis of CODATA, PDG-2016, nuclear excitations and nuclear binding energy could provide a further development of the Standard Model.

References

1. F. Wilczek, Nature 520 (2015) 303.
2. S.I. Sukhoruchkin, Statistical Properties of Nuclei, p. 215. Pl. Press, 1972.
3. P.J. Mohr et al., Rev. Mod. Phys. 88 (2016) 035009.
4. S.I. Sukhoruchkin, Nucl. Part. Phys. Proc. 282–284 (2017) 189.
5. S.I.Sukhoruchkin, Proc. QCD-17, Montpellier, France (in press).

# MULTI-OBJECTIVE DECISION-MAKING FRAMEWORK FOR EFFECTIVE WASTE COLLECTION IN SMART CITIES

*L. Manqele*[1], *M. Dlodlo*[2]

[1] CSIR,
[2] University of Cape Town

lmanqele@csir.co.za

There are metropolitan areas in smart cities that are experiencing waste collection challenges through ineffective methods of waste collection in resource constrained environments. This paper identified an opportunity to investigate efficient decision-making ways that will make use of data generated by IoT-enabled objects, taking into account the multi-objective goals in a smart city through addressing data loss challenge. Having the list of decision-making algorithms is one thing but choosing which algorithm to use requires intelligence. There is a need for decision-making algorithms that will be sufficiently dynamic to address different levels of data loss inherent in IoT data collection. This paper presents the framework for the sufficiently robust algorithm(s) that will enhance the smarter decisions in the smart city.

# RDA INTEREST GROUP ON AGRICULTURAL DATA (IGAD)

*I. Subirats*

FAO of the United Nations

imma.subirats@fao.org

The Interest Group on Agricultural Data (IGAD) promotes good practices in the research domain, including data sharing policies, data management plans, and data interoperability. As a forum for sharing experiences and providing visibility to research and work in agricultural data, IGAD has become a space for networking and blending ideas related to data management and interoperability. It also provides fertile ground to reach out and promote projects among other international organizations and institutions working in agricultural research and innovation. On a logistical level, one of IGAD's chief roles is to serve as a platform that leads to the creation of domain-specific Working Groups.

To this end, IGAD has taken a multi-faceted approach, splintering into specific focus groups each with their own tangible goals. In keeping with RDA strategy, IGAD has formed one successful Working Group (WG) – the Wheat Data Interoperability WG and two key discussion groups – on Rice Data Interoperability and Agrisemantics. Soil experts are also committed to partnering with IGAD.Under RDA guidelines, WGs have 18 months to achieve concrete deliverables and each of IGAD's groups are steadily moving towards this. At the RDA P9 the Agrisemantics group talked through its efforts to consolidate a portal to integrate existing ontologies, while the Rice Data Interoperability WG committed to a five point action plan that includes a survey to gather different types of rice data and best practices in digitalization aimed at opening up access to Thailand's national data legacy.

Objectives To promote good practices in the research domain: data sharing policies, data management plan, data interoperability To provide a platform for networking and cross-fertilization of research ideas in data management and interoperability To solicit and promote interactions and projects among the major international institutions and groups worldwide which work on agricultural research and innovation To achieve data interoperabilittEvents at RDA The RDA holds its plenary meetings every six months at different locations around the world. Built around the RDA Working and Interest Group meetings the sessions are designed to consolidate the RDA's work and that of its interest groups in building social and technological bridges towards open data sharing. IGAD's role since its inception at the RDA Gothenburg has expanded with every successive Plenary. The outcomes of the Wheat Data Interoperability Group have proved to be a success and the challenge is now on to replicate those successes with the Rice Group.

IGAD has used the Plenary meetings as a place to reach out and forge new alliances with other groups, as well as creating new offshoot groups aimed at specific solutions. During Plenary sessions IGAD has hosted a wide array of speakers and discussions and it continues to work alongside major international initiatives such as GODAN, GACS, CIARD, CGIAR, FAO of the UN, INRA, CYMMIT among others.

# INCOMPLETE DATA RECONSTRUCTION ALGORITHM APPLIED TO INTERBANK MARKET

*V. Y. Guleva, V. V. Povazhnyuk, K. O. Bochenina*

ITMO University

valentina.gul.va@gmail.com

Anticipating banking crises requires the understanding of processes in complex interbank market, understanding the influence of the components on the structure of interbank market and resulting topological dynamics due to complex interplay of different components of the system. Since banking system demonstrates emergent behaviour and the structure of interbank interactions may affect shock propagation in unpredictable way, resulting in different levels of systemic fragility, the observable patterns seems to be mathematically unpredictable. Therefore the modeling of possible scenarious is required. Nevertheless, data of interbank transactions are not public, therefore we have to guess network structure using other available data, which cause uncertainses during the simulation. The incompleteness of data sometimes significantly hamper the exploration of system properties and modeling across all research areas. Nevertheless, lots of available data allows for some desired data reconstruction with help of observable patterns and implicit properties. Modeling and exploration of banking crises, and the examination of systemic fragility causes, also faces the problem of data incompleteness. In this particular case, when the subject of our study is interbank market and data available are from annual reports and empirical research, we are interested in reconstruction of interbank network structure. We present an algorithm of graph structure reconstruction using node states dynamics data (represented by bank balance sheets for our particular case) and graph topological properties. The method presented was compared with three other methods commonly used for interbank network reconstruction problem, namely, maximum entropy, minimum density and low density methods.

We applied the metaheuristic algorithm to interbank market reconstruction problem, namely, we used a simulated annealing algorithm. And it is Russian interbank market that was chosen for algorithm examination. A network contained the 504 largest Russian banks to was compared with corresponding empirical results obtained by Leonidov and Rumyantsev (2013). The topological properties of a graph to be fitted was average in- and out- degree, density and average clustering coefficient. The proposed algorithm of network reconstruction was compared with popular maximum entropy method, and two methods provided by Anand (2015), minimum density and low density methods. Results shown the efficiency of the approach.

# EVALUATING TRUST THROUGH TEMPORAL INFORMATION

*J. Lee[1], V. Rivera[1], M. Mazzara[1], P. Dondio[2], L. Longo[2]*

[1] Innopolis University
[2] Dublin Institute of Technology

j.lee@innopolis.ru

As online interactions among virtual entities become more and more popular, the matter of online trust gained significant importance. There exist many reputation systems to measure trustworthiness of online users which utilize different types of information and facilitate different purposes. For example, in online market places such as Amazon and eBay, it is reasonable to measure reputation of sellers and buyers based on their previous transactions while in malware detection systems we should prioritize current behaviors to old behaviors since many attacks are usually brand new. Therefore, depending on applications, many reputation algorithms have been proposed and they require different domain-specific information to compute reputation to perform better than others. We propose to use a reputation scheme that utilizes only temporal information of interactions to compute reputation. We explore the ideas in existing works and evaluate the performance with a new dataset from StackOverflow. We chose StackOverflow for the ease of getting datasets as well as the size. The main temporal features we consider to compute reputation of users are regularity, frequency, presence and activity. Other types of information such as nodal attributes and relationship of users has rich value since a reputation value is influenced by context of interactions, feedbacks of interactions from the neighbors of users, how many neighbors a user has, to name a few. However, such valuable information is hard to collect and domain dependent. Our goal is to design a general model which not only stand alone but also can be embedded to existing reputation systems regardless of domains. In order to achieve the goal, we need domain independent information to compute reputation for users. We believe that the four aspects of time information are able to capture trustworthiness of users. We compare the reputation results with other methods using non-temporal information to demonstrate the effectiveness of the proposed scheme.

# DATA GENERATION THROUGH GIS AND REMOTE SENSING TOOLS TO ANALYZE AND ASSESS MULTI HAZARD RISKS AND MAKING DISASTER RESILIENT COASTAL COMMUNITIES IN BANGLADESH

*A. Q. Mahbub*

University of Dhaka

aqmahbub@yahoo.com

Bangladesh, with an area of only 147570 $km^2$, is one of the most populous (162 million) and densely inhabited (over 1200 people per $km^2$) place on earth. The country is also known as one of the most disaster prone nations in the world. The high magnitude of the country's disaster problems are largely due to the geographical location and physical settings along with various anthropogenic factors (e.g. low HRD, mass poverty, social inequality, very high density of population, lack of good governance etc.). Being located at the largest delta on earth (Bengal Delta), the most part of the country (almost 80%) has had been formed through the fluvial process of the mighty GBM (Ganges-Brahmaputra and Meghna) river system. In the coastal part the formation of the delta is still remains active. The physical characteristics causing a wide range of disasters are considered to be the deltaic land with very flat topography, hot and humid climate, riverine country with fragile and open coastal belt, drainage congestion, low river gradients, heavy monsoon rainfall, enormous discharge of sediments, funnel shape and the relative shallowness of the Bay of Bengal, etc. All these factors together makes Bangladesh vulnerable to various natural hazards such as flood, cyclone and associated storm surge, tornado, riverbank erosion, drought, etc. In fact, the country has become a playground of a wide variety of hydro-meteorological hazards which occur here almost routinely. In addition, due to very low height deltaic flatland and sea side location, Bangladesh has been considered as a leading affected country cause by global warming and associated sea level rise.
The impacts of natural hazards and climate change are observed all over the country. However, the country's highly economic potential but fragile and very low-lying coastal region which accounts for 20% of the country's total area, 22% of total population, demonstrates a sever impacts. The fate of economic development, employment generation, port facilities, food security, and geo-political strength of Bangladesh heavily depends on disaster friendly sustainable development of this vital coastal region of the country. Disaster and development linkages are now well established fact and given this, Bangladesh's future is greatly depends on the proper management of our blue economy. In the background of this notion, this paper is going to focus mainly the scenarios of natural hazards and climate change impacts in the coastal region of Bangladesh and how the adverse impacts of hazards and sea level change can be tackled by undertaking research using cross border hydro-metrological data as well as available statistics collected and stored by different institutions like academic, professional and government departments. Using various data like afforestation of mangrove forest, establishment of cyclone shelter, and construction of embankments or polders etc. in the coastal region, the paper is going to highlight the challenges of combating cyclonic hazards in the Bay of Bengal. The paper also brings out some schemes for scientific analysis of data that reflect the contemporary challenges of data management, share and exchange among the stakeholders and across the borders.

# INSTITUTIONAL RESEARCH DATA MANAGEMENT SERVICES AT PURDUE UNIVERSITY

*M. Witt*

Purdue University

mwitt@purdue.edu

Five years ago, the Purdue University libraries, information technology division, research office, and sponsored programs services collaborated to design and implement a campus research data repository service as a core university research facility. The service, called the Purdue University Research Repository (http://purr.purdue.edu), helps university researchers write and implement effective data management plans and provides them with a online platform to collaborate with others on research, manage and publish their data in a scholarly context, archive their data, and report the ongoing impact of sharing their data. The process of collaboration helped to introduce and connect people with expertise related to data as well as other services and infrastructure that are distributed across campus.

