# Gap Analysis on Open Data Interconnectivity for Global Disaster Risk Research

CODATA Task Group

Linked Open Data for Global Disaster Risk Research (LODGD)

Date: May 20, 2016

Version 16.0

# Study Panel

## Contributing Authors

Brenda K. JONES, USGS EROS(Earth Resources Observation and Science) Center,USA

Carol SONG (Co-Chair), Purdue University,USA

Edward T.-H. CHU, National Yunlin University of Science and Technology, Taiwan of China

FAN Jinlong, National Satellite Meteorological Center,China

HUANG Shifeng, China Institute of Water Resources and Hydropower Research,China

HUANG Mingrui, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences,China

LI Guoqing (Co-Chair), Institute of Remote Sensing and Digital Earth, CAS,China

LI Xiaotao, China Institute of Water Resources and Hydropower Research,China

Michael RAST, ESA-ESRIN(ESA's Centre for Earth Observation),Italy

Masaru YARIME, The University of Tokyo, Japan.

QING Xiuling, National Science Library, Chinese Academy of Sciences,China

Susan L. CUTTER, University of South Carolina,USA

Siyka ZLATANOVA,Delft University of Technology, The Netherlands

XIE Xinlu, Chinese Academy of Social Sciences,China

ZHANG Hongyue, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences,China

## Review Panel

To be announced

## Contact Information

Review comments and suggestions are greatly appreciated. Please send them to:

LI Guoqing (ligq@radi.ac.cn)
Carol SONG (carolxsong@purdue.edu)
ZHANG Hongyue (zhanghy01@radi.ac.cn)

# Table of Contents

# 1 Open and Linkable: New Strategies of International Science Data Management[1]

Disasters are sudden calamitous events that bring great damage, loss or destruction to large populations and regions. They are most often caused by natural hazards such as flood, hurricane, fire, earthquake, etc., and their damaging impacts on people's lives and properties are often aggravated due to inappropriately managed risks. Disaster research aims to better understand the process and interaction of various natural phenomena among themselves and with human activities, to better quantify vulnerability and risks, and to gain and disseminate knowledge to aid decision-making in reducing risks and helping people cope with disasters and their aftermath.

More and more research domains are becoming data driven, especially in the face of vastly improved technology and infrastructure that can collect and make huge amounts of data available. Similarly, research on disasters relies heavily on scientific data, including both observations, analysis and simulation data, that are multidisciplinary, heterogeneous, and dispersed across institutional and country boundaries. Increasingly diverse sources of data, including unstructured data such as information from communications, social media, etc., are also beginning to play an important role in disaster studies. The need for open data and data interconnectivity, i.e., accessible and usable by researchers, decision makers and the public, is nowhere as critical as in the area of disaster research, management and mitigation.

Reports (Munich Re,2004) have indicated that more natural disasters have occurred in the past sixty years and that the economic and societal impact of disasters has climbed up by five times in the same period of time. In-depth analysis of the current state of disaster scientific data management and acquisition patterns indicates a greater need for interconnection of dispersed scientific data related to disaster risk assessment and mitigation. Today, large amounts of disaster related scientific data exist, such as data from monitoring equipment, base maps, evaluation, progress, socio-economic statistics, and so on. They are typically dispersed geographically and owned by various government agencies, research centers, groups and, sometimes, individuals around the world. Researchers often find it difficult, if not impossible, to discover relevant data needed for their study. Even when they identify the data sources, they may not be able to obtain the data due to ownership issues, or the lack of tools to successfully select, transfer, interpret and use the data with their applications.

Gaps in data infrastructure, data sharing policies and data use governance must be addressed to unleash the potential of disaster research in helping regions, especially the developing countries, to improve risk assessment, reduction and mitigation. Two related areas are being studied: *open data*, and *data interconnectivity*. The data-driven nature of disaster research demands open access to scientific data, as it is impossible to fully understand the cause and impact of a disaster event without consulting multiple types of data. In addition to open data, disaster researchers face perhaps a greater challenge –to find relevant data sets in a "sea" of distributed and disparate data resources. The next generation data infrastructure must provide linkage of data, helping researchers to find relevant data across distributed data holdings.

As stated in the Sendai Framework, disaster risk reduction requires a multidisciplinary approach and decision-making based on the open exchange and dissemination of disaggregated data. It is urgent to enhance the scientific and technical work on disaster risk reduction and its mobilization through the coordination of existing networks and scientific research institutions at all levels and in all regions. In answering this call to the science community, it is of utmost importance to promote and practice the collection, management, opening up and sharing of scientific data related to disaster risk research, as well as the employment of relevant technologies and applications, consistently and globally. This

[1] Please refer to the ICSU international accord on open data for definition (Appendix A)

white paper, supported by ICSU and CODATA communities, aims at systematically analyzing the needs for the data infrastructure proposed in the Sendai Framework and providing the conceptual building blocks to help realize the Sendai imperative.

Our vision for the next generation disaster risk research data infrastructure is an interconnected, collective repository of observational and derived disaster-related data that is open, discoverable, and easily accessible and usable by all, enabled by the revolutionary digital technologies today and open access policy embraced by users and providers.

This paper aims at identifying the gaps in technology and relevant policies that prevent effective interconnection of disaster related data and information for use in research, education and public engagement. It examines the current state of information technology for data management and sharing, as well as policies regarding data availability at various levels, and discusses potential solutions and examples toward open data and data interconnectivity for disaster research.

## 1.1 Background

The many benefits of open data, especially scientific data, both observational as well as simulation and analysis data, have long been recognized. Open data typically refers to data that is available to anyone who wishes to view, analyse and utilize. As early as 1957, the World Data Centre system was established to support open access to scientific data collected from the observational programs of the 1957–1958 International Geophysical Year. Originally established in the United States, Europe, Russia, and Japan, the World Data Centre system has since expanded to other countries and to new scientific disciplines. Its holdings include a wide variety of data that cover timescales ranging from seconds to millennia. These data provide baseline information for research in many disciplines, especially for monitoring changes in the geo-sphere and biosphere.

While the idea of open science data has been actively promoted, the rise of the Internet, web services, and the declining cost of computing and storage hardware has significantly lowered the barrier to publish or obtain data, bringing us much closer to a reality of broadly sharing scientific data to help improve society and people's lives.

In 1995 GCDIS (US) stated its position clearly in On the Full and Open Exchange of Science data (a publication of the Committee on Geophysical and Environmental Data - National Research Council): "The Earth's atmosphere, oceans, and biosphere form an integrated system that transcends national boundaries. To understand the elements of the system, the way they interact, and how they have changed with time, it is necessary to collect and analyze environmental data from all parts of the world.

Studies of the global environment require international collaboration for many reasons: (a) to address global issues, it is essential to have global data sets and products derived from these data sets; (b) it is more efficient and cost-effective for each nation to share its data and information than to collect everything it needs independently; and (c) the implementation of effective policies addressing issues of the global environment requires the involvement from the outset of nearly all nations of the world. (d) International programs for global change research and environmental monitoring crucially depend on the principle of full and open data exchange (i.e., data and information are made available without restriction, on a non-discriminatory basis, for no more than the cost of reproduction and distribution) "(Li Juan et al,2009).

The International Organizations advocating open data strategies include the Organisation for Economic Co-operation and Development (OECD), and some of the specialised agencies of the United Nations, such as the United Nations Educational, Scientific, and Cultural Organisation (UNESCO), the International Council for Science (ICSU), the interdisciplinary Committee on Data for Science and Technology (CODATA), the InterAcademy Panel on International Issues (IAP), and the Academy of Sciences for the Developing World (TWAS). GEO, International Charter, IRDR, and Future Earth. They have proposed their respective principles on how to access and share the data.

(1) GEO data sharing principle:

- There will be full and open exchange of data, metadata and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation.
- All shared data, metadata and products will be made available with minimum time delay and at minimum cost.
- All shared data, metadata and products being free of charge or for no more than cost of reproduction will be encouraged for research and education.

(2) The International Council for Science (ICSU) has successively established the World Data Centre (WDC) and the Committee on Data for Science and Technology (CODATA) for information collection, exchange, service and sharing. ICSU is devoted to researching, observing and evaluating relevant data and information as well as their relations with decision-making, with open access of data comprising an important aspect; CODATA is committed to uplifting the quality, reliability, management and accessibility of data that holds great significance for the whole scientific community. The report on CODATA Strategy shows that data sharing holds a very important position in CODATA's strategic planning:

The mission of CODATA is to strengthen international science for the benefit of society by promoting improved scientific and technical data management and use. Looking across and beyond specific scientific programs such as IRDR and Future Earth, CODATA is well positioned to promote coordination with key international initiatives and programs such as GEO, the European Union's Global Research Data Infrastructures (GRDI2020) project, the Eye on Earth initiative, the WSIS implementation and follow-up process, and the proposed Data Web Forum. For example, there are many potential synergies between the development of the GEOSS Disaster Societal Benefit Area and IRDR data needs, and between a range of GEOSS data and services and Future Earth activities. CODATA could play a lead role in harmonizing data policies, improving data access and interoperability, and developing long-term strategies for data stewardship.