# THE ARCTIC FUTURES INITIATIVE – A SYSTEMS ANALYSIS PERSPECTIVE ON THE PLAUSIBLE FUTURES OF THE ARCTIC

*A. Reissell*

International Institute for Applied Systems Analysis (IIASA, Austria)

reissell@iiasa.ac.at

The Arctic is rapidly changing. Despite an ever-increasing volume of high-quality research, phenomena that are studied independently often interact in overlooked ways. These interactions call for a systemic analysis that does not yet exist. The analysis should be international – involving data and expertise from researchers in the Arctic states and other countries – as well as transdisciplinary: integrating sciences related to the environment, economics, anthropology, technology and geopolitics. The combination of these factors and how they will interact is currently missing, but is crucial to forming a clear picture of options on the future of the Arctic. A holistic, inclusive approach to such a synthesis would contribute to informed decision-making as human choices determine the future of the Arctic.

The Arctic Futures Initiative (AFI) is a project for a balanced Arctic future in a global context. It is a platform launched by the International Institute for Applied Systems Analysis (IIASA), a neutral interdisciplinary scientific institute whose mission is to provide guidance to decision makers worldwide.

Global socio-economic changes, such as increasing global demand for new energy and natural resources, growing global population, increasing emissions of greenhouse gases, as well as technological changes, all affect Arctic societies, climate and environment. Due to the complex intertwining of such factors, the Arctic Futures Initiative answers a demand for a holistic, integrated approach to the development of the Arctic, its variety of actors with differing values, and diverse national strategies.

AFI focuses on providing decision-makers with options that balance environmental protection, economic prosperity and societal well-being for the rapidly changing Arctic.

Globally, the unique capacity and methodology to work in a neutral, integrative and inclusive manner lies with IIASA, housing the expertise and integrated methodology applicable at the global, regional and national scale on a wide range of interlinked issues, all critical to the potential development options for the Arctic.

The AFI process identifies gaps in knowledge, shared challenges and opportunities across the Arctic. As a forum for analysis and synthesis, AFI engages stakeholders and builds capacity, aiming to establish a new generation of thinkers for the Arctic. The outcomes are synthesis reports and an Arctic Integrated Systems Assessment to be used by public and private decision-makers in the Arctic and non-Arctic nations, in order to serve a sustainable Arctic future.

This presentation on AFI will highlight selected scientific results and starting work, inform on activities, and elaborate on plans for the Initiative in the short and long term.

# Research Data in Chemistry: Past and Future

*D. P. Martinsen[1], L. McEwen[2]*

[1]David Martinsen Consulting
[2]Cornell University

dmartinsen@consultdpm.com

Like many other disciplines, e.g. crystallography, astronomy, and physics, chemistry has a somewhat long history of publishing data in repositories that grew in response to improvements in technology and to meet the needs of both commercial and academic researchers. Spectroscopic databases, funded by publishers and government agencies, were common in the 1970s. By the late 1980s, JCAMP-DX was proposed as a mechanism for researchers to export and archive and share data from their instruments. In the mid-1990s, JCAMP-DX was taken over by IUPAC, and has since been expanded to a number of additional analytical techniques. During the same time period, the standard has not kept pace with recent developments in the areas of metadata requirements, provenance, and attribution. Vendor enhancements have resulted in a divergence of key metadata content across different instruments.

More recently, ThermoML was developed by NIST and IUPAC as a standard for capture and dissemination of thermodynamic parameters. The effort was driven by the need for consistent and complete reporting important for compiling and critically evaluating data for re-use in engineering and other applications. In an effort to enable authors of articles in a number of relevant journals to submit their data in the standard format, a software interface, Guided Data Capture, was developed. As it turned out, authors were unwilling or unable to use the software. NIST employed a number of interns to capture the data from those journals for standardized entry into the evaluation engine. An interesting outcome of the effort was that the capture of the data in a standard format allowed algorithmic analysis of the data and highlighting of possible errors in the articles. This is somewhat similar to the way in which checkCIF is used to raise attention of problems in crystallographic data. The development of tools to analyze discipline-specific data to aid in the validation and curation of research data is an important area of research as data publication and reuse becomes more of an expected outcome of scientific research.

Reuse of data, by both humans and machines, is a key component of research data publication. Depending on the complexity of the experimental technique, complex software may be required to extract and view/interact with the data. Mestrelab, a prominent vendor of spectra analysis software, has piloted an interesting approach to coupling published data with analysis capability. In their Mnova software, they have added a feature which allows an approved publisher to digitally sign an NMR FID dataset. Unlicensed users of Mnova opening the digitally signed dataset are able to view and interact with that dataset using the full power of Mnova. This approach is somewhat different than a software-lite edition, that allows general users a less powerful version of software. Mestrelab's rationale is that users seeing the full power of the package may be converted into purchasers while data publishers can make their research data available to a broader audience of users. This model could serve to address concerns about sustainability in a world where open access to data is expected.

Finally, there is a trend from the highly curated, discipline specific repositories mentioned above to more generic repositories where any scientist may upload and publish uncurated content. Critical questions regarding how best to sustain the validation and curation efforts, which are fundamental to reproducible science and dependable re-use of data, can benefit both from the perspectives of traditional data publishers alluded to above as well as newer data publication advocates.

# PROCESS OF EMERGENT SELF-ORGANIZATION IN CRITICAL URBAN SITUATIONS

*T. Fatkulin, D. Voloshin, N. Butakov*

ITMO University

mellowstripe@gmail.com

To date, mathematical modeling and data analysis methods are widely used in the field of study and forecasting of various critical situations, such as man-made and natural disasters, acts of aggression, and social cataclysms. Yet, emergent processes of spontaneous self-organization remain poorly understood in the research literature. Such processes can be used to increase effectiveness of resource management during critical situations, whereas previously functionable systems (infrastructure, social, economic) are temporarily unable to solve problems in the new circumstances. One recent example of this type of critical situations, which became the catalysts for the process of self-organization are tragic events that occurred in St. Petersburg on April 3 2017, when as a result of the attack on the subway station the work of the public transportation system was considerably limited. As a result, social services for dissemination of information on the possibility to offer and use taxi services spontaneously organized in several social networks. To the best of our knowledge there are no works on modeling of similar processes, as well as theoretical frameworks that successfully describe such phenomena. In the scope of ongoing research, we propose to address this problem through analysis and modeling the process of self-organization mechanisms emergence using data received from social networks and messengers, which served as platforms for the coordination of people united within one self-organizing community.

The key objective of our research is to build a model of self-organization process based on the analysis of dynamics and properties of social media discussions. The foundation for such analysis is served by an event-oriented approach, when each user message is treated as a separate event. We perform analysis of the formation of semantic classes of textual Information (topics) and discussions which reflect dynamic of information exchange within the set of previously identified topics. Identification of the stages of self-organization formation and stable role patterns of interaction between individual participants is performed as a part of the discussion analysis and has a goal to build a model for the evolution of a rule-based discussion Interactions between users.

We propose to use neural-network-based approach for location extraction from weakly structured social media text data. It allowed us to approximate the routes (start point, destination point) of corresponding spontaneously organized taxi services. The data at use is collected from the telegram channel created for Coordination of actions in conditions of transport restrictions caused by the critical situation in St. Petersburg on April 3 2017 (collected data from April 3 - the moment of creation of the group, On April 11). Text-based analysis of discussion participants allowed us to infer distinctive stages of online discourse evolution with the set of corresponding role patterns distributed among users. This role patterns will serve as a foundation for the model of discourse evolution process, and will help to investigate various characteristics of self-organization in different conditions.

# PRACTICE AND IMPACT OF DIGITAL DATA CITATION TASK GROUP: STATE OF THE PRACTICE AND RELEVANT IMPACT FACTORS WHEN WORKING WITH RESEARCH DATA

*F. A. Linares*[1], *A. Vahed*[2], *J. Brase*[3]

[1] Information International Associates (IIa),
[2] CSIR,
[3] Universität Göttingen - Georg-August-Universität Göttingen

falinares@iiaweb.com

The CODATA Task Group for Data Citation (2010-2016) produced a far-reaching report, developed principles for data citation (in collaboration with several partners) and contributed extensively towards the general knowledge and understanding of the practice of data citation. The last two years of the Task Group's work was dedicated to raising the international awareness of data citation. The task group has not yet looked at the impact of the work completed or at the impact these activities have had on data citation. It is therefore necessary to now investigate the impact all these activities have had on the actual data re-use and citation. Funders, of scientific research, have begun to require data management plans as part of their selection and approval processes. As part of such plans the open availability of the data is often also a requirement. It is important that the necessary standards, incentives, and conventions to support measuring the impact of re-use of data, data citation, preservation, and accessibility be put into place. Currently the measurement of impact, of citing online data, is complicated by the lack of established standard metrics.
The proposed next phase of the CODATA Task Group, organized jointly by several CODATA committees and International Council for Scientific and Technical Information (ICSTI), together with representatives from several other organizations, will follow up on previously completed work and will examine a number of key issues related to the impact of data identification, attribution, citation, and linking.
The main objectives during the next 2-year period are to (a) to understand the impact of data citation practices in the research policy and funding communities throughout the world, (b) to build and disseminate knowledge about the importance and impact of data citation, and (c) to identify and understand metrics for data citation and uses of these metrics. The intention is to prepare and document several knowledge sharing products. This paper provides insight in what has been accomplished so far by the task group, focusing on the state of the practice and factors that impact work when working with research data.

# DATA IN IMPROVING QUALITY OF EDUCATION

*A. Reddy*

National University of Educational Planning and Administration (NUEPA)

anugula.reddy@gmail.com

UN while adopting SDGs gave a call for data revolution to stimulate to help monitor and the progress. Steps are being taken to evolve new data architecture to collect high volume data, high frequency data across several fields. The data are collected to monitor and also to make policies to achieve SDGs. In education also data are being increasingly sought to be used to change the way schools function and the way educational systems are governed. The global learning crisis i.e. large number of children particularly in developing countries are failing to learn basic numeracy and literacy skills even after attending schools for many years has triggered the debate on functioning and governance of schools and education system. It is frequently lamented that there is no adequate data on learning levels of children, and the limited data that are available is not comparable across different contexts and over the years. Against this background, attempts are being made globally to develop learning metrics which is expected to help in global monitoring of progress towards quality primary and secondary education. This paper explores how data on learning achievement can be used to reform education and schools using evidence from India. India though not participating in international achievement tests such as PISA to collect data on learning but is not immune from assessment learning levels of children either. It has instituted National Learning Assessment (NAS) tests to assess learning levels. This paper examines the how data on learning levels are used in policy making in education in India and gaps and mismatches would be identified. The paper also examines how to align India's effort to collect data on learning achievement with global efforts to develop learning metrics.

# ANALYZING SOCIAL PATTERNS WITH WIFI

*V. Romanov, G. Succi*

Innopolis University

v.romanov@innopolis.ru

The total computational capacity to process new information grows rapidly, and the topic of big data is the long accepted necessity rather than a concept. However, the research in some areas can be stalled due to either the shortage of relevant data or missing links between available pieces of information. The study of social interactions is vastly underdeveloped, and it suffers from both of problems above. We propose to direct more attention to the information from ubiquitous communication networks such as cellular or WiFi. Recent studies showed that such networks, WiFi in particular, can be used to infer simple patterns in human behavior. We believe that proper system design and non-invasive network monitoring techniques allow collecting an immense amount of new information. When incorporated into the big data paradigm, the latter can provide additional insights into both single person behavior and the collective behaviour pattern in small social groups. In this paper, we are going to describe possible knowledge that is underutilized in existing communication networks.

# A DATA DRIVEN APPROACH TO TRANSFORM THE RURAL TOWN OF KOMBAT INTO A SMART CITY

*W. Nekoto*

Kombat Village, Grove Mining Namibia Pty Ltd.

wilhelmina0711@gmail.com

Urban migration in Africa has been on the rise in the last few years. Key factors that drive urbanisation are amongst, quality of education, health, housing, service, employment to mention but a few. National statistics also show that 2/3 of the Namibian population still lives in the northern rural region of the country. A home to culturally appreciative domestic farmers and indigenous groups. Infrastructure and service are still underdeveloped, in comparison to continuous developments that take place in the capital city of Windhoek and the fishing and mining coastal towns such as Swakopmund and Walvisbay.

The gap of inequality widen with urbanisation. Data ranks Namibia with the largest inequality gap. As push and pull factors of urbanisation attract a large community to the city, the inequality in these cities continues to grow. Additionally, the cultural identity of citizens in adapting to the cultural appropriateness of the cities is also affected. Kombat, a rural town in Namibia, is centrally located between the northern and central region of the country, mostly surrounded by the Northern region. The town has recently been acquired by a young business man with the vision to develop a culturally smart city to cater for large population in the northern region and Namibia. The paper will describe, a multidisciplinary need, to address the three core objectives to developing Kombat and how it is leveraging on data, open science initiatives, evolutionary psychology for the "human-in-the-loop" approach, towards the development of a competitive smart city. The paper addresses the FAIR data concept, and its application, which has thus far helped in identifying key sectors, to further inspire smart developments in the town. Surrounded by a mine, meteorite, and tourist attractions such as the Etosha and Okavango Delta, among these sectors are the Tourism (Astro-tourism), Agriculture, Mining(Astro-Mining, copper, lead, silver) and Logistics sector, which are key to attracting local and international human capacity to the town. The paper discusses how it has thus far, and will use open source technologies and approaches such as the Kappa Architecture and open source tools for its data streams/collection, processing, storage, its accessibility and reuse to the open community for similar initiatives in Africa. Thus by encouraging citizen participation to codevelop services, and leveraging on the communication of citizens, such a workflow encourages good governance in real-time.

In conclusion, the paper encapsulates its challenges and achievements thus far in using social media for citizen participation, the social challenges and perception to technology amongst some citizens, and alternative data sources, approaches and technologies used for gather sentiments, towards a Gross National Happiness goals. Key to culturally appreciative smart cities and to meet the challenges that exist within urban culturally appropriate cities.

# DATA-DRIVEN MODELING AND SIMULATION OF COMPLEX HEALTHCARE ENVIRONMENT FOR P4 MEDICINE

*S. Kovalchuk*[1], *A. Funkner*[1], *O. Metsker*[1], *A. Yakovlev*[2]

[1] ITMO University,
[2] Federal Almazov North-west Medical Research Centre

olegmetsker@gmail.com

Today P4 (personalized, predictive, preventive, participatory) medicine and healthcare are rapidly developing and applied in diverse applications for care of in- and out-patients. Contemporary research solutions deal with multiple scales of healthcare process ranging from single patient and disease to city-scale or even global populations. Developed solutions are often built using diverse data sources and models which enable describe, predict, or optimize certain aspect and scales of the system. Nevertheless, complexity of system-level modeling and simulation, high uncertainty of the disease development, and informal structure of care process lead to significant complexity of system-level solutions development. Within the proposed approach a hybrid dynamic approach for modeling and simulation of complex multi-scale healthcare system is proposed. The approach is based on integration and automatic control of various data sources and models within a single solution with dynamic identification of the composite model structure and iterative improvement of the model through data (observations) assimilation procedures. Key role is played by mining techniques (process mining, data mining, text mining) which can be combined for detailed identification of processes in healthcare (characterized by high uncertainty, informal or unknown structure, high variability, multiple aspects and sub-processes, etc.). Dynamic implementation of mining techniques enables automatic working with diverse patient flow with certain level of personalization, flexible classification and predictive modeling of care processes. Data-driven modeling and data assimilation techniques enable reactive model tuning during an ongoing patient care process (e.g. predict length of staying in hospital, clinical outcomes, and complications). This leads to enhanced support of decision making (both clinical and organizational). Scaling this approach and extension with various simulation solutions (e.g. based on system dynamics, discrete-event simulation, agent-based modeling) gives opportunity to investigate specific structure of patient flow with diverse patients under high-uncertainty conditions. This reflects intentions of P4 medicine in contrast to more common evidence-based medicine (dealing with "averaged" patients). Moreover, analysis and simulation of patient flow on multiple scales enable stochastic, ensemble-based, fuzzy approaches to be applied for quantitative analysis of patient flow and, in turn, decisions support. Thus, combination of data-driven models, simulation solutions, and model control procedures reveals ways to investigate complex healthcare environments and develop innovative solutions for translation of technologies into medicine and healthcare area, decision support systems, virtual environments, etc. Within the presented work several applications in medicine and healthcare we developed on the basis of the proposed approach to support care of patients with cardiological diseases: acute coronary syndrome, arterial hypertension, etc. The applications were developed in tight collaboration with Federal Almazov North-west Medical Research Center (FANWMRC) (Saint Petersburg, Russia), one of the leading cardiological centers in Russia.

# INTEGRATING HETEROGENEOUS EARTH OBSERVATION DATA FOR ASSESSMENT OF HIGH-RESOLUTION INUNDATION BOUNDARIES GENERATED DURING FLOOD EMERGENCIES

*E. Sava[1], G. Cervone[1], A. Kalyanapu2, K. Sampson[3]*

[1] Pennsylvania State University,
[2] Tennessee Technological University,
[3] National Center for Atmospheric Research

esava1128@gmail.com

Flooding is one of the most damaging hazards that cause extreme devastation worldwide every year. The increasing trend of flooding events, paired with rapid urbanization and an aging infrastructure is projected to enhance the risk of catastrophic losses and increase the frequency of both flash and large area floods. One of the key factors in preventing and reducing losses is to provide reliable information about the risks associated with flooding. Accurate predictions of flood extents are required to understand and mitigate the risk, before, during, and after an event. This prediction is a critical safety tool for flood preparedness, response and recovery, and provides actionable data for officials and the general public. Large volumes of data sets — commonly referred to as "big data" — derived from sophisticated sensors, mobile phones, and social media feeds are increasingly being used to improve citizen services and provide clues to the best way to respond to a situation through the use of visualization and GIS mapping. Such data, coupled with recent advancements in data fusion techniques of remote sensing with near real time heterogeneous datasets have allowed decision makers to more efficiently extract precise and relevant knowledge and better understand how damage caused by disasters have real time effects on urban population. This research utilizes multiple data sources generated during emergencies to improve the identification of flood mapping. Specifically it presents a fusion technique using satellite remote sensing imagery coupled with non-authoritative data such as Civil Air Patrol (CAP) and social media sources. It assesses the feasibility of integrating multiple sources of contributed data into hydrodynamic models for flood inundation simulation and estimating damage assessment. The goal is to augment remote sensing imagery with new open-source datasets to generate flood extend maps at higher temporal and spatial resolution. The proposed methodology is applied on two test cases, relative to the 2013 Boulder Colorado flood and the 2015 floods in Texas.
Results show that this methodology is particularly effective in urban areas, where hydrological models tend to contain more errors due to the complexity in simulating an urban environment with managed waste and drainage networks. The urban environment is also where most social media data are available, due to the higher population density and availability of data networks.

# MINING OF BIOSYNTHETIC PATHWAY DATA FOR GENOME GUIDED DISCOVERY OF NOVEL NATURAL PRODUCTS

*D. Mohanty*

National Institute of Immunology (New Delhi, India)

deb@nii.ac.in

The secondary metabolites biosynthesized by microbial, fungal and plant species constitute a large repertoire of pharmaceutically important natural products. Hence, availability of genome sequences of various organisms has opened up the opportunities for identification of novel natural product biosynthetic pathways and deciphering the chemical structures of genome encoded secondary metabolites by *in silico* genome mining. Recent studies have revealed that, apart from polyketides and non-ribosomal peptides which are of nonribosomal origin, very large number of peptide natural products are also biosynthesized using ribosomal machinery. In contrast to nonribosomal peptides, these posttranslationally modified peptide natural products are called, RiPPs (Ribosomally synthesized Posttranslationally modified Peptides). Genome mining studies have revealed that, completely sequenced genomes of various microbial and fungal organisms harbor biosynthetic gene clusters which can potentially biosynthesize polyketides/nonribosomal peptides as well as RiPPs [1, 2]. However, only for a small fraction of these biosynthetic gene clusters (BGC) chemical structures for biosynthetic products have been experimentally characterized. Therefore, it is necessary to develop powerful in silico methods, which can link genomic space to chemical space and can predict chemical structures of the putative metabolic products for these large number of BGCs.

Taking advantage of the continuous growth in completely sequenced genomes and experimental characterization of metabolic products of increasing number of BGCs by high resolution mass spectrometry, during the last decade ours as well as other research groups have developed data driven computational methods for deciphering chemical structures of polyketides, nonribosomal peptides and RiPPs [3,4]. These data driven computational methods use a knowledge based approach which derive prediction rules from analysis of genes to metabolite relationships in experimentally characterized PKS/NRPS and RiPP gene clusters. Mining of available data on large number of experimentally characterized secondary metabolite biosynthetic pathways has helped in relating the enzymatic domains present in these gene clusters to the chemical moieties they add during biosynthesis. The talk would give an overview of the data driven systems chemical biology framework to integrate information on diverse chemical structures of known secondary metabolites with protein sequence and structure databases to derive predictive rules for relating genes to metabolites. These computational frameworks have been implemented in RiPPMINER (A machine learning based resource for predicting cleavage and cross-links in RiPPs) and SBSPKSv2 (A structural bioinformatics resource for deciphering chemical structures of polyketides and nonribosomal peptides) [5,6]. RiPPMiner (http://www.nii.ac.in/rippminer.html) derives its predictive power from machine learning based classifiers, trained using a well curated database of more than 500 experimentally characterized RiPPs. The backend database of RiPPMiner catalogs information about modification system, precursor sequence, leader and core sequence, modified residues, cross-links and gene cluster for more than 500 experimentally characterized RiPPs. RiPPMiner can predict leader cleavage site and complex cross-links between post-translationally modified residues starting from genome sequences. Benchmarking of prediction accuracy of RiPPMiner on a large lanthipeptide dataset indicated high sensitivity, specificity, accuracy and precision. SBSPKSv2 (http://www.nii.ac.in/sbspks2.html) has been developed based on comprehensive analysis of sequence, structure and secondary metabolite chemical structure data from 311 experimentally characterized PKS/NRPS gene clusters with known biosynthetic products. The

various different types of catalytic domains present in these 311 gene clusters comprise of 1150 KS, 939 AT, 628 DH, 193 ER, 950 KR, 1292 ACP, 193 TE, 127 MT, 742 C, 717 A, 840 PCP, 23 CHS and also many examples of unusual domains like EC (Enoyl CoA hydratase), PH (Phytanoyl CoA hydroxylase) and PS (Pyran synthase ) etc. It also allows the user to compare the chemical structure of a given secondary metabolite to the chemical structures of biosynthetic intermediates and final products.
RiPPMiner and SBSPKSv2 are valuable resources for discovery of novel natural products by genome mining and rational design of novel secondary metabolites by biosynthetic engineering.