(3) CODATA promotes direct cooperation among scientists and engineers by facilitating their participation in international data activities, has basically established a worldwide exchange system of science data via Internet, organizes and supports taskforces, work teams, committees and groups working on specific data issues to carry out international cooperation. The plan on implementation of data-sharing policy concerns two aspects:

- Policy and Institutional Frameworks for Data, including the task of "Establish Data Policy Committee (DPC), Organize forum on open access, Organize forum based on OECD guidelines and principles, Expand data policy role in Future Earth, IRDR,and other ICSU programs and initiatives" ;
- Data Strategies for International Science, including the task of "Help IRDR develop and implement a coordinated data strategy, Help Future Earth develop and implement a coordinated data strategy and Formalize relationships with key international data initiatives"

## 1.2 Open Data Strategies for Natural Disaster Research

Natural disasters result from interaction between different spheres of the earth and human activities. Ever since the time humans began to inhabit the earth, they have been subject to ceaseless confrontation against natural disasters and constantly draw lessons from their experience in coping with disasters. As people, countries, and the world become increasingly connected in many ways, understanding and responding to natural disasters has also become a global issue. The data-driven nature of the scientific research on disasters demands open access to data as it is impossible to fully

understand the cause and impact of a disaster event without consulting multiple types of data. We submit that international cooperation based on open data is crucial to the advancement of scientific research on disasters.

Numerous technological advances in instrumentation and computation have enabled the observation and recording of various data before, during and after each disaster, creating a valuable database of information for research into natural disasters. The most recent improvement in satellite observation, aerial photo grammetry, in development of ground observation stations and advancement of various measurement instruments have significantly improved our ability to obtain information on disasters and increased our capacity to keep the exponentially growing amounts of data. This accelerated growth in data size is due both to the higher ability of data collection and to the increased number of disaster events in recent years.

The data challenge for disaster researchers comes from several aspects: the massive quantity of data, the distributed nature of these data, the heterogeneity and diversity of the data. Compounding these challenges is the lack of data sharing. Due to both policy and technology limitations, it is often difficult to share and access data across disciplines, organizations, and distant geographic locations. The reality of utilizing all the relevant data to better understand, respond to and mitigate disasters is closer today than ever before – progress in cyberinfrastructure (e.g., computing, data management/federation/movement) as well as maturation of the Internet and web technologies have laid the groundwork for a framework of open access data to support disaster research and other relevant stakeholders. It should be recognized that open access and sharing of data is more than a concept; it needs to be implemented at both policy and technology levels, thus requiring collaboration among disciplines and cooperation among international groups and organizations.

Sharing of science data on disasters is realized through cooperation among states, institutions and organizations, which has received a lot of concern from international organizations and governments. International data organizations and disaster research entities show keen interest in open access of data; they have proposed relevant strategic plans and enhanced contact with policy makers and research institutes to implement the plans. Among them,

(1) The International Charters for Space and Major Disasters, an entity promoting the policy of open access of data, aims to authorize users to provide the only spatial data access and payment system to areas suffering from natural or artificial disasters. Each member organization pledges to provide necessary resources in support for implementation of the rules set in the Charter, to help reduce losses of human lives and properties in the wake of disasters, and thus to save valuable time and provide an efficient means for emergency access to data on unexpected natural disasters.

(2) With regard to disaster research, ICSU, ISSC,UNISDR and RADI are jointly engaging in a 10-year project, *Integrated Research on Disaster Risk (IRDR*); their working group on disaster loss data is devoted to sharing of multidisciplinary data on disasters, including collection, storage and dissemination of data related to disaster loss.

(3) The *Integrated Risk Governance Project* (IRG) is a research plan of Future Earth, which aims to enhance global risk management capability through a ten-year effort. Achieving the goal necessitates application of data from multiple disciplines, such as weather, climate, ecology, hydrology, geophysics and environmental science.

Promoting open access and sharing of even more disaster related data has become part of the core contents in various joint international research plans and coordination mechanisms on dealing with disasters.

## 1.3 The Future: From Open Data to Data Interconnectivity

The future of data-driven science lies in the open and easy access of data as science is becoming increasingly more collaborative, both within and across disciplines and geographic boundaries of states and countries. The terms Open Data and Data Interconnectivity emphasize two different aspects of data sharing. The term "open data" refers to the availability of data, which should be freely available to all, while the term "data interconnectivity" emphasizes accessibility and usability, that open data should also be easily accessible by all and usable in modelling, further analysis, validation, etc. The value of science data manifests in its application. Compared with special access to exclusively held data, open access of data promotes and enables data reuse and repurposing across disciplines. With the campaign for open access of data going on, the ratio of exclusively held data will be in steady decline while open access data will increase, with the ultimate emergence of a worldwide open and accessible Internet space of science data. Currently, there is no unified definition of "open data" which is interpreted differently by various organizations and institutions, for example, the Open Data Centre Alliance regards open data as the Company's IT infrastructure, or an application mode and solution of cloud computing; Scholarly Publishing and Academic Resources Coalition (SPARC) regards open data as a new mode and notion of academic publishing of science data; the open data campaign advocated by W3C adopts the RDF data model to establish RDF links among data entities of different types and sources, so as to guide users through ordinary HTML web pages and structured data by specific semantic web explorers or search engines, and finally enable all to have free access to their desired data. The definition embraced by each group or domain serves its specific needs. However, we see "open data" as a guiding principle, both philosophically and practically, that scientific data should be freely accessible by anyone without restrictions such as copyright, patent and other mechanism.

Open access of scientific data and linkage of open data represent a natural progression of data sharing. Open data access serves as the foundation of data sharing: the democratization of data – access by all without restrictions. Technology implementation is needed to make it easy for users, especially the non-domain experts, to obtain the data. Once open access is achieved, usability becomes the central issue: what does one need in order to make use of the data? Data users face a number of challenges because of complexity in data collection instrumentation and the physics represented by the acquired data, diversity in data formats and tools that deal with the variety of formats, difference in the spatial resolution of the data, and so on. On top of these, researchers face difficulties when dealing with data from a different discipline because of differences in vocabularies employed by different disciplines in describing data. Our concept of *data interconnectivity* focuses on the interconnection of datasets, bringing together data from diverse sources, disciplines and data providers to enable interoperability to support disaster research and applications.

In the area of disaster research, the international community has come to a consensus on the openness of relevant data to support disaster research, and, as a result, many data banks are available now. Due to reasons related to technology, policy and culture, various data are yet to be effectively connected, which has led to the low utilization rate of open data.

The Linked Open Data for Global Disaster Risk Research (LODGD) Working Group of CODATA aims to promote the interconnection and cross-domain utilization of disaster data.

The idea of LODGD is initialled from the experience of the well known GEO data sharing principle contributed by CODATA, also based on the experience of the technical implementation on distributed data linking. LODGD is started from Earth Observation community and now extending much widely to cross-disciplinary communities. The LODGD Task Group will study the mechanism for connecting dispersed disaster related science data to enable easier and faster discovery, access and to significantly reduce the barriers that researchers are facing today due to limited interconnection of various existing disaster-related data.

There are two levels in LODGD concept model: data characterization (lower level) and data connection (upper level). At the lower level is the knowledge about disaster taxonomy and data dependency on disaster events. As a scientific study, this level aims to understand and present the correlation between specific disaster events and science data though integration of literature analysis and semantic knowledge discovery. The upper level deals with technical methods to connect the distributed data resources identified by the lower level knowledge given a specific disaster type.

# 2 Status of Open Data for disaster research

It is necessary to clearly understand the status of open data that is relevant to disaster research before we analyze the gaps toward a comprehensive and interconnected data infrastructure. In this section, we describe the technical connotation of the concept of open data and examine the current state of open data in a number of scientific disciplines.

## 2.1 Issues related to open data

### 2.1.1 Data Accessibility

Accessibility is the driving force toward the landscape of open science data. Data sharing becomes a reality when there are few or no barriers to obtaining the data by experts and non-experts alike.

More than a decade ago, the issue of open access to research literature was discussed and debated. The work from initiatives such as Budapest Open Access Initiative (February 2002), Bethesda Statement on Open Access Publishing (June 2003), and Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (October 2003) directly prompted open access to scientific articles. The view held by this community is as follows:

There are many degrees and kinds of wider and easier access to research literature. By "open access" to this literature, it means its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited. Here's how the Bethesda and Berlin statements put it: For a work to be OA, the copyright holder must consent in advance to let users "copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship."

The "open access" movement initially focused on research literature, such as journal publications, and emphasized on barrier-free reuse of, reference to, revision, coping or and dissemination of scientific or research data or taking some measures to simply ensure users' better access to data. This concept has since been broadened to include scientific data, observational or derived, especially data that has been obtained through publicly funded research.

In the past few years, ICSU began to discuss access to data and information, and in particular, science data and information, as reflected in ICSU's 2013-2017 Strategy Plan.

The long-term ICSU vision is for a world where science is used for the benefit of all, excellence in science is valued and scientific knowledge is effectively linked to policy-making. In such a world, universal and equitable access to high quality science data and information is a reality and all countries have the scientific capacity to use these and to contribute to generating the new knowledge that is necessary to establish their own development pathways in a sustainable manner. Science is

increasingly dependent on access to, and integration of, large data-sets from multiple sources. Both the access to and the interoperability of these data sets present major challenges. Maintaining and expanding a quality-assessed science data bank is a multi-faceted challenge that must be addressed if we are to make scientific progress in areas such as earth system sustainability research.