References

1. Wang H, Sivonen K and Fewer DP (2015) Genomic insights into the distribution, genetic diversity and evolution of polyketide synthases andnonribosomal peptide synthetases. Curr Opin Genet Dev. 35:79-85. doi: 10.1016/j.gde.2015.10.004.
2. Zhang Q, Doroghazi JR, Zhao X, Walker MC, and van der Donk, WA (2015) Expanded natural product diversity revealed by analysis of lanthipeptide-like gene clusters in Actinobacteria. Applied and environmental microbiology, 81(13), 4339-4350.
3. Medema, MH, and Fischbach, MA (2015) Computational approaches to natural product discovery. Nature chemical biology, 11(9), 639-648.
4. Anand S and Mohanty D (2011) Computational methods for identification of novel secondary metabolite biosynthetic pathways by genome analysis. in Handbook of research on computational and systems biology: Interdisciplinary applications (Eds. Limin Angela Liu, Dongquing Wei and Yixue Li) Medical Information Science Reference (IGI-Global), Hershey, PA, USA. pp 380-405.
5. Agrawal P, Khater S, Gupta M, Sain N and Mohanty D (2017) RiPPMiner: A bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links Nucleic Acids Res 45 (W1): W80-W88. DOI: 10.1093/nar/gkx408
6. Khater S, Gupta M, Agrawal P, Sain N, Prava J, Gupta P, Grover M, Kumar N and Mohanty D (2017) SBSPKSv2: structure-based sequence analysis of polyketide synthases and non-ribosomal peptide synthetases. Nucleic Acids Res 45 (W1): W72-W79. DOI: 10.1093/nar/gkx344

# TEACHING ADVANCES STATISTICS FOR DATA SCIENCE STUDENTS WITH A SOFTWARE ENGINEERING FLAVOUR

*V. Ivanov, T. Stanko, G. Succi*

Innopolis University

nomemm@gmail.com

Innopolis University is meant to be a brain force, an integral element of a new Innopolis city. Mission of Innopolis project is to create opportunities for the economic growth of the Russian Federation through further development of Information Technology, improving welfare of the nation and creation of a highly professional and intellectual society. High-quality education is a keystone of this vision. Master degrees of Innopolis University offer a high-quality IT education for young specialists and development engineers in the following programs: Software Engineering, Cyber Security, Data Sciences, and Robotics. All the courses are taught in English by foreign or Russian professors. The Data Science programme of Innopolis University focuses on large-scale data management and data-driven analytics by conducting fundamental and cutting-edge research in databases, data mining and machine learning.
One of the core cources of the curriculum is Advanced Statistics. The course deals with complex problems of big data processing using modern statistical approaches. The goal of the course is to help the students understand the power of non-parametric Statistics including various resampling methods, non-parametric classification and regression. These methods are applied in modern scientific and data-driven applications to solve a number of problems. The practical part of the course is built on top of modern software statistical software packages such as SciPy (www.scipy.org). The strong side of the course is its ubiquity and connections the most elements of the programme, other courses, such as Machine Learning, Optimization, and the diploma project. An important feature of the course is the constant evolution of curriculum such as addition of new topics that are closely related to the research directions of Innopolis University. For instance, one of the lectures is related to a very complex problem of software reliability research at large scale. During the talk we will report on the experience of teaching the course and the lessons learned.

# INFORMATION ABOUT GEOMAGNETIC DISTURBANCES DERIVED FROM GLOBAL GEOMAGNETIC OBSERVATION NETWORK: AE INDEX, DST INDEX, AND WP INDEX

*M. Nosé[1], T. Iyemori[1], O. Troshichev[2], D. Sormakov[2], J. Matzka[3], G. Bjornsson[4], G. Schwarz[5], S. Nagamachi[6], P. Kotze[7], H. Theron[7], L. Wang[8], S. Egdorf[9], S. Gilder[9], J. J. Curto[10], A. Segarra[10], C. Çelik[11]*

[1] Kyoto University,
[2] Arctic and Antarctic Research Institute,
[3] GFZ German Research Centre for Geosciences,
[4] University of Iceland,
[5] Geological Survey of Sweden,
[6] Japan Meteorological Agency,
[7] South African National Space Agency,
[8] Geoscience Australia,
[9] Ludwig Maximilians Universitat,
[10] Observatori de l'Ebre,
[11] Boğazici University

nose@kugi.kyoto-u.ac.jp

The AE index has been used to identify substorms or to estimate magnitude of ionospheric convection for more than three decades. This index is derived from the horizontal component of the magnetic field variations from 12 stations in auroral/subauroral latitude (61-70 degrees geomagnetic latitude (GMLAT)). These stations are Abisko [operated by SGU, Sweden], Dixon Island, Cape Chelyuskin, Tixie, Pebek [AARI, Russia], Barrow, College [USGS, USA], Yellowknife, Fort Churchill, Sanikiluaq (Poste-de-la-Baleine) [CGS, Canada], Narsarsuaq [DTU Space, Denmark], and Leirvogur [U. Iceland, Iceland]. Most of the stations are operated rather well and keep sending data to Kyoto University in quasi-real-time, which make it possible to provide the real-time AE index with science community. The provisional AE index is calculated by a few month delay, because it takes time to receive definitive data or visually check artificial noises with baseline correction. Digital data of the provisional AE index is available from http://wdc.kugi.kyoto-u.ac.jp. The Dst index has been widely used to identify geomagnetic disturbances in low- and mid-latitude, in particular, development of geomagnetic storms. This index is derived from the horizontal component of the magnetic field variations from 4 stations in mid-latitude (|GMLAT|=21-34 degrees), which include Kakioka [KMO, Japan], Honolulu, San Juan [USGS, USA], and Hermanus [SANSA, South Africa]. Quasi-real-time data have been transferred from these observatories to Kyoto University with little problems, resulting in continuous derivation of the real-time Dst index that is available from http://wdc.kugi.kyoto-u.ac.jp. The provisional/final Dst index is calculated after the definitive data are released from all of stations, thus it delays about 1-2 years. Geomagnetic field data with high time resolution (typically 1 s) have recently become more commonly acquired by ground stations. Such high time resolution data enable identifying Pi2 pulsations which have periods of 40–150 s and irregular (damped) waveforms. It is well-known that pulsations of this type are clearly observed at mid- and low-latitude ground stations on the nightside at substorm onset. Therefore, with 1-s data from multiple stations distributed in longitude around the Earth's circumference, substorm onset can be regularly monitored. We propose a new substorm index, the Wp index (Wave and planetary), which reflects Pi2 wave power at low-latitude, using geomagnetic field data from 10 ground stations. The stations are San Juan, Tuscon, Honolulu [USGS, USA], Canberra, Learmonth [Geosciences Australia, Australia], Kakioka [KMO,

Japan], Iznik [Boğazici University, Turkey], Fürstenfeldbruck [Ludwig Maximilians Universität, Germany], Ebro [Universitat Ramon Llull, Spain], and Tristan da Cunha [GFZ, Germany]. Digital data of the Wp index are available at the Web site (http://s-cubed.info) for public use. These products would be useful to investigate and understand space weather events, because substorms cause injection of intense fluxes of energetic electrons into the inner magnetosphere and potentially have deleterious impacts on satellites by inducing surface charging. In the talk, we will review the present status and future perspective of AE/Dst/Wp index derivation.

# OBTAINING OBSERVATIONAL DATA OF THE CHARACTERISTICS OF THE EARTH'S SURFACE IN THE SHORT-WAVE RANGE OF RADIO WAVES

*S. Yu. Belov*

M. V. Lomonosov Moscow State University (MSU, Russia)

belov_sergej@mail.ru

The problem of remote diagnostics of a "rough" earth's surface and dielectric subsurface structures in the shortwave radio wave band is considered. A new incoherent method for estimating the signal-to-noise parameter is proposed. Specification was carried out for the ionospheric case. This range makes it possible to diagnose a subsurface layer of the earth, since the scattering parameter is also formed by inhomogeneities in the dielectric permeability of subsurface structures. By using this method in the organization of monitoring sounding, it is possible to identify the areas of variation of these media, for example, for assessing seismic hazard, hazardous natural phenomena, changes in ecosystems, and also for some extreme events of anthropogenic nature. Also, these techniques can be used to develop a system for monitoring, monitoring and forecasting emergencies of natural and man-made nature, as well as for assessing the risks of emergencies. The idea of the method for determining this parameter is that, by having synchronous information about a wave reflected from the ionosphere and about a wave reflected from the earth and the ionosphere (or having passed the ionosphere twice when probing from a satellite), it is possible to extract information about the scattering parameter. The paper presents the results of recording the quadrature components of the signal by means of the ground measuring complex of installation of coherent sounding in the short-wave range of radio waves at the test site of the Moscow State University (Moscow). A comparative analysis is performed and it is shown that according to the analytical (relative) accuracy of the definition of this parameter the new method is an order of magnitude larger than the widely used standard method. An analysis of the analytical errors in estimating this parameter allowed us to recommend a new method instead of the standard one.

References

1. Belov S.Yu. 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. Prague, ISBN 978-80-906655-2-1. 2017. P. 161.
2. Belov S.Yu., Belova I.N. Environmental aspects of the use of remote sensing of the earth's surface in the short-wave range of radio waves. // IGCP 610 Third Plenary Conference and Field Trip "From the Caspian to Mediterranean: Environmental Change and Human Response during the Quaternary". Moscow, MSU. 2015. P. 29–31.
3. Belov S.Yu., Belova I.N. The analysis of methods of determination the scattering parameter of the inhomogeneous fluctuating ionospheric screen. // Atmosphere, ionosphere, safety. Kaliningrad, 2016. ISBN 978-5-9971-0412-2. P. 435–440.
4. Belov S.Yu., Belova I.N. Methods of obtaining data on the characteristics of superficial and subsurface structures of the earth by remote sensing in the short-wave range of radio waves. // IGCP 610 project "From the Caspian to Mediterranean: Environmental Change and Human Response during the Quaternary" (2013-2017), GNAS Tbilisi, Georgia, ISSN 978-9941-0-9178-0. 2016. P. 26–29.
5. Belov S.Yu. The analysis of monitoring data of the parameter scattering power the earth's surface in the short-wave range of radio waves. // Data Intensive System Analysis for Geohazard Studies, Geoinformatics research papers, eISSN: 2308-5983, DOI: 10.2205/2016BS08Sochi. Vol. 4, No. 2, BS4002, 2016. P. 50.

6. Belov S.Yu., Belova I.N., Falomeev S.D. Monitoring of coastal ecosystems by method of remote sensing in the short-wave range of radio waves. // Managing Risks to Coastal Regions and Communities in a Changing World. ISBN 978-5-369-01628-2, DOI: 10.21610/conferencearticle_58b4316d2a67c, St. Petersburg, 2016.
7. Belov S.Yu. The program of registration quadrature a component of the n-fold reflected radio signal from a terrestrial surface. The certificate on registration of the right to the software No. 2016612172 of 19.02.2016.

# SINOBON: BIODIVERSITY MONITORING IN CHINA

*X. Zheping*[1], *X. Xuehong*[2], *M. Keping*[2], *C. Xuefei*[1]

[1] National Science Library, Chinese Academy of Sciences (China)
[2] Institute of Botany, Chinese Academy of Sciences (China)

xuzheping@126.com

As one of the richest biodiversity countries in the world, China now encounters more and more challenges for balancing human well beings and biodiversity conservation. The government need more information about the dynamic changes of biodiversity and ecosystems. Therefore, more and more observation data should be collected from different field stations and integrated them at national level.

There are currently two observation networks in China. One is China BON led by Ministry of Environmental Protection (MEP) of the People's Republic of China. Currently, the biological taxa for monitoring include mammals, birds, amphibians and butterflies, and 441 sample sites have been established with 9000 line transects and point transects.

The other observation network is Sino BON established by the Chinese Academy of Sciences (CAS) following the basic principle of "sound planning and unified layout for biodiversity observation at national scale". It is a priority project supported by the CAS as an observation and research platform at the academy level. It became a member of AP BON and GEO BON.The Sino BON includes a zoological diversity center, a botanical diversity center, a microbiological diversity center and a synthesis center. In zoological diversity center, there are six thematic networks: the mammal diversity observation network, the bird diversity observation network, the amphibian & reptile diversity observation network, the freshwater fish diversity observation network, the insect diversity observation network and the soil invertebrate diversity observation network. In the botanical diversity center, there are three thematic networks: CForBio, the steppe & desert biodiversity observation network, and the forest canopy biodiversity observation network. Microbiological diversity center has a soil microbial observation network. The synthesis center is in charge of standards & criteria formulation, data management & sharing, and remote sensing with LiDAR. After the collection and integration of raw data, these data should be mapped into Essential Biodiversity Variables (EBVs),which will be facilitated to merge data in a regional and global level.

In 2013-2015, 84 million RMB (approximately 11.8 million EUR) grants were supported by the CAS, most of them are for kinds of observation settings. In 2016-2020, a budget about 100 million RMB (approximately 14 million EUR) from the CAS was approved, not only for settings and human resource, but also for data collection, publication and services.

Sino BON research should integrate multi-source mapping biodiversity data, including community assembly, biogeography, conservation planning, phenology, climate change and so on. It not only demonstrates monitoring methodologies with general equipments, but also develops new approaches to speed up biodiversity discovery and monitoring.

# NETWORK ANALYSIS OF INTERNATIONAL MIGRATION

*A. Rezyapova, F. Aleskerov, S. Shvydun, N. Meshcheryakova*

National Research University Higher School of Economics (Russia)

annrezyapova@gmail.com

During last years, migration has attracted a lot of attention and has been examined from many points of view. The migration theory has a long history of establishment starting from Adam Smith in the nineteenth century. Since that time a huge number of theories in various fields of science was developed. Migration was studied from the prospect of its causes, gravity modeling, push-pull factors theory and human capital approach. These works form the stem of migration theory and explore it on different levels. However, they assume migration as a bilateral process and migra-tion flows between any two countries are studied independently from the flows be-tween other countries.

We model migration in the framework of network analysis. It allows to present all countries involved in the international migration as a graph, where nodes are coun-tries and edges correspond to migration flows between them. The networks are con-structed for each annual dataset on international migration flows. Our work is aimed to detect the countries with the highest level of importance in the international mi-gration process. For this purpose, we evaluate the classical centrality indices: degree and weighted degree centrality, closeness, eigenvector and PageRank. Unfortu-nately, classical centrality measures do not reveal the most influential countries in the international migration process. This fact refers to the following reasons. Clas-sical centrality measures do not consider individual properties of the nodes, group influence and indirect connections.

Individual characteristics of countries, for instance, population are essential, be-cause countries with different population hosting the same amount of migrants ex-perience different impact of this flow on their population. Influence of group of countries in the process of international migration can be crucial, if we consider the example of migration from African region to certain European country. Single Af-rican country may not be influential in the network. However, when migrants from these countries form a huge share in the population of receiving country, it can in-dicate severe importance of the group of countries in the process. The indirect con-nections between countries are important to be considered for the following reasons. First, migration between any two countries may occur not directly, but through mi-gration route. In this case we need to detect the initial country that generates migra-tion flow. Second, the flow between any two countries can lead to the emergence of new flows between any other countries. For example, the country experiences huge immigration wave and it causes the emigration of its citizens to any other countries.

To consider all the relevant issues, we constructed our own index that accounts for the distinctive features of the problem. We apply Indices of Short-Range (SRIC) and Long-Range Interactions Centralities (LRIC). SRIC examines only first-long-range connections in the network and LRIC accounts for s-long-range routes. The network of countries is constructed based on the annual data (from 1970 to 2013) on international migration flows provided by the United Nations Organization (United Nations (2009, 2015). International Migration Flows to and from Selected Countries). The process of international migration is studied in the dynamics. The ranking of most influential countries in the process is provided for each year with the attention to important country characteristics and specificity of the process. The main trends and changes in the international migration are analyzed. These results are important in order to provide countries highly involved in the process of inter-national migration with the relevant migration policy.

# HOW TO ACCOMPLISH INTEROPERABILITY OF HETEROGENEOUS CONCEPTS? MULTIFUNCTIONAL SENSORS AND SYSTEMS FOR IN-SITU MONITORING OF MARINE ENVIRONMENT

*M. Wichorowski, S. Sagan*

Institute of Oceanology, Polish Academy of Sciences

wichor@iopan.pl

Growing computation power and storage equipment capabilities give opportunity of increasing data volume archived and processed in data centres. This trend is well known as big data paradigm in financial and business applications, but is observed in environmental sciences as well. Aggregation of data originating from variety of disciplines and exploring information not existing or visible in the past gives new possibilities for researchers. Exploding data volume also lead to demand of new services e.g. multidimensional analysis and data visualisation. The key factor for development of technology and data processing possibilities is interoperability on different abstraction layers: cooperation of data centres, data policy enforcement, deployment of standards, development of controlled vocabularies, transmission protocols, unification of data models and common sense of datalife cycle plan.
The Common Sense project well demonstrates applied interoperability of processes covering specification, design and prototyping of new multifunctional sensors for marine environment monitoring purposes. Sensors, as a part of bigger systems and infrastructures interoperating each other have to conform regulations formulated in legal acts in the context of parameters measured, condition of operation, interaction with environment as object of examination, transmission protocols and data collections terminating data acquisition, processing, storage, information extraction and exchange processes.
The trigger of development process was pulled by enforcement of variety of legal acts leading to achieve good environmental status. Couple of these regulations are well visible/known by the community: Marine Strategy Framework Directive (MSFD), the Common Fisheries Policy (CFP), Reduction of Hazardous Substances (RoHS), INSPIRE. The general objective was to provide support for implementation of these European Union marine policies and response for requests for integrated and effective data acquisition systems.
The way to accomplish objectives was increasing the availability and interoperability of standardised data on eutrophication; concentrations of heavy metals; micro-plastic fraction within marine litter; underwater noise; and other parameters such as temperature and pressure. This was facilitated through the development of a sensor web platform, called the Common Sensor Web platform.

References

1. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, Version 3.1, 25 August 2016
2. Guidelines on FAIR Data Management in Horizon 2020, Version 3.0, 26 July 2016
3. OGC® Sensor Web Enablement: Overview And High Level Architecture. Mike Botts, George Percivall, Carl Reed, John Davidson. OGC 07-165, 2007
4. Directive 2011/65/EC of The European Parliament And Of The Council of 8 June 2011 on the restriction of the use of certain hazardous substances in electrical and electronic equipment (recast) (Text with EEA relevance)
5. INSPIRE Monitoring Indicators – Guidelines Document, Monitoring and Reporting Drafting Team and European Commission
6. COMMON SENSE: Cost-effective sensors, interoperable with international existing ocean observing systems, to meet EU policies requirements. John Cleary, Margaret McCaul, Dermot Diamond, María Begoña González García, Cesar Díez, Concepció Rovira, Mike Challiss, Yassine Lassoued, Alberto Ribotti, José Sáez

Acknowledgements
This publication is result of Common Sense, the EU 7th Framework Programme project no 614155.

# A BRIEF HISTORY OF RESEARCH DATA POLICY

*P. Uhlir*

Consultant on information policy and management

pfuhlir@gmail.com

The history of scientific, or more broadly, research data policy is quite recent. Prior to the early 2000s or so, research data were generally not considered to be a separate topic for policy-making or much attention by research managers or researchers themselves. There were a few discipline-specific statements, such as the Bromley Principles on Access to Global Change Research Data [Bromley 1989] and the Bermuda Principles regarding the rapid deposit of human genome data in Genbank [Bermuda Principles 1996]. The open access movement, which began in the mid-1980s with Open Source Software and was extended to scientific, technical, and medical (STM) journals in the late 1990s, did not include the underlying data within the scope of consideration. Research data were subsumed as a part of each nation's government information law and policy, and they were an integrated into of science policy, but were not considered to be a separate element for analysis and management in the research process.

That began to change quite rapidly in the mid 2000s as a result of a confluence of factors. From a technical perspective, the internet and WWW revolution that enabled digitally networked research was taking off. Although scientists were among the first adopters of digital networks, there were many barriers to new perceptions of the role of data in the research process, that had to be overcome (and still do). There also were other related technical infrastructure factors that needed to be addressed, but the time was ripe for a new phase in research [Atkins 2003].

Scientifically, a new research paradigm was emerging–data science or data-intensive science–that generated, networked, and used the rapidly accumulating data resources in novel ways. A seminal report was published in 2009 by Microsoft Research, which described the new and faster research exploring and using massive data sets [Hey, Tansley, and Tolle 2009].

From an organizational standpoint, data centers to store and make available the vast data reserves were sprouting up in all OECD countries and serving most disciplines. At the same time, many research organizations worldwide–governmental, intergovernmental, non-governmental, and discipline-specific–issued statements, declarations, and principles in support of more open access to research data [Doldirina, C., et al. 2017].