Access to data and information is critical to the whole scientific enterprise. In particular, it is a rate-limiting step for scientific development in many poorer countries. A World Data System is being established to ensure long-term stewardship and open worldwide access to essential data-sets and data products.

The first session of the United National Environment Assembly of the United Nations Environment Programme concluded that the state of environment should stick to the following core principles:1)enabling open access to data resources from the government, research projects, non-government organizations, communities and traditional knowledge; 2)sharing data for multiple purposes; 3) reliable data and information management; 4) archiving and tracking the use of data and information; 5)support the public's access to data and information by various means.

The requirement for data accessibility has evolved from the narrow connotation of making data available based on agreed-upon inter-organizational protocols a decade ago. Today, data accessibility almost always means timely access over computer networks. Any implementation of open data access will have to support online, on-demand access methods, as well as access through web services and programmable application interface (API).

### 2.1.2 Data Sharing

Data sharing is the practice of a data holder/owner making its data available to other users. The sharing aspect deals with issues on data ownership, restrictions (or the lack of) for others to access the data, the provisions of data use and acknowledgement. The broader scientific community recognizes that the value of data is in its utilization and demonstrated impact as a result of that use. Discussions at various forums aimed at working out the principles and technicalities that would guide the broad sharing of scientific data.

In the 2010 Beijing Declaration, GEO members committed themselves to implement the GEOSS Data Sharing Principles by developing flexible policy frameworks that enable a more open data environment, which has influenced national and regional data policies including INSPIRE and Copernicus in Europe and Landsat in the United States.

The GEOSS Data Sharing Principles are as following:

- There will be full and open exchange of data, metadata and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation;
- All shared data, metadata and products will be made available with minimum time delay and at minimum cost;
- All shared data, metadata and products being free of charge or no more than cost of reproduction will be encouraged for research and education.

One of the first accomplishments of the Group on Earth Observations was the acceptance of a set of high level Data Sharing Principles as a foundation for GEOSS. Ensuring that these principles are implemented in an effective and flexible manner remains a major challenge. The 10-Year Implementation Plan says "The societal benefits of Earth observations cannot be achieved without data sharing". Further, the sharing of GEOSS data will in some cases be subject to important exceptions such as the protection of national security, privacy and confidentiality, indigenous rights, and threatened ecological and cultural resources.

The National Science Foundation (U.S.) requires a two-page data management plan for all proposals submitted. The agency stipulates that "all investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples,

physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing." This requirement started in 2011, and in just three years, the impact is quite visible. Investigators not only are giving more thoughtful consideration to satisfy this requirement, many incorporate data sharing and accessibility into their proposals as a key component for dissemination of research outcome. This will accelerate in the coming years as cyberinfrastructure for data sharing becomes more mature and widely adopted in the U.S.

### 2.1.3 Data Interconnectivity

A key challenge in disaster research, whether analysing causal relationship of various impacts or modelling to predict magnitude and scale of future disaster events, is to make use of multiple data sources, synthesize and discover the underlying relationships. At the very basic level, how does a researcher of a particular specialty find datasets that may be relevant to his/her study?

The academic community recognizes the importance of the interconnection of datasets, which often come from different scientific disciplines (e.g., hydrology, meteorology, climate, civil engineering, land use, and public health). **The term *data interconnectivity* is about connecting data from diverse sources, disciplines and data providers, and enabling common understanding and interoperability through community-endorsed standards and methods for description and access. The realization of a data interconnectivity landscape will allow scientists to easily and efficiently discover, understand, access and utilize data across disciplines in disaster research, leading to new discoveries and solutions for mitigating impact of disaster and improving readiness in communities that are vulnerable to natural hazards.**

Internationally, current efforts include those of the LODGD of CODATA, a research group dedicated to promote the linkage of disaster data, Experts in this group has formed a knowledge network and is researching the technology framework that will serve the disaster research community by connecting relevant open data from repositories at national, institutional and research group levels. Additionally, IRDR has established Disaster Loss Data (DATA) Working Group. IRDR-Data focuses its effort on issues related to the collection, storage and dissemination of disaster loss data. We envision a new kind of data infrastructure for disaster research that will connect disaster related datasets of observations, analyses, statistics, etc., from multiple scientific disciplines as well as through citizen participation.

## 2.2 The Challenges of Disaster Data

We examine the current state of open science data as related to the research of disaster risk reduction, including data from various disciplines such as social economy, earth observation, hydrology, meteorology, earthquake, geography and disaster loss statistics. While all categories of data listed here present challenges for cross-domain utilization, each also presents unique challenges due to its characteristics.

### 2.2.1 Social and Economic Data

Societal factors intervene between nature (and the natural processes)and the built environment to redistribute the risk prior to an event, and to amplify or attenuate the losses after an event (Cutter, 2010).[2] Economic development, social structure, culture, ethics, legal system are important background knowledge to understand the interaction, and to reduce vulnerability and build resilience to hazards. In assessing risks and vulnerability, the place-based(positioned) disaster loss data, as well as population, infrastructure, land use, building codes, socio-economic statistics, and disaster

[2]S. L. Cutter. Social Science Perspectives on Hazards and Vulnerability Science .T. Beer (ed.), *Geophysical Hazards*, International Year of Planet Earth, 17. DOI 10.1007/978-90-481-3236-2_2, © Springer Science+Business Media B.V. 2010

insurance data all contribute to the validity and accuracy of the assessment. In most cases, systematic information, such as the societal losses triggered by geo-hazards, is available (Cutter,2010).

The degree of open access to data varies among data owners/holders. Open access is also subject to the capability of the available data infrastructure, the legal system, as well as influence of the culture, public participation, and consideration for privacy, security or economic benefits. According to the *open data barometer: 2013 global report*[3], the availability of truly open data remains low, with less than 7% of the dataset surveyed in the Barometer published both in bulk machine-readable forms, and under open licenses. The developed countries provide more open available data than the developing countries do.

Governments collect a great deal of data, and make a portion of the data available as a public service. In some countries, datasets can be made open only when the public request them. however, such requests may not be approved by the government. In other cases, processed data, rather than raw data, is open for public access, but such data does not meet the needs of scientific research. Disaster loss data is treated as a kind of sensitive information which is a metric on a government's performance, or highly relevant to disaster compensations, and as a result, such data (is or are)only selectively made open to control public opinion. Data collected by research institute or universities under national funded programs is open to public in developed countries where information laws regulate the rights and responsibilities to make research data available. Disaster data collected by commercial entities, e.g., insurance companies, are typically not available or shared outside their own organizations as they are closely related to business strategies and practices.

Unstructured data relevant to disaster risk and loss assessment are scattered on web sites, media outlets and social media. Such data have a high degree of openness but require much more efforts to utilize in research. Understanding how such data can be connected together or with other types of data to benefit disaster research is in itself a research topic.

Under the initiative of the US government in 2011, the Open Data Coalition is taking form in the world, involving more than 40 countries and regions. "Open Data Barometer: Global Report 2013" evaluates the openness of state governments' data from such perspectives as the completion of laws on information disclosure, formulation of policy on open access of data, demand of social organizations and professionals for open data, points out the developed countries outperform developing countries in opening the data, as reflected in the fact that America and Europe are the first echelon, followed by the Asia-Pacific Region, Mid-east, Central Asia and then Africa. Across the world, the implementation of open government data policy remains at the state level, yet to reach the city level.

*Challenge: Most of the social and economic data is not open. They tend to be unstructured, and not easy to use.*

### 2.2.2 Remote Sensing Data

When natural disaster happens, various kinds of observable objects are involved including weather information such as rainfall, snowfall, wind, etc, environment elements including topography, river and so on; exposure such as infrastructure, human population, livestock, etc. Earth observation has the advantage of monitoring all these objects without unobtrusively. Based on the different electromagnetic radiation characteristics of disaster relevant objects, the different features can be detected by remote sensors and can be recorded in the form of images.

Earth observation data plays a crucial role of providing information throughout the chain of pre-warning, disaster response, post-disaster reconstruction, and disaster relief. Satellite and other remote sensing instruments may provide accurate, near real-time Earth's surface information over the world. At present, the most commonly used remote sensing data acquisition platforms are satellites

[3]Tim Davies,2013.

and aircraft. Satellites platforms including optical satellites, Radar satellites, Geosynchronous satellites, small satellite constellations, et al. Aerial remote sensing , has the advantage of mature technology, large-scale imaging, high-resolution, suitability for large scale terrain mapping and detailed investigation of the small area. however, there are limitations in flight altitude, endurance, altitude control, all-weather performance ability as well as a wide range of dynamic monitoring. Aerial and satellite remote sensing can play complementary roles in disaster monitoring.

As technologies in space exploration, photoelectric, microwave, and computing have advanced in the past decades, remote sensing technology has entered a new phase where it is possible to provide multi-resolution, multi-band, multi-polarization, multi-temporal earth observation data in time for disaster mitigation. In order to fully utilize the vast amounts of earth observation data, coherent data infrastructure for storing, transmitting and processing needs to be developed to enable more efficient distribution, computation, and fast access for disaster mitigation.