Finally, legal and policy innovations, modeled after the open licensing of software and the STM journal literature, was applied to scientific databases as well [Reichman and Uhlir 2003; Creative Commons 2006]. This was coupled by a growing set of research agency and foundation guidelines, and related requirements in contracts and grants, to make the data generated in the course of research available to other researchers for reuse, especially after the publication of their research results. Various science policy organizations and learned societies issued advisory reports that helped make the case for open data for research [CODATA 2015].

By about 2013 and onward, many of these trends became greater and more pervasive, a maturation of the movement for open data for data science and related applications. The focus has shifted from answering the question of, why?, to an examination of the details of how research data can be made most productive and useful. This presentation will describe these three phases of research data policy and provide some lessons learned and possible future developments.

References

1. 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).
2. Bromley, A. (1991). Principles of Access to Global Change Research Data
3. CODATA (2015). The Value of Open Data Sharing. Paper commissioned by the Group on Earth Observations, Geneva, CH.
4. Creative Commons (2017). http://creativecommons.org.
5. Doldirina, C., Eisenstadt, A., Onsrud, H., Uhlir, P.F. (2017). Legal Approaches for Open Access to Research Data.
6. Hey, T., Tansley, S., and Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Press.

# ADAPTING UNDERGRADUATE PROGRAMS OF DATA SCIENCE TO THE PROFESSIONAL REQUIREMENTS OF THE INDUSTRY

*N. Ahituv, A. Hasgall*

[1] Tel Aviv University,
[2] The Center for Academic Studies

ahituv@post.tau.ac.il

Data Science (DS) has developed in recent years mainly because of the need to make decisions based on huge amounts of data – Big Data, and because of the development of technology that allows to create patterns and provide relevant meanings to information.
A salient problem in the relationship between the scientific discipline of DS and the industry, which needs to employ practitionnairs in the area, is how to transfer the high level theoretical knowledge of DS to an undergraduate academic program that will meet the growing demand for data analysts in the industry. The purpose of this paper is to demonstrate how the scientific discipline of Data Science fits into an undergraduate academic program intended to prepare data analysts for the business, public and government sectors.
The article explains why the demand for Data Scientists and Data Analysts is growing tremendously. It delineates the professions and industries where DS experts are required. It reviews a number of DS programs among academic institutions worldwide, and defines the components constituting a DS program.
The article then classifies DS programs into three types, based on the depth of learning and knowledge required. It argues that the second type, type B, is the most required type, since it meets the most prevailing organizational needs.
Following that conclusion, the article divides the curriculum of type B program into a number of categories and lists out the courses included in each category. Since DS is a new academic discipline, this article is intended to assist academic institutions in introducing a DS program into their studies.

# OPEN SCIENTIFIC LITERATURE MANAGEMENT AND APPLICATION IN CHINA—GOOA

*X. Chen, J. Huang, F. Wang, Z. Xu*

National Science Library, Chinese Academy of Sciences

chenxuefei@mail.las.ac.cn

In China, open innovation, open science, open resources are valued by the government, enterprise and scientific community. The Chinese Academy of Sciences and National Natural Science Foundation of China declared their open access (OA) statement separately in 2014, which required scientists to open and store their papers and associated data in public repository. Meanwhile, the unprecedented volume of open scientific literature data featured of scattered, easy to loss and instability around the world needs to be integrated and managed, which could help Chinese researchers to reuse and explore as much data as possible.

To solve this problem, we start the “Open Access Journals Integrated Service System Project (GoOA)” (http://gooa.las.ac.cn) funded by Chinese Academy of Sciences. GoOA selects high-quality 2,174 OA journals from 125 reputable publishers, and collects 429,581 papers and 433,648 associated data sets from these journals. GoOA serves for >80,000 CAS researchers and graduate students, as well as Chinese scholars from Chinese universities. Through GoOA, We try to explore the solution of evaluation, selection, and repository of high-quality OA journals and associated data, devote to studying a sustainable utilization of open scientific literature data, including:

1) Take different acquisition strategies depending on different rights license protection levels of publishers. Publishers who provide open interfaces and structured data on websites, for instance, GoOA takes active acquisition and harvests all from web sites or API; for most publishers, a cooperation agreement we provide would be assigned, OA journals and their data would be deposited by publishers on their own to GoOA.

2）Select high-quality OA journals and their data, "quality, not quantity". OA journals are selected and collected in GoOA annually. Selected journals in GoOA will be evaluated and scored with three different-weighted indicators, including publishing quality, academic influence and openness to content. According to final scores, these journals are rated and divided into three levels. Based on the levels of journals, we can also conclude the levels of open data in the journals. Furthermore, the work of OA journal evaluation is the fruit of the combination of qualitative and quantitative methods as well as objective assessment and review by experts. The usage statistic and comments from users are also taken into evaluation system.

3）All data in GoOA is described by semantic and structured. The GoOA system embeds the domain ontology to provide navigation and retrieval of the concept of knowledge. In order to achieve intelligent knowledge discovery, GoOA stores data in structured format and organizes by semantization, describes the digital object into each chart and data unit, establishes the relations between literatures and associated data.

4) Ensure Long term preservation and aim to reuse. Because OA journals and their data is unstable, GoOA deposits all selected journals and their data locally and backups. GoOA also open all data to public, calls on researchers to reuse GoOA data (open paper, open data, open charts, etc.) to produce new products by their own data analysis tools. For example, by exploring GoOA and other sources, a thematic data set of rice cultivars is extracted by researchers for scientific discovery.

In the next step, GoOA will deposit and collect as many types of data as possible, and integrate into a data hub.

# DEVELOPMENT OF EARTH REMOTE SENSING DATA ANALYSIS TOOLS FOR VERY LARGE DISTRIBUTED DATA ARCHIVES

*E. Loupian, A. Kashnitskiy, M. Burtsev, A. Proshin, I. Balashov, V. Tolpin*

burcev@d902.iki.rssi.ru

Constant growth of Earth remote sensing data volumes available to researchers and experts makes effective data management methods more and more necessary. A lot of methods used nowadays originate from times when satellite data was expensive, hard to obtain and had low temporal resolution and coverage, i.e. it was suitable for regional but not for global monitoring tasks. Today the situation has changed. Very large Earth remote sensing data archives are accumulated and maintained in many organizations, thus making the data generally available to end users but the data management methods are still lacking effectiveness. The data user usually still has to search, download, store and process all the data himself wasting time and resources instead of focusing on the problem.

The paper proposes new data management and analysis methods developed by the authors specifically for very large and constantly updated data archives. These methods make possible development of information systems providing both access to data and data processing features just with a web browser, i.e. – fully functional web-based GIS integrated with very large data archives. The data within such systems is stored in distributed archives and all the user driven processing is performed on the storage datacenter computational infrastructure but not on the user's one. This prevents heavy network and optimizes management of large data volumes. The functionality and number of implemented data processing features is rapidly evolving and now can compete with many desktop satellite data processing software. The obvious advantages of the proposed methods include immediate access to large data volumes without necessity to download it locally, possibility to use the datacenter computing resources and possibility of performing all the data search, access, processing and analysis operations only with a web browser.

The proposed methods are successfully implemented in a number of web information systems for satellite data analysis including the VEGA-GEOGLAM (http://vega.geoglam.ru/) agriculture monitoring system developed in framework of EC FP7 SIGMA project. Other operational applications of these methods include forest cover dynamics monitoring and analysis, wildfires detection and estimation of consequences, volcanic activity monitoring and research, ocean research and many others. The work was cofunded by the RFBR (grant No. 16-37-00427 mol_a).

# EVIDENCE BASED DISASTER RISK REDUCTION BY IMPLEMENTATION OF SENDAI FRAMEWORK 2015-2030

*S. Nakamura, K. Blanchard, V. Murray*

Public Health England

Virginia.Murray@phe.gov.uk

The Sendai Framework for Disaster Risk Reduction 2015-2030 [1] was adopted by UN Member States in 2015 at the Third UN World Conference on Disaster Risk Reduction in Sendai City in 2015 [1]. The Framework is the first major agreement of the post-2015 development agenda and aims to reduce disaster losses in lives, livelihoods, and health in economic, physical, social, cultural and environmental assets of persons, businesses, communities and countries [1-3]. The four priorities stated in the framework are:

1. Understanding disaster risk;
2. Strengthening disaster risk governance to manage disaster risk;
3. Investing in disaster risk reduction for resilience and;
4. Enhancing disaster preparedness for effective response and to build back better in recovery, rehabilitation and reconstruction [2].

To achieve these targets and priorities, the DRR community needs robust scientific data, an agreement on how to collect and analyse data and access to the tools in which to facilitate those measurements [3, 4]. A scientific approach is therefore vital in order that the community meet this criteria and is crucial to build evidence based policy and practice (3, 5, 6). Furthermore, implementation of these policies will require a collaborative effort across public, private sectors, civil society organizations, academia and science research institutions [1, 4]. This collaboration will be one of key challenges to overcome whilst we implement the Sendai framework 2015-2030.

The Sendai Framework articulates the requirement for this within the following paragraph:

- "the need for improved understanding of disaster risk in all its dimensions of exposure, vulnerability and hazard characteristics…;
- the strengthening of disaster risk governance, including national platforms;
- accountability for disaster risk management;
- preparedness to "Build Back Better";
- recognition of stakeholders and their roles;
- mobilization of risk-sensitive investment to avoid the creation of new risk;
- resilience of health infrastructure, cultural heritage and work-places;
- strengthening of international cooperation and global partnership, and risk-informed donor policies and programs, including financial support and loans from international financial institutions" [1].

There is also clear recognition of the Global Platform for Disaster Risk Reduction [8] and the regional platforms for disaster risk reduction as mechanisms for coherence across agendas, monitoring and periodic reviews in support of UN Governance bodies [1]. Implementing the Sendai framework would enable more effective and coordinated disaster risk reduction management.

1. WHO. Sendai framework for disaster risk reduction 2015 - 2030. 2005.
2. UNISDR. Sendai Framework for Disaster Risk Reduction 2015 [cited 2017 3 July]. Available from: http://www.unisdr.org/we/coordinate/sendai-framework.
3. Aitsi-Selmi A, Murray V, Wannous C, Dickinson C, Johnston D, Kawasaki A, et al. Reflections on a Science and Technology Agenda for 21st Century Disaster Risk Reduction. International Journal of Disaster Risk Science. 2016;7(1):1-29.

4. Karmen P, Montserrat MF, Tom DG, Ian C. Science for Disaster Risk Management 2017: Knowing better and losing less. Publications Office of the European Union; 2017.
5. Murray V, Maini R, Clarke L, Eltinay N. Coherence between the Sendai Framework, the SDGs, the Climate Agreement, New Urban Agenda and World Humanitarian Summit, and the role of science in their implementation. 2016.
6. Aitsi-Selmi A, Murray V, Heymann D, McCloskey B, Azhar EI, Petersen E, et al. Reducing risks to health and wellbeing at mass gatherings: the role of the Sendai Framework for Disaster Risk Reduction. International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases. 2016;47:101-4.
7. UNISDR. Sendai Declaration. 2015.
8. UNISDR. Global Platform for Disaster Risk Reduction 2017 [cited 2017 24 July]. Available from: http://www.unisdr.org/conferences/2017/globalplatform/en.