The Earth observation community recognizes the importance of data sharing. The GEOSS 10-year Implementation Plan, endorsed by all GEO Members, states: "The societal benefits of Earth observations cannot be achieved without data sharing." The Earth observation community is at the forefront of open and shared data. At present, the kilometer level low-resolution data are almost fully open and available; the 10-30 meter level resolution data are becoming more open; the 1-meter high-resolution data are available through the commercial way. Including higher resolution commercial satellite data, some major space-based observation data for disaster application can be acquired through international organizations such as Disaster Charter, et al and made available free of cost.

However, in order to retrieve timely information from various satellite data, it is in urgent need of remote sensing data with the open and standardized format such as GSFC, ISO19000 series and so on, as well as the data processing tools.

*Challenges: Higher resolution data is often necessary for disaster research and mitigation, but they are not widely available. Using such data, especially real time data, requires significant resources and expertise to process and analyze*.

### 2.2.3 Hydrological Data

Hydrological data are an important part of national basic information resources. Hydrological data obtained through long-term and continuous observations are vital to disaster prevention and reduction, as well as management of water resources. Hydro-graphic stations mainly measure water level, discharge, sediment, precipitation, evaporation and water quality. Hydrological data are the basis for evaluation of water resources, hydrological calculation and forecast, water environment evaluation and climate change. Hydrological affairs in many countries are managed in different layers from the central government to local government.

Most of the countries in the world began continuous hydrological observation only in recent years except that some started decades ago, therefore the long-term hydrological data series are very rare. Comparatively speaking, developed countries, such as the United States, Germany and Japan, have wide coverage of hydrologic station networks with high density distribution and a high degree of automation, while such network of hydrologic stations are new and sparse in African countries.

In some countries, hydrological data are widely shared. In the United States, USGS provides both real-time and historical hydrological data online and through web services (http://www.usgs.gov), including both time series stream flow and geospatial vector data. China has also built a website for sharing hydrology and water resources data (http://www.hydrodatanet.gov.cn)that various services, including online data query and download, off-line data product and yearbook printing, are provided. The sharing data include basic information

of hydrology and water resource, real-time discharge, water level and rainfall information, as well as compilation and analysis results.

Under the framework of cooperation among countries and regions, sharing and exchange of hydrological data over a basin or large area is critically important for improving the ability of countries along the shoreline to prevent and mitigate disasters, and ultimately to reduce human injuries, and loss of lives and properties. However, due to the lack of unified policy, standard and technical platform for data sharing, cross-region and cross-sector hydrological data sharing has not been truly realized at present.

Current efforts by various countries have provided examples for an international mechanism for exchange and sharing of hydrological data. More than 140 countries have river agreements among them, most of which include detailed provisions on international river monitoring, hydrological data collection and sharing. The International Commission for the Protection of the Rhine River (ICPR), established by several European countries, has technical and coordination working groups and more than 20 flood warning centres. A computer network connects these centres, hydrological stations and meteorological stations for sharing warning and other information. The 1995 cooperation agreement for Sustainable Development of Mekong River Basin requires that countries along the Mekong River regularly exchange the necessary data and information according to the Procedures for Data Exchange and Sharing among member countries.

*Challenges: Hydrology data vary significantly in formats and data representation, hence making it difficult to use across regions and scientific domains.*

### 2.2.4 Meteorological Data

Meteorological data refers to temperature, air pressure, moisture, wind, radiation, etc, that is measured in the natural environment with instruments at meteorological stations, radars, etc.. Typically, the collection of meteorological data is conducted by government agencies or departments. This practice results in a well-developed (often with sustained funding) global network for collection, dissemination, exchange, storage and sharing of meteorological data, which also explains the high degree of availability and accessibility of the meteorological data in many countries.

A sound coordination mechanism underlies the open access of meteorological data around the world. The sustained efforts have promoted the open access of meteorological information across the world.

The World Meteorological Organization (WMO) promotes cooperation in the establishment of networks for making meteorological, climatological, hydrological and geophysical observations, as well as the exchange, processing and standardization of related data, and assists technology transfer, training and research. WMO facilitates the free and unrestricted exchange of data and information, products and services in real- or near-real time on matters relating to safety and security of society, economic welfare and the protection of the environment. WMO RESOLUTION 40 (Cg-XII) adopts the following policy on the international exchange of meteorological and related data and products and WMO commits itself to broadening and enhancing the free and unrestricted1 exchange of meteorological and related data and products. Global Climate Observing System(GCOS) is a joint undertaking of the World Meteorological Organization (WMO), the Intergovernmental Oceanographic Commission (IOC) of the United Nations Educational Scientific and Cultural Organization (UNESCO), the United Nations Environment Programme (UNEP) and the International Council for Science (ICSU). GCOS is to provide comprehensive information on the total climate system, involving a multidisciplinary range of physical, chemical and biological properties, and atmospheric, oceanic, hydrological, cryospheric and terrestrial processes. It includes both in situ and remote sensing components, with its space based components coordinated by the Committee on Earth Observation Satellites (CEOS) and the Coordination Group for Meteorological Satellites (CGMS). GCOS is intended to meet the full range of national and international requirements for climate and

climate-related observations. As a system of climate-relevant observing systems, it constitutes, in aggregate, the climate observing component of the Global Earth Observation System of Systems (GEOSS).

At the national level, developed countries and major developing countries have established open data infrastructure with clear policies on data sharing, making available meteorological data ranging from ground-based observation statistics to satellite data.

As for disaster research, it needs to meet the demand for disaster prevention and reduction by further opening meteorological data. The layout and distribution of ground-based meteorological stations remain to be critical issues in meeting the high demand for disaster prevention and reduction. For instance, higher density of common meteorological stations and faster observation frequency will improve data collection, which in turn will help regions, especially in the vast number of developing countries, to deal with emergency disasters and unexpected local disasters. Furthermore, although globally shared meteorological data is within convenient reach thanks to efforts by world meteorological organizations, meteorological information systems in different countries are not yet connected. Significant gaps exist in terms of integrated use and coordination of data from multiple sources.

*Challenges: There is inadequate data collection in developing countries. The data tends to be very large and have diverse formats, presenting barriers for researchers in other domains to use in their research.*

### 2.2.5 Seismic Data

Seismic data comes from two areas: seismology and engineering seismology. The former is a comprehensive discipline about the occurrence patterns of earthquakes through the solid medium of earth, the transmission patterns of seismological waves and the macroscopic consequences of earthquakes. Engineering seismology studies the consequences and impact of earthquakes on urban areas and civil infrastructures using seismological theories and methods with the goal of improving the design and construction of structures to be earthquake resilient.

Opening of seismic data within the country boarder had been adopted around the world, helping nations to build their national earthquake early warning and reduction system. Taking China's National Seismic Science Data Sharing Centre as an example, it is the main agency undertaking the seismic science data sharing around the country. Provincial-level data sources feed data into the national data sharing centre. The dissemination and sharing of seismic science data are classified into four tiers: the first-tier data is open to the public, the second one is open to domestic and international users, the third is for domestic users, and the fourth is for exclusive use by authorized users.

Sharing of seismic data is often a sensitive issue. There have been efforts to exchange seismic data internationally for decades. Although an international site-record sharing and exchanging mechanism is in operation, the number of sites and variables of exchanged data are very limited compared with nation level sharing. The primary reason is because of potential use of seismic data could be used to for sensitive purposes, such as monitoring the occurrence of nuclear tests.

*Challenges: Countries and regions restrict the dissemination of seismic data because of sensitive nature of their potential usage.*

### 2.2.6 Geographic Data

Geographic data is information related to location and space on land surface. A comprehensive collective of geography information, referred to as the "fundamental geography data", describes the earth's measuring points of control, drainage system, residents and facilities, transportation, pipeline system, the boundary and the administrative region, geomorphology, vegetation and soil, cadastral, toponym, etc., which is related to natural and social elements of the location, shape and attribute

information. This data is not only critical to research in geo-sciences and other scientific domains, but also, perhaps more importantly, critical to the design, management and risk mitigation of urban infrastructure such as municipal facilities, transportation, etc. (Although most of these data sets are processed data using various technologies mentioned above, they represent a category of information that can stand on its own in this study.)

In most countries, this type of data is collected and managed by government agencies. Accessing to such datasets are restricted with different authorization mechanism in most countries. Certain low sensitivity data are accessible by the public. The definition of free accessible dataset varies significantly among countries.

The past decade has seen progress to open more fundamental geography datasets. The United States Geological Survey (USGS) provides free access to datasets of digital elevation model (DEM), digital orthophoto quadrangle (DOQ), Digital Line Graph (DLG), Digital Raster Graphic (DRG) data ranging from 1:24,000 to 1:250,000, as well as the LANDSAT 7 satellite remote sensing data, land use data, population densities data, soil surveys data and so on. Land Information New Zealand (LINZ) established a topographic databases (NZTopo) with scales ranging from 1: 50,000 to 1: 4,000,000. More than 40 datasets have been released on the web (LINZ Data Service), including LINZ topographic maps, New Zealand offshore islands, Pacific region topographic maps, and so on. Geospatial Information Authority of Japan (GSI) has completed the national 1:25,000 topographic maps, and provides various maps, including the topographic maps, land use maps, vegetation and biological maps, and the volcano correlation maps, at different scales.As developing countries, China has completed its national foundation geographic information system ranging from 1: 1,000,000, 1: 250,000, 1:50,000, and 12.5m grid interval digital elevation models of the key guard areas of the 7 main rivers, and 1m resolution numeral orthogonal projection likely database, which covers the national 9,600,000 square kilometre national territory area. As of October 2014, China's science and technology resource sharing network (escience.gov.cn), has shared 1789 datasets, ranging from geosciences, health, agriculture, forestry, weather, hydrology, materials science, etc.

*Challenges: Restricted access is often placed on these datasets. The diversity of data formats and expertise in geospatial analysis and processing are also barriers to researchers.*

### 2.2.7 Disaster Loss Data

Data on disaster losses refer to statistics of various costs related to disasters such as casualties, damaged buildings, GDP and other economic losses. Compiling such socio-economic data typically needs to consult and use a wide range of scientific data before, during and after disasters occur.

Access to disaster loss data is always viewed a sensitive issue and subject to governmental policies in all countries. Many governments use a top-down approach, collecting comprehensive data and generating statistical information within government operated organizations. Such data typically contain highly sensitive information on deaths, damages and losses; they are used in government-sanctioned domains, and generally not accessible with protected from public.

However, the trend around the world is toward open access and sharing of disaster loss data. The Working Group on Disaster Data (WGDD), which consists of Asian Disaster Reduction Centre (ADPC), Centre for Research on the Epidemiology of Disasters (CRED), Global Risk Identification Programme (GRIP), Red de Estudios Sociales en Prevención de Desastres en América Latina (Network of Social Studies in the prevention of Disasters in Latin America - LA RED), Munich Re, and United Nations Development Programme (UNDP), proposed the Disaster Loss Data Standards, to improve the sharing and interoperation of disaster loss data via networks. An example in this regard is the US disaster loss databank established by Hazards & Vulnerability Research Institute of University of South Carolina. The data sharing platform of SHELDUS™ provides loss data for 18 different natural hazard events types to users in the U.S. Section 6.1 of this paper describes this platform and its applications in more detail.

Under the present management mechanism for disaster loss data, an urgent priority is constructing a unified metadata repository and realizing open access to the metadata information. This step will significantly help improve disaster emergency management, response, and disaster scientific research on long term issues.

*Challenges: This type of data almost always has sensitive information about the region and country, hence, most of this data is not openly accessible. There is also lack of data standard for this type of data, making it difficult to use.*

## 2.3 Status of Open Data in Developing Countries

While almost all countries around the world suffer from natural disasters (earthquakes, volcanoes, floods, droughts, storm surges, etc), the impact of such disasters on the developed and developing countries are quite different: heavier human losses in the developing countries while greater economic losses in the developed countries. The low quality infrastructure, lack of effective mechanisms and polices to manage disasters and a lower degree of social mobilization are often attributed to the greater loss of human lives and properties in the developing countries. International communities on major natural disaster reduction research and practice recognize that the scientific use of disaster data can improve the timeliness, accuracy and effectiveness of disaster mitigation decision-making, and thus greatly reduce disaster losses.

But for developing countries, usually, the effect of technology achievements and joint international action cannot be fully reflected. Developing countries should draw on the technological progress of mankind to effectively construct their own disaster reduction capacities. Speaking from the international community, disaster data sharing capacity bears on the corresponding countries' economic development and scientific strength. As developing countries have no independent remote sensing satellites and can't timely and effectively respond to natural disasters, equal access to data is not guaranteed. UN ECA holds that timely acquisition of remote sensing data is a powerful instrument to promote regional sustainable development. And theoretically, remote sensing technology provides both developed and developing countries with data of the same quality at the same frequency. But the high cost remains as a barrier for developing countries. The international society may immediately start to facilitate transparent share and transfer of technical achievements from developed countries to developing countries through "data democracy".

Some developing countries are making progress toward this goal. As an example, China, one of the largest developing countries, has employed the following three mechanisms to improve the capability of opening and sharing disaster data with the international community.

### 2.3.1 National coordination mechanism to fully use the internal data resources

After Wenchuan Earthquake, the Chinese government launched a project of Disaster Emergency Data Reservoir (DEDR). DEDR targets studies of emergency collaborative planning and scheduling technology, emergency data sharing technology and emergency cooperation mechanisms. In 2011, it was put into operation with standardized and stable connection with Chinese Metrology, Ocean and Land satellite data centres, as well as aerial remote sensing data centres and UAV operation companies. As part of a national mitigation data support system, DEDR receives international support through cooperation with the International Disaster Charter, UNSPIDER as well as IRDR. When the huge earthquake shocked Sichuan Province in Lushan County (latitude 30.3, longitude 103.0), Ya'an City at 8:02am on April 20 of 2013, DEDR started the emergency services at 11:50 to coordinate the sharing of data from domestic satellites and aerial remote sensing sources as well as international satellites. Within 12 hours after the earthquake, it harvested and published the pre-event dataset of the disaster-hit area through data centres of CRESDA, CMA, CEODE, BJ-1, as well as SPOT、LANDSAT、SJ-9A satellites. Subsequently, DEDR received disaster observation data from every related agency. Up to 10:00 a.m. on April 26, 2013, the Reservoir had collected 112 GB satellite and

aerial imaging disaster data, of which approximately 61 GB was about pre-disaster and 51 GB post-disaster. These data were immediately used by 19 ministries and local governments. The total download reached over 2TB about this event via the DEDR platform.

### 2.3.2 Regional cooperation on disaster data sharing

In cooperation with China and India, the United Nations Economic and Social Commission for Asia and the Pacific (UN-ESCAP) is promoting a Regional Cooperative Mechanism for Drought Monitoring and Early Warning in Asia and the Pacific (The Drought Mechanism), for development and operational provision of Earth Observation-based products and services to relevant countries in the region. Under the Mechanism, a pilot project for Mongolia and Pakistan is underway, and a project involving other countries is under consideration. It can help the less-capable countries to share the benefit of regional progress.

### 2.3.3 International assistance mechanism data for disaster

On May 12, 2008, a Richter 8.0 Earthquake struck Wenchuan and its surrounding areas in Sichuan, China. The earthquake and subsequent disasters, like geological disasters, caused heavy casualties and huge property loss. China used 15 satellites for emergency observation, including Fengyun meteorological satellite, CBERS, Beijing-1 small satellite, and airborne instruments. Even so, they could not meet the extraordinary demand of coverage and timeliness required by disaster analysis, and data processing, and urgently needed assistance from the international community. Through CEOS, International Disaster Character, UNSPIDER and other coordination mechanisms, China established point-to-point connection with most of the world's leading earth observation organizations. International assistance remote sensing data continued to be provided to the Chinese side from NASA, USGS, JAXA, ESA and etc. Other international agencies also provided data assistance through other channels indirectly.

# 3 Gaps and Challenges in Linking Open Disaster Data

To enable broader use of open science data in the research of natural hazards and disaster prevention and reduction, it is critical to expand open access and data sharing from current practices and provide better linkage of data, especially datasets from different organizations and disciplines. Many barriers still exist today for majority of researchers, especially those in the developing countries, to find relevant datasets and access them easily. In this section we discuss the major gaps and challenges related to these barriers.

## 3.1 Technology Gaps and Challenges

Information technology is the backbone of the existing and certainly future generation of disaster data infrastructure. The technical elements comprising the data infrastructure include *data management, data discovery, data interoperability* and *data services*. These must work seamlessly together to manage and serve the diverse, distributed, multi-disciplinary disaster related data and information to various stakeholders (researchers, decision makers, and the public).

Today's data infrastructure that host and serve data to disaster studies have been developed by various groups, projects, institutions and agencies. Efforts have been directed at establishing coordination among different data repositories within a country and across country boundaries to share data. However, significant gaps remain, preventing the effective use and broader applications of these important data resources. As data acquisition techniques continue to improve, newer, higher resolution and more varieties of data will be generated, pushing the limit of today's data infrastructure.

The first major challenge is the management of the large amounts of diverse, heterogeneous datasets that are needed by the disaster research community and the public. This requires both hardware and software, such as storage systems (including storage media ,interconnect fabrics and data structuring) and data **management** software, as well as people expertise in running the facilities. In addition, analysis, processing and visualization capabilities (e.g., high-performance computing, big-data computing) are necessary to support of the use of the data. The challenge for various organizations to operate a large data facility lies not only in the initial investment and setup, but also the on-going maintenance and technology update necessary to support a reliable and on-demand data service. Although not unique to the disaster research community, this gap is exacerbated by the scale, time-criticalness and the broad involvement of multiple disciplines and sectors of the society of the domain. As more and more unstructured data, such as social media communications, are being considered and used in disaster studies, data management systems need to adapt to support such data and provide corresponding data services on the unstructured data.