# DATA-DRIVEN CLIMATE MODELING AND PREDICTION

*D. Kondrashov*

University of California, Los Angeles

dkondras@atmos.ucla.edu

Global climate models (GCMs) aim to model a broad range of spatiotemporal scales of climate variability with state vector having many millions of degrees of freedom. On the other hand, while detailed weather prediction out to a few days requires high numerical resolution, the major fraction of large-scale climate variability can be predicted in a much lower-dimensional phase space. Low-order models can simulate and predict this fraction of climate variability, provided they are able to account for linear and nonlinear interactions between the modes representing large&slow scales of climate dynamics, as well as their interactions with a much larger number of modes representing fast&small scales.

The recently introduced Multilayer Stochastic Modeling (MSM) framework [1] emphasizes the ubiquitous role of nonlinear, stochastic as well as memory effects for the derivation of data-driven low-order models with good skill in simulating and predicting main dynamical features of the targeted spatiotemporal field as an output of a nonlinear, high-dimensional geophysical (climate) model, or as a set of observations. However, if the input data are not numerous enough and exhibit mixture of different spatiotemporal scales, the analysis may reveal multiple predictors and complex model structure.

The novel time series analysis technique of data-adaptive harmonic decomposition (DAH) provides an attractive alternative that reduces the data-driven inverse modeling effort to elemental MSMs with fixed and much smaller number of coefficients to estimate [2]. The key numerical features of DAH rely on the construction of covariance matrix that exploits time-lagged cross-correlations in space. Eigenmodes associated with DAH covariance matrix form an orthogonal set of oscillating data-adaptive harmonic modes (DAHMs) that come in pairs and in exact phase quadrature for a given temporal Fourier frequency. Moreover, the pairs of data-adaptive harmonic coefficients (DAHCs), obtained by projecting the input dataset onto DAHMs, can be effectively modeled within a universal parametric family of simple nonlinear stochastic models – coupled multilayer Stuart-Landau oscillatory models (MSLM) stacked per frequency, and synchronized at different frequencies by the same noise realization. DAH-MSLM applications on didactic examples, as well as climate modeling and prediction will be presented. In all cases, the DAH decomposition allows for an extraction of spatiotemporal modes revealing key dynamical features in the embedded phase space. The DAH-MSLMs are shown to successfully model the typical patterns of the corresponding climate fields, as well as their key statistics.

In particular, DAH-MSLM has been successfully applied to a difficult problem of modeling and prediction of Arctic sea ice data [3]. Decline in the Arctic sea ice extent (SIE) due to global warming has profound socio-economic implications and is a focus of active scientific research. SIE has dramatically decreased over recent decades, and especially in summertime with extreme minimum events in recent years. Of particular interest is prediction of September SIE on subseasonal time scales, i.e.~from early summer into fall, when sea ice coverage in Arctic reaches its minimum. Forecasting of September SIE is very challenging due to the high variability of ocean and atmosphere over Arctic in summer, as well as shortness of observational data and inadequacies of the physics-based models to simulate sea-ice dynamics. The real-time DAH-MSLM prediction outperformed most statistical models and physics-based models in 2016 Sea Ice Outlook [4] — a collaborative effort to facilitate and improve subseasonal prediction of September SIE: $4.79 \cdot 10^6$ km$^2$ vs. actually observed $4.7 \cdot 10^6$ km$^2$. The key success factor is associated with DAH-MSLM ability to efficiently disentangle and model

complex spatiotemporal dynamics of SIE.

References

1. D. Kondrashov, M. D. Chekroun, and M. Ghil, Data-driven non-Markovian closure models, Physica D, 297, 33–55, 2015.
2. M. D. Chekroun and D. Kondrashov, Data-adaptive harmonic spectra and multilayer Stuart-Landau models, preprint: arXiv:1706.04275, 2017
3. D. Kondrashov, M.D. Chekroun, X. Yuan, M. Ghil, "Data-adaptive Harmonic Decomposition and Stochastic Modeling of Arctic Sea Ice" in Advances in Nonlinear Dynamics by Springer Nature, Ed. A. Tsonis, doi:10.1007/978-3-319-58895-7.
4. Sea Ice Outlook: https://www.arcus.org/sipn/sea-ice-outlook/2016/post-season.

# CHALLENGES IN REGIONAL COLLABORATIONS FOR DATA SCIENCE - INSIGHTS FROM THE BIOLOGICAL METAGENOMICS PROJECT

*K. Mineta*

King Abdullah University of Science and Technology

katsuhiko.mineta@kaust.edu.sa

Recent advancement of the biological technologies such as the new generation of the sequencing devices provides a massive amount of the data to the researchers. The big data is now a common topic among biological research as well as other research fields. To enhance the research activities in biology, regional collaboration for big data is now an essential part of the research projects. We are currently conducting a big data project, performing marine metagenomics of the Red Sea that is surrounded by the African continent and the Arabian Peninsula. Due to its unique features such as high temperature and salinity, the Red Sea is an excellent resource for the bioprospecting such as hunting the novel and useful genes and gene products. Comparing the data from the different regions as collaboration is the most efficient way to find those genes unique to the Red Sea. Thus, regional collaboration for data exchange and sharing is essential to our research project. Accordingly, the database is a key to this collaboration. In this presentation, the challenges in the regional collaboration will be discussed through our current research activities and experiences.

# A GLOBAL MAP OF AGRIFOOD DATA STANDARDS

*V. Pesce*

Global Forum on Agricultural Research

valeria.pesce@fao.org

The GODAN Action project, a three-and-a-half-year programme launched by the UK's Department for International Development, has published a global map of data standards relevant to the exchange of agri-food data. Quoting what the Dublin Core Metadata Initiative says about their DCMI Registry, the reuse of existing vocabularies or classes / properties therein is essential to standardization, and promotes greater interoperability between vocabularies and datasets. The discovery of existing vocabularies is an essential, and prerequisite, step in this process. This map promotes the wider adoption, standardization and interoperability of vocabularies by facilitating their discovery and reuse. In addition, it provides a useful overview of what exists and helps identify overlaps, duplication, gaps and limits to adoption, hopefully encouraging practitioners not to duplicate efforts and to collaborate to both develop and use common standards.

The map was published under the Agrisemantics domain because of its purpose being very close to the scope of the Research Data Alliance Agrisemantics Working Group. The map is available at http://vest.agrisemantics.org.

The map builds on two existing portals: the AIMS VEST Registry of FAO (a metadata registry, now fully merged with the new map) and the AgroPortal of University of Montpellier / Stanford University (an RDF vocabulary repository, in turn building on the NCBO BioPortal).

Regarding the types of data standards covered, within the scope of this map by "data standards" we mean "vocabularies" tout court, in the broad sense in which vocabularies are defined by W3C, including schemas (metadata element sets, models, ontologies) and value vocabularies (sets of controlled values, from code lists to thesauri to authority data). In addition, while we do not include file formats among the data standards, we do include some non-semantic data formats that are very much used and considered full standards in certain disciplines or for certain types of data (e.g. NetCDF for observations). Otherwise, file formats are used as annotations for the data standard, both for the format in which the data standard is serialized and for the format in which the data using the standard is expected to be serialized.

Regarding the domains and the types of data covered, since the agri-food sector spans across several disciplines (plant sciences, farming systems, natural resources management, forestry, all disciplines and activities involved in the food supply chain...) and is also closely interlinked with neighboring disciplines (climate, environment, geospatial, biology...), we included standards covering all of these disciplines. For the description of data standards, the core properties used in the map were drawn from existing standards for describing vocabularies (VOAF, VOID, VANN, OMV, DCAT). Besides those essential properties, we added properties that support the evaluation of the quality, usability suitability and openness of the featured standards.

For the categorization of the data standards, the main categorizations (domain, type of data, format, license) are based on existing classifications or authority lists whenever possible, like FAO and USDA domains, GODAN AgPack types of agri-food data, IANA and W3C formats, the Open Definition list of licenses).All metadata are exposed as a triple store, using the above mentioned vocabularies, assigning URIs to all categorization terms and linking them to external concepts whenever possible. For vocabularies that are also hosted at the AgroPortal, which is a repository of RDF vocabularies, queries can be extended to the full RDF of the vocabulary. We see this map as a key infrastructural component for improving the interoperability of agri-food data and we work on ensuring its sustainability, so that it is maintained beyond the end of the project, as a common asset of the community working with agricultural and nutrition data.

# WEB-BASED EARTH OBSERVATION DATA ANALYSIS SYSTEM VEGA-GEOGLAM IN SUPPORT OF GLOBAL AGRICULTURAL MONITORING RESEARCH AND DEVELOPMENT

*Sergey Bartalev, E.Loupian, V. Tolpin, E. Elkina*

Space Research Institute of the Russian Academy of Sciences (SRI RAS, Russia)

bartalev@smis.iki.rssi.ru

The web-based land cover and vegetation monitoring service VEGA (http://sci-vega.ru/), in operation since 2011, has been developing by the Space Research Institute (IKI) to provide the monitoring facilities based on Earth observation (EO) data. Based on the VEGA technological platform and infrastructure with support of the EC FP7 SIGMA project and Russian Ministry of Education and Science (project ID is RFMEFI61615X0063) the web-service VEGA-GEOGLAM (http://vega.geoglam.ru/) was build. VEGA-GEOGLAM is a global agricultural monitoring service aimed to perform cropland mapping and assessment using EO and in-situ data analysis over JECAM (Joint Experiment of Crop Assessment and Monitoring, http://www.jecam.org/) test-sites. The system is based on the concept of geospatial information web-service, which gathers satellite and other geospatial data from different sources and provides a worldwide user access to them.

VEGA-GEOGLAM provides an access to near-real-time updated EO data collected over JECAM test-sites. Users have an access to high resolution (e.g. Landsat, Sentinel-1&2) data as well as to moderate resolution (MODIS, Proba-V) data along with various derived products. MODIS surface reflectance products and Landsat-TM/ETM+ images are available for a period starting from year 2000 with continuous daily update. MODIS data pre-processing automated chain produces time-series of weekly cloud-free image composites and various vegetation indices. The Landsat imagery is also a subject of pre-processing routine for radiometric calibration, cloud and cloud-shadow screening, as well as image compositing. Being daily updated both these datasets are available for users through web interface together with thematic products and data analysis tools.

MODIS derived thematic maps available through the VEGA-GEOGLAM web-service have a strong focus on vegetating cover status and dynamics and include products such as e.g. land cover, arable lands, crop types and some other maps. Depending on the thematic content the products are updated daily, weekly or annually in automatic or semi-automatic mode, which together with the advanced analysis tools provides an opportunity to evaluate land cover dynamics in various aspects. The data analysis tools available on the VEGA-GEOGLAM service allow extracting multi-annual temporal profiles of NDVI (or other vegetation index) in an "on-the-fly" manner for any area of interest indicated by user. The service gives a rapid access to all available data for the certain area as well as tools for land cover change visualization using selected multi-temporal data. These and other visualization, measurement and image classification tools provide an unique opportunity for integrated analysis of moderate- and high-resolution data highly exploiting their complementarities in terms of temporal frequency and spatial accuracy of different EO instruments.