The second major challenge is that disaster related data is **difficult to discover and access across multiple repositories and disciplines**. As pointed out in earlier sections, disaster related information resides in various geographically and organizationally distributed repositories, using different data formats, access protocols and polices, managed by institutions with very different governance models. The lack of interoperability standards makes it hard for individual researchers to access existing data resources. Further compounding the issue is the multidisciplinary nature of disaster research. For example, when studying an earthquake, a researcher may also need to find data related to landslide, heavy rain, flood, or volcano eruption. It would be very difficult, if not impossible, to find all or most of the datasets relevant to a study topic – finding one dataset does not lead to other information that are directly and indirectly related (e.g., geographically and/or temporally). Even when one identifies a data repository that likely contains the datasets needed, the user typically has to spend significant amount of time to extract the right dataset (e.g., entering search criteria acceptable to the interface, which is non-trivial to researchers not familiar with that particular data resource and interface), to be able to obtain the data (e.g., having the network bandwidth and local storage to download the data. This segues naturally into the next major challenge related to the utilization of data.

Disaster data is often **difficult to use across applications and disciplines.** When researchers are able to access and obtain relevant data, they may not have adequate description of the data (lack of standards on data formats), and may not understand the data representation due to disciplinary background. The variety of data formats presents a significant challenge. As described in Section 2, disaster research demands many different types of data, such as data from remote sensing observation, meteorology, hydrology, geography, socio-economic statistics, and social media information. Some fields do not have established standard formats, making such data more difficult to discover and use because of the lack of broadly recognized vocabulary. Progress in the domains of taxonomy, ontology, data structuring and mapping between data structures, etc., is expected to help address some of the challenges.

The fourth major challenge is the poor interoperability of **'cross-disciplinary'** datasets for disaster research. This issue stems largely from the multidisciplinary nature of disaster data. Data from both natural sciences and social sciences are needed for disaster research. For instance, in forecasting hazardous events due to weather, observational data such as satellite data, meteorological and hydrologic data will be considered; in assessing vulnerability, risks, and estimating potential damages (to aid local government in decision-making), various statistics such as urban infrastructure, population, and other socio-economic data will need to be considered together with hazard related data. These two types of data are often collected using very different approaches, governed by different policies, and managed by different types of organizations (research/technical groups vs. government/bureaucratic institutions). An even bigger issue is related to different spatial scales of the

datasets – the socio-economic data are often at country or province level while observational data can be as high resolution as a meter. Downscaling of the coarse-scale data often uses unrealistic assumptions (e.g., distributions of infrastructures, economic activities and population at various times of the day), leading to inaccurate assessment and estimate.

In addition, increasingly large amounts of unstructured data, such as news reports, social media chatters, pictures, that certainly contains relevant information and may also contribute to vulnerability and potential loss analyses. Social science researchers are also using "storylines" to describe disaster events based on their observation and analysis "on the ground" where these events occur. Other researchers conduct quantitative analysis and collect data through interviews and questionnaires. The quality of such data may be debated as the method of collection and interpretation appears to be more subjective than objective, compared with observational data from physical sciences. The challenges of effectively utilizing unstructured data remain research topics today, including methods to collect, index, manage the data, and make them discoverable, accessible and usable by the broad research community.

Finally, **data services** will be a critical part of disaster research. The word "service" has at least several connotations such as reliability, persistence, and stable interfaces. Disaster data service providers will need to address these aspects in addition to storing, managing and supporting access to the data itself. Disaster studies are highly data-driven and data-dependent, and the user community require multiple interfaces for reaching at the data, e.g., database query, web service, and graphical user interface. Federation of data service providers is a major challenge due to lack of standards. Other areas of considerations include data validation, citation, and community feedback.

## 3.2 Policy and legal Gaps

The Sendai Framework for Disaster Risk Reduction 2015-2030, adopted at the Third UN World Conference in Sendai, Japan in March 2015 as the result of stakeholder consultations through the United Nations Office for Disaster Risk Reduction (UNISDR), establishes a global policy agenda (United Nations Office for Disaster Risk Reduction, 2015). It aims to achieve a substantial reduction of disaster risk and losses in human lives, livelihoods and health, and in the economic, physical, social, cultural and environmental assets of persons, businesses, communities and countries over the next 15 years. Compared with the Hyogo Framework for Action (HFA) 2005-2015: Building the Resilience of Nations and Communities to Disasters, the Sendai Framework puts a strong emphasis on the application of natural and social sciences driven by data.

One of the priorities for action in the Sendai Framework is understanding disaster risk. To achieve this goal, it is vital to promote the collection, analysis, management and use of relevant data and practical information, and ensure broad dissemination of such data to meet the needs of different categories of users. At the same time, it is also crucial to promote real time access to reliable data, make use of space and in situ information, including geographic information systems (GIS), and use information and communications technology innovations to enhance measurement tools and the collection, analysis and dissemination of data.

While data and information sharing is understood to be critical to scientific progress, academic scientists do not necessarily share data or information with their colleagues (Thursby, Thursby, Haeussler, Jiang, 2009). Often, incentives exist to encourage scientists not to disclose research. Their reluctance to share data and materials is increasingly considered as a major problem (Cohen and Walsh, 2008). Scientists' willingness to share data is also highly specific to the context in which they conduct research (Haeussler, Jiang, Thursby, and Thursby 2014). Therefore, it is important to understand the benefits and costs of sharing data to scientists and consider carefully how policy can influence decisions on sharing and providing access to data.

Policies on open access to scientific data would take the form of mandatory rules, infrastructure, or incentives (OECD, 2015b). Mandatory rules could be implemented through requirements in

research grant agreements or national strategies or institutional policy frameworks. Infrastructure, which could take the form of soft or hard one, would include initiatives undertaken to develop an open science culture. Since the Organisation for Economic Co-operation and Development (OECD) established principles and guidelines on access to public research data in 2007, the member countries have made efforts to adapt legal frameworks and implement policy initiatives to encourage greater openness in science (OECD, 2015a). Development of the skills necessary for researchers to share and reuse the research data produced by others and data management guidelines for universities and public research institutes are also important infrastructure. Incentives would be provided by financial support to cover the cost of releasing data sets, proper acknowledgment of the efforts of researchers for data citations, and career advancement mechanisms based on metrics taking into account data-sharing efforts.

Compared with mandatory rules or infrastructure, however, incentives mechanisms for researchers involved in open data activities have not been widely introduced. Evaluations of universities and researchers are still mostly based on research bibliometric indicators, with little value attributed to the sharing of pre-publication inputs and post-publication outcomes including data. And data cleaning and curation through the development of metadata, which requires a substantial amount of resources, is not acknowledged adequately in evaluation mechanisms or grant allocation procedures. This issue could be addressed to a certain extent by extending citation mechanisms to data sets. To scientists it is crucial that their work is recognized in their communities, and hence reward to them would be scientific acclaim or a prize, such as the Nobel Prize. Increasingly scientists could consider reward in the form of financial return, such as income from commercial application of the scientist solution. Hence a significant challenge in policy making is to incorporate these incentives to scientists into practical measures and instruments for encouraging sharing and providing access to data.

Tackling grand challenges such as disaster risk reduction requires close cooperation and collaboration on access to and sharing of data beyond sectoral or geographical boundaries. On the other hand, when research activities are conducted in partnership with the industrial sector, commercial interests would not be ignored, and consequently the mode of sharing research results could be different from the case in which only public actors are involved. And certain classes of data, such as medical records and national security, need to be treated with great care in their access and sharing, due to potential concern about sensitivity or confidentiality.

Open data access needs protection and guarantee from laws and policies, including information security and privacy. The absence of strong Right to Information Laws may prevent citizens from using open data to hold government accountable, and weak or absent Data Protection Laws may undermine citizen confidence in open government data initiatives (Open Data Barometer,2013). In the developed countries, laws of data access and data protection have been well established (e.g., in the U.S., the Freedom of Information Act provides that "that any person has a right, enforceable in court, to obtain access to federal agency records, except to the extent that such records (or portions of them) are protected from public disclosure by one of nine exemptions or by one of three special law enforcement record exclusions"), while they are lacking in most developing countries today. The lack of protection for public's right to information impedes work in the area of disaster research, for example, as it is difficult for non-governmental organizations, research institutions, and the public to obtain certain high resolution and important disaster information. In many countries, the increase in the demand for data sharing exceeds the pace at which the laws governing data access and sharing are established. Citizen's right to information (similar to United State's Freedom of Information Act) will ensure open access to government data for effective transparency and accountability (Open Data Barometer,2013). Policies and technical implementations are lacking in addressing the dynamic nature of data access. For instance, data that is normally restricted will need to be made available openly to aid in emergency situations.

While we advocate broader sharing of multidisciplinary data to support disaster research and practices, we also recognize that new issues may arise as a result. For example, a dataset that was acquired originally for a specific purpose or by one type of application could be used in a different domain and disseminated to a completely different audience. Examples of such data may include medical records and statistics, and social media interactions. Data may be used in applications beyond the original intended use, thus expanding the scope of use. Laws and regulations to protect intellectual property and scope of use are needed to ensure appropriate use of data.