The VEGA-GEOGLAM system is used for global monitoring of agricultural production and yield forecast in the framework of the GEOGLAM Crop Monitor initiative (https://cropmonitor.org/).

# TOWARDS AN E-INFRASTRUCTURE FOR OPEN SCIENCE IN AGRICULTURE

*O. Hologne, M. Huber*

French Institute for Agricultural Research (INRA, France)

odile.hologne@inra.fr

The European project e-ROSA (Towards an e-infrastructure Roadmap for Open Science in Agriculture) seeks to build a shared vision of a future sustainable e-infrastructure for Open Science in agriculture & food. It aims at facilitating the co-development of a common roadmap by and for involved research communities and key stakeholders related to scientific data and research infrastructures, in line with the EOSC vision, agenda and architecture.
An e-infrastructure is seen as a combination of digital technologies (hardware and software), resources (data, services, digital libraries), communications (protocols, access rights and networks), and the people and organisational structures needed to manage them.
The agri-food sector is dealing with an increasing amount and variety of data due to:

- The multidisciplinary nature of agri-food science, which is adopting a more and more systemic approach;
- The automation of data collection thanks to robots, sensors, etc., as well as new engineering tools such as in the omics field;

The development of new types of data sources and providers: e.g. Internet of Things, citizen science, voice- and image-based applications, micro-blogging, etc.
In addition, we have increasing computational capabilities for data collection and analysis, which can support more efficient knowledge generation and decision-making. As such, we today have the opportunity to rely on more complex and integrated reasoning and modelling, requiring the integration of these numerous, dispersed and heterogeneous data.
Many initiatives and infrastructure services already exist and need stronger and coordinated support to promote Open Science for agri-food and implement the European Open Science Cloud (EOSC). There is a pressing need for the development of a common e-infrastructure framework in order to:

- connect data and connect infrastructures,
- integrate existing initiatives into a common framework at a global level,
- share efforts and resources,
- support a collective change of practices through the adoption of shared standards,
- provide a pre-competitive space for sharing data and speeding up innovation processes.

Various issues need to be taken into account when envisioning an e-infrastructure, including:

- The articulation between general e-infrastructure issues (technical and non-technical) with specific scientific data-related needs (e.g. thanks to Virtual Research Environments);
- The easy access to and use of the e-infrastructure by researchers;

Challenges linked to a distributed organisation (i.e. in nodes):

- Agree on data access and interoperability policies;
- Secure sustainable funding for long-term common resources;
- Agree on an efficient division of labour to maximise total impact;
- Synchronise technical updates (e.g. at a local level, when implementing improvements of a common data model);

The identification of needs and services to be covered by the e-infrastructure;
The evaluation of an e-infrastructure (e.g. assess impact on change in practices). A shared vision of what could be this e-infrastructure is crucial, and the development of demonstrators in addition to a better understanding of the data ecosystem could be useful to build it, facing technical and cultural challenges.

# NEW QUALITY OF PREDICTIVE ANALYTICS BY USING NEW DATA SOURCES: SOLUTIONS AND EXPERIENCE

*M. Ageykin, S. Shvey, V. Shvey*

EC-leasing

mageykin@ec-leasing.ru

According to IBM strategic forecast, all companies in the next 5 years will be divided into winners and losers depending on quality of making corporate decisions. Research and case studies provide evidence that a well-designed and appropriate computerized decision support system can encourage fact-based decisions, improve decision quality, and improve the efficiency and effectiveness of decision processes. There is resource that we all have aplenty: a large amount of open data, both structured and unstructured. This report introduces the concept of acquiring data from big data sources such as social media, news, mobile and smart devices, weather information, and information that is collected via sensors and using this data to get new quality of predictive analytics. Predictive analytics today in many companies is perceived as an evolutionary step in business analytics and is used primarily to build a forecast based on the same data on which reports are built. Nevertheless, this does not take into account the enormous importance of external factors in forecasting and nowcasting. In this report will be shown cases from various areas where external data are the basis for predictive analytics and allow gain results unattainable to the forecast based only on enterprise data. Multiple scenarios for storing, handling, preprocessing, filtering, and exploring big data in predictive analytics are described. Techniques are described that help to form homogeneous (homogeneous) groups of data, based on data from various sources. An important part is working with text documents and gaining information for use in predictive modeling. At the end of the report you can see examples of predictive models based on external data which allow the companies using them to generate additional profits and outrun their competitors.

# DATA ANALYSIS, EVENT RECOGNITION AND APPLICATIONS

*E. Uchaikin*[1], *N. Kudryavtsev*[1], *D. Kudin*[2]

[1] Gorno-Altaisk State University,
[2] Geophysical Center of the Russian Academy of Sciences (GC RAS, Russia)

dvkudin@gmail.com

The paper describes an algorithm that estimates thunderstorm activity within a radius of up to 2500 km with help of local station of the WWLLN Network. The determination of the thunderstorm activity index of the selected WWLLN station is based on the image processing of spectrogram of the VLF antenna signal, which generated every minute. The presence of high-amplitude signals in the frequency range 5-15 kHz indicates the reception of a large number of atmospherics on the antenna, while the signals of atmospherics are 3500-12000 km is recognized by the built-in software of the WWLLN network with output data to "R" text files. The proposed algorithm uses "R" text files to ignore long-range strikes when analyzing the spectrogram. Additionally, the local meteorological data and the data of a horizontal long-wire electric antenna are used, to refine the final estimation. The algorithm designed to real-time work on web site with ability to notify about high thunderstorm activity.

# PRIORITIES FOR UNDERSTANDING ARCTIC OCEAN/ATMOSPHERE/CRYOSPHERE SYSTEMS

*G. Boulton*

University of Edinburgh & CODATA

G.Boulton@ed.ac.uk

The fast pace of warming in the Arctic is associated with major changes in the coupled ocean-atmosphere-cryosphere system, which is the subject of intensive scientific monitoring and modelling. Understanding these changes is important because of their role in the global climate system, because of impacts on regional communities and because of economic opportunities. Much research in this domain is directed towards improving our understanding of the impact of a changing Arctic on Northern Hemisphere weather and climate; the safety & wellbeing of people in the Arctic and across the Northern Hemisphere; reducing the risks associated with Arctic operations and resource exploitation; and supporting evidence-based decision-making by policymakers worldwide. Although there is great focus on efficient retention of and access to the resultant large data volumes now being acquired, the efficient use of this data is sub-optimal because of limited capacity for data integration.

# SYSTEM OF INTEGRATED MONITORING OF CATASTROPHIC PHENOMENA BASED ON SATELLITE AND GROUND DATA

*V. G. Bondur, M. N. Tsidilina, E. V. Gaponova, K. A. Gordo, O. S. Voronova*

Institute for Scientific Research of Aerospace Monitoring AEROCOSMOS (Russia)

vgbondur@aerocosmos.info

In view of growing annual number of natural and technogenic emergencies, increase in their scale, growth of human losses and economic damage, the issues of prevention and mitigation has become particularly urgent. Over the past 50 years, the number of such catastrophic phenomena has increased by 3 times, and caused damage has increased by 9 times [http://geoenv.ru]. Natural catastrophes cause explosions, fires, floods, destruction of residential and industrial buildings, dam failure, etc. In this regard, one of the most important tasks of ensuring the safety of the population is early detection, prediction and prevention of catastrophic natural phenomena such as earthquakes, tsunamis, wildfires, volcanic eruptions and typhoons [Bondur, 2015; Bondur, Ginzburg 2016; Bondur et al., 2016; Bondur, Tsidilina, et. al., 2017; Bondur, Gordo, 2017; Bondur, Voronova, 2017].

Significant progress achieved in observing various natural disasters, including those recorded by satellite methods, as well as historical data sets enable a conscious approach to the creation of integrated monitoring systems for catastrophic phenomena and give optimistic prospects of their practical application.

At present, ISR“AEROCOSMOS” is developing a system of integrated monitoring of catastrophic phenomena to reduce their effects based on satellite and ground data. An important element of the system performance is satellite monitoring instruments, which provide necessary information about the key parameters of the environment with the required regularity and large spatial coverage [Bondur, 2014, 2016].

The key principle of building the mentioned system is also the ability to comprehensively analyze satellite and ground data with various spatial and temporal resolution for monitoring fires, seismic areas, volcanic eruptions and typhoons, as well as their consequences that are often catastrophic.

The system of integrated monitoring of catastrophic phenomena based on satellite and ground data is built on the basis of a multilevel hierarchical principle through the functional integration of data flows coming from various satellites equipped with remote sensing instruments, from ground stations and from scientific data archives, and processed data [Bondur, 2014].

The use of satellite and ground data, as well as methods for their integrated processing and analysis aimed to monitor catastrophic phenomena is a challenging task. The solution of this task will allow to combine and systematize diverse data that will provide an opportunity for online monitoring and prevention of dangerous catastrophic phenomena.

Acknowledgements

This research was performed under the grant from the Russian Science Foundation (project No 16-17-00139) at ISR «AEROCOSMOS»

Keywords: integrated monitoring system, satellite data, ground data, satellite monitoring, earthquakes, wildfires, typhoons, volcanic eruptions.

References

1. Bondur V.G. Satellite monitoring of trace gas and aerosol emissions during wildfires in Russia // Issledovanie Zemli iz kosmosa. 2015. No6. p.21-35.

2. Bondur V.G., Ginzburg A.S. Emission of Carbon-Bearing Gases and Aerosols from Natural Fires on the Territory of Russia Based on Space Monitoring // Doklady Earth Sciences. 2016. Vol. 466. No. 2. P. 148-152. DOI 10.1134/S1028334X16020045.
3. Bondur V.G., Tsidilina M.N., Gaponova E.V., Voronova O.S. Joint analysis of various precursors of seismic events using remote sensing data at the example of earthquake in Italy (24.08.2016, M6.2) // 17TH INTERNATIONAL MULTIDISCIPLINARY SCIENTIFIC GEOCONFERENCE SGEM 2017 Conference proceedings. 2017. pp. 149-163.
4. Bondur V.G., Gordo K. Satellite monitoring of wildfires and their effects in the northern Eurasia // 17TH INTERNATIONAL MULTIDISCIPLINARY SCIENTIFIC GEOCONFERENCE SGEM 2017 Conference proceedings. 2017. pp. 227-239.
5. Bondur V.G., Voronova O.S. Using remote sensing data to monitor volcanic activity: Mount Etna case study //17TH INTERNATIONAL MULTIDISCIPLINARY SCIENTIFIC GEOCONFERENCE SGEM 2017 Conference proceedings. 2017. pp. 339-347.
6. Bondur V.G., Zverev A.T., Gaponova E.V. Precursor variability of lineament systems detected on satellite images during strong earthquakes // Issledovanie Zemli iz kosmosa. 2016. No 3. p.1-10.