### 3.3 Governance and Cultural Gaps

Data may be considered as property of a particular organization in countries where management of data is distributed among different government departments, universities and research institutes without an authoritative coordination department（Fan Xiu'er, 2006）. A similar situation exists for disaster loss data. Disaster loss data at the provincial level are available in the statistical yearbooks, while the high-resolution data is not fully open. When high-resolution data at other departments is needed, administrative coordination is necessary in order to access the data. This process is time consuming and labor intensive. Many developing countries lack a basic foundational framework for managing governmental digital data, and need to build government data collection and management capacity (Open Data Barometer, 2013).There is no systematic open-access inventory or accounting for individual nations or aggregated to the global scale of hazard events and losses by location or by hazard agent (Cutter, 2010).

Cultural and societal ethical standards also have a significant influence on the degree of data sharing and openness. Countries such as those in Europe have a high degree of open dissemination of governmental data as it is considered an integral part of civil rights, and having great benefit to governance, democracy and social development (Liu Kejing, 2007). In some culture, governments are reluctant to publish negative information, such as losses caused by disasters, for fear of the perception of incompetency on the government's part.

The level of awareness on citizen's right to information varies among countries of different cultural backgrounds. In some parts of the world, people recognize that governmental data, especially those related to disasters, should be openly accessible by anyone and they demand data transparency from their government. As a result, they often do have access to such data. In other countries, citizens are less active in public participation of disaster management. The lower levels of awareness, transparency, and public participation are barriers to disaster data openness.

## 4 Scientific Issues behind Data Interconnectivity

There is a wide range of active research topics related to the use of scientific data. In this section, several describes the scientific issues related to open disaster data: autonomy of disaster data resources, data classification based on dependency relationship of disaster event-supporting data, and the conflict between specialists and the masses toward open access of disaster data.

### 4.1 Data Dependency

Hazards of different types lead to disasters, such as flood, droughts, heavy rain, snowstorms, earthquakes, typhoons, landslides, wildfires and infestations, etc., causing great loss of life, damage and hardship. Research on each type of disasters involves (1) pre-warning, prevention and simulation prior to the disaster event; (2) the cause, process, and occurrence mechanism while the disaster is unfolding; (3) emergency responses, assessment of loss, post-disaster recovery and reconstruction. Research on long term strategies of vulnerability assessment and mitigation, as well as communication with the public, aims at making the society more resilient to natural hazards and reducing or eliminating loss of life and property. Disaster research is multi-disciplinary by nature,

conducted through a complex process involving multiple domains, systems and data sources. One of the most notable characteristics of this research domain is its high degree of dependency on data. There is a tendency of comprehensive usage of data from multiple domains. Multiple natural processes and hazards may contribute to the same disaster event, e.g., landslides and flooding that occur at the time of an earthquake, and it would not be possible to study such phenomena without access to relevant datasets.

We looked at a number of SCI academic papers about earthquake and flood disaster events and examined the data used by researchers in studying different types of disasters. We found that the researchers of earthquake disasters relied on primarily eight types of data, namely, geological data, geophysical data, ground observation data, basic geographical data, earth observation data, space physics data, clinical medicine and socio-economic data. The researchers of flood studies relied on eight types of data including earth observation data, hydrology data, meteorology data, basic geographical data, ground observation data, biological genetics data, geophysical data and socio-economic data. Collection of disaster relevant data often depends on the condition of the natural environment, available technology, and economic strength of the affected area. Hence, data dependency may differ across regions, even for the same disaster event, due to temporal and regional characteristics. To consider the comprehensive impact from multiple systems, researchers rely on multi-source data. For example, Rutger Dankers utilized data on climate, geography, plantation and land cover in his simulation of flood disasters; Uwe Ulbrich utilized the socio-economic data, hydrologic data and observation station data in the study on the formation of precipitation and flood.

There is a significant technical challenge in managing the wide variety of disaster related data and making it possible for researchers to find related datasets.

The first challenge is the lack of consistent data representation. Disaster data comes from a wide range of disciplines. The methods of data collection vary significantly, from field monitoring, remote sensing observation, macro statistics, on-site investigation, to models and simulation. Data heterogeneity not only lies in the differences in measurement unit, recording method, description of data or metadata, and format, but also in the diversity of systems of storage, discovery and access.

Secondly, different disciplines classify the data in different ways. The current classification system of disaster data comprises the four-category, five-category and seven-category classifications . The field needs to have a classification standard that is accepted by the disaster research community and implemented in data management systems. Such a standard not only considers scientific classification of disaster data, but also take into account the overlap among data from different disciplines, and more importantly, ensures effective classification of data in practical use to promote exploration for more data. Developing an appropriate classification system holds the key to providing the optimal data plan to enable broad use of open disaster data and accelerate research in dealing with natural disasters.

## 4.2 Specialists and the Public toward Disaster Data

Disaster events have far-reaching impact, often affecting lives of tens of thousands, even millions, of people who are victims at the forefront of natural disasters. As they bear the impact and participate in mitigation of disasters, the public are both data collectors and consumers. As disaster data plays a very important role in all the phases of disaster reduction, one of the goals of sharing of disaster data is to obtain data from the public and also serve the data to the public people.

Before the era of industrialization, disaster reduction is a social management conduct with a high degree of social mobilization. Due to the lack of necessary technical means, all members of the society needed to be mobilized and various techniques and skills were employed to observe as well as forecast disasters and conduct relief efforts. Industrialization and, most recently, development of computation and information technology, disaster forecast and research to understand the various hazards have become more and more specialized and professional, leading to rather restricted access

to disaster data. Data sharing has been interpreted narrowly as sharing among specific communities specializing in the relevant techniques and usage.

Advancements in information technology, especially the widespread use of the Internet, have made it easier to put data and information in the hands of the masses, creating new opportunities for both research and commercial interests. The open data movement in disaster data management is driven both by the need for increased transparency from the public and by the need for public's participation to improve disaster management.

One of the most recent examples is related to air pollution in Chinese cities. Heavy smog caused by air pollution across China in recent years has been shown to have negative impact on people's health and daily lives. Many people check air pollution information several times a day – a use case where hundreds of thousands of users accessing real-time air quality data via the Internet as evidence of a strong need to support public data consumption.

Another example is the industry of tornado forecast. In countries and regions prone to tornados, the public constantly check tornado related information, which propels further opening of data and improvement in access.

Meanwhile, the public are becoming disaster information collectors as well as consumers. The participants in disaster data related activities now extend from a group of scientists to the public, which poses a big challenge in terms of both technology and policy. In a project sponsored by the Colorado-based satellite company Digital Globe, Tomnod utilized crowd sourcing to identify objects and places in satellite images. After MH370 went missing, Tomnod swiftly pooled global remote sensing images and distributed them online so that anyone could examine and mark all possible information related to MH370. Prior to that, more than 10,000 people from various walks of life had participated in data processing for the website.

Making data openly accessible by the masses brings major challenges and incurs additional costs to address these challenges. While the data specialists have the tools and knowledge in dealing with complex and heterogeneous data, the data infrastructure to support public participation of data collection and data consumption remains a challenge. When the public participates in data collection and data consumption, policy issues related to privacy protection and proper use of data rise to be a top priority. We also recognize that disaster data may reveal different issues to different stakeholder groups. For example, data producers and users may have different considerations due to their different interests. Disaster data comes in all forms, shapes and places. The geographic distribution of the data resources may cause problems related to timeliness and accuracy. The lack of standards for integration of static and dynamic data and the co-existence of massive amounts of heterogeneous data present significant challenges in reconciling and utilizing the data. The full extent of issues, both scientific and technological, involving the public as a key component in disaster data collection and usage should be studied.

## 4.3 Autonomy of Disaster Data Resources

Today, disaster data resources typically reside within the confines of agencies or institutions, often with technology implementations and polices specific to the hosting institutions. In the long term, it is necessary to grant autonomy to data resources. Autonomy of data recognizes the independence and transparency of data and promotes self-governance of data management and services. Progress in this direction will depend on technology advancement and, more importantly, policy shifts toward open access.

From the perspective of technology, the needs from specific applications and government mandate often drive the initial development of a data infrastructure. The resulting infrastructure is often constructed to meet the immediate demands by sacrificing portability and reusability of data resources. Compounding this deficiency is the policy requirement, e.g., restriction on the data fields to

be exposed to open access, that results in systems that are not broadly usable. This creates significant barriers to the utilization of the data resources, including:

- non-standard representation of data and meta data
- poor interoperability of data resources in different systems and applications
- inconsistent access methods, difficult for broad dissemination of data
- high cost of system development and customization to serve application-specific needs.

While we recognize the benefits of application-driven technology development, the multidisciplinary nature of disaster research can serve as an effective driver to increase usability and interoperability of data. By realizing autonomy of disaster data resources, the focus on data resource will now be shifted to establishing service-oriented data facilities to help the user community effectively use, reuse and share data resources. The neutrality of data service facilities will directly impact the open access and linking of data resources. Unlike the traditional data holdings, autonomous data facilities will be built with open access, interoperability, and policy management in the design at the beginning.

The autonomy of disaster data poses two main challenges: *technical* and *policy-wise*. First, there needs to be an appropriate framework to define policies related to access, rights, credit, etc. Such policies will be integrated in the data infrastructure for presentation, enforcement and update. Second, the data resources must follow standards in order to support heterogeneous systems, including standards for metadata, access, sharing, etc. This construction of data standards need to bebased on existing data standards and embraced by the community. To access data via the heterogeneous data sharing system can selectively get the meta-data expression and description corresponding to application needs, and thus meet the specific demand of the application system.

Management of disaster data in autonomy is the basis for long-term sustainable opening and sharing of data to the largest extent, thus a deeper understanding of autonomy will promote reforms of culture, policy, interest relations and technological conditions related to disaster data.

# 5 Cyberinfrastructure for disaster data interconnectivity

Technological barriers have been a major gap in our ability to link scientific data from multiple domains and types of instruments effectively for disaster researchers, as described in Chapter 3. However, the extraordinary innovations in information technology and their rapid adoption by researchers and public alike have reached the point where many of these barriers can now be addressed by a new and enhanced data infrastructure capable of supporting data producers and consumers of diverse sources and types of scientific data. This section we discuss a number of key enabling technologies that could contribute to such a cyberinfrastructure that will help realize our vision of an advanced data infrastructure of interconnected cross-domain disaster data for research and knowledge dissemination.

Although a new term a decade ago, *cyberinfrastructure* (*e-Science* in some countries) is widely used to refer to a collective of interoperable information systems, data and software that is fundamental to scientific discovery and collaboration. Analogy to the traditional physical infrastructure of roads, bridges, power grids, and telephone systems, cyberinfrastructure encompasses a set of complementary and interconnected areas including computing systems, data repositories and other information resources, networks, digitally enabled instruments and sensors, connected through interoperable software and tools, and services (Atkins et al ,2003). Applications of specific communities can be built on top of and utilize resources in a cyberinfrastructure. The architecture of a data infrastructure would typically consist of various layers from networking, computing, digital data, to interoperable services. In this section we highlight several key areas of cyberinfrastructure relevant to an interconnected disaster data infrastructure.

## 5.1 Networking and Data Movement

While the desire to allow scientific data to be freely accessible has been around for a long time, the digital revolution in the past two decades that has put network access to almost any content into the hands of millions of people promises to make the vision of open access to science data a reality. Network connectivity has become ubiquitous thanks to the increased connectivity among institutions, academic campuses, government agencies, and even homes, and to the proliferation of personal computation and communication devices in recent years. In the context of a data-centric cyberinfrastructure, the networking layer provides the foundation to data accessibility.

Computer networks are the most mature in terms of information standardization. They use well-defined formats, or protocols, for exchanging messages, although the underlying implementation may vary from device to device. A Wide-Area Network (WAN) spans regions, countries and even the world, connecting localized networks (e.g., LANs, MANs) and transmitting data over long distances and between different local networks.

Any initiative of open data and data interconnectivity relies on well-connected network and adequate data transfer rate in order to support movement of data from data providers to data users and sharing of data among sites and users. As data size increases rapidly, data transfer rate and methods of transfer have important implication to success in linking and sharing of data. For example, it takes one day to transfer 1 terabytes of data over a 100 Mbps (megabits per second) network, in comparison to approximately 20 minutes over a 10 Gbps network (assuming a disk fast enough to store the incoming data).

Data transfer refers to the transmission of data from one physical location to another. File sharing is a primary example of transferring large amounts of data across the Internet. While emailing files as attachments or putting them up on web sites for download remain to be popular ways of sharing data, most datasets needed in disaster research are too big and complex to be shared effectively by these means. More recent technology and services, e.g., cloud storage such as Dropbox, Google Drive, and Amazon S3 to name a few, have rapidly changed the way people share personal files, such as photos, videos, and other files, across computers and mobile devices. These services allow a user to put files in a remote storage server, and still have control over who can create, access and delete the files. Amazon S3 supports up to 5 terabytes in size. The cost and transmission rate remain in researchers' ability to share files of much larger sizes, from hundreds of gigabytes to terabytes. High-performance parallel file transfer methods (such as graft using parallel TCP streams and multi-node transfers to achieve high throughput) have been used by academics for many years. Services created based on these methods, such as Globus Connect (Foster et al,2011), iDrop Desktop (as part of iRODS (Moore et al, 2010)), Phedex used by the LHC Experiment (Egeland et al,2010) to name a few, are being used by research institutions and individuals for big data transfers across the Internet. There is a well-established body of work that can be leveraged in building the next generation of disaster data infrastructure where supporting big data from multiple domains is one of the key requirements.

## 5.2 Advanced Computing

The need for computing power is everywhere, especially in the area of disaster research, from modelling and simulation, to data processing and visualization. The power of a single computer has steadily increased while the cost of the hardware has come down, making computing more affordable than ever. Today, the processing power in an iPhone, for example, rivals that of a supercomputer 30 years ago. A single multi-core server can meet the needs of simulation and data analysis for many researchers. However, significant computational power is critical to meet the needs of simulation and modelling at the regional and global scales, using high and super-high resolution data, and dealing with large amounts of data of different formats and sources.

There are several forms of computation: serial, parallel, and distributed computing, all of which are applicable for various problems in the disaster research domain.

*Serial computing* refers to software written as a serial stream of instructions executed on a CPU (central processing unit) on a single computer. The instructions are executed one by one in sequence. Software is easier to write, but for large problems, e.g., dealing with many points on a grid or a long simulated time period, the program may need to run a long time.

*Parallel computing* paradigm breaks up a problem into many smaller tasks and execute them concurrently on multiple CPUs. These CPUs may be on a single computer, as in multi-processor, multi-core machines, or on multiple computers. While executing the subtasks simultaneously may speed up the calculations, the amount of time needed for the subtasks to communicate with each other may affect the speed-up of the total processing time. Applications whose subtasks need to communicate with each other frequently are referred to as being *tightly coupled*. Modern supercomputers, such as Tianhe-2 (China), Titan (USA), and K-computer (Japan)[4], with high-speed (often custom designed) interconnects between the computing nodes are designed to support the tightly coupled parallel applications. Applications whose subtasks rarely or never have to communicate are called *loosely coupled*, or more often, *embarrassingly parallel* applications. These applications can run efficiently on systems connected with slower and less expensive interconnects (such as gigabit Ethernet). They lend themselves well on distributed computing systems (aka high-throughput, grid computing) where the computers may locate in geographically distributed locations. Load balancing, job and workflow management systems such as HTCondor[5], Pegasus[6]are widely used to harvest idle computer cycles, map tasks to distributed resources, and manage input and output data.

Affordability of today's cluster computing systems, along with the de facto standard operating system Linux and job management software on these systems, computing power is much more accessible to researchers in the disaster research domain than ever before.

## 5.3 Data-intensive computing

Computation in the domain of disaster research is typically data-centric. Whether researchers are modelling natural phenomena, or synthesizing and analyzing data from multiple sources and at multiple scales, they not only need computing power, but also need a system that can get large volumes of data into their application and store output data efficiently. Applications that deal with large volumes of data (TB or more) and spend most of their execution time on data input/output (I/O) and processing are considered data-intensive, in contrast to the computation-intensive applications that spend most of their execution time on calculations.

Analogous to compute-intensive applications, parallelization of data-intensive applications typically involves partitioning the data into multiple segments or regions which can be processed independently using the same executable application program in parallel on an appropriate computing platform, then reassembling the results to produce the completed output data. Such parallelization tends to be straightforward and can normally scale linearly according to the size of the data.

Major challenges for data-intensive computing from the domain of disaster research are managing and processing exponentially growth in data volumes, significantly reducing associated data synthesis and analysis cycles to deliver results on-demand, e.g., assisting decision making, in a timely manner.

Innovations in system architecture and hard ware are extending the capabilities of supercomputers to better support data-intensive sciences. The emerging new systems aim to better

[4]Top500 supercomputer sites http://www.top500.org

[5]HTCondor open source distributed computing software http://research.cs.wisc.edu/htcondor

[6] Pegasus open source scientific workflow management software http://pegasus.isi.edu

meet the demands of the rapidly growing data-intensive scientific applications, for example, by including massive high-performance storage for data storage, large-scale Flash memory capable of reading and writing files at ~1 terabytes (TB) per second, processor accelerators (such as Intel Phi and NVIDIA GPU) to support data parallelism, as well as software tools and databases to support big data analytics and transfers.

## 5.4 Cloud Computing

The term *cloud computing* may mean different things to different people. It is as much a computing paradigm as an IT service model. Cloud computing builds on a number of technologies that came before the new term was coined, including distributed systems (remote computers), utility computing (service provisioning), and virtualization (multiple virtual machines running on the same physical hardware). Many IT services are now "in the cloud", such as provisioning of computation time, data storage, and applications (e.g., Amazon EC2, Microsoft Azure, Gmail, Microsoft Office 365, and Apple iCloud apps).

Five characteristics are considered essential in cloud computing by NIST (National Institute of Standards and Technology): On-demand self-service, network accessibility, resource elasticity, resource pooling, and metered service. Traditionally, research computing and business applications have always been very complicated and expensive. The amount and variety of hardware and software required to run them are daunting. A team of IT experts, such as system architects and administrators, is needed to design, install, configure, test, run, secure, fix and update the systems and software, a major barrier to many researchers and organizations of various sizes. In the world of cloud computing, service providers deploy tens of thousands computer servers in distributed data centres. Users of all sectors can access computing cycles, storage and applications remotely, wherever and whenever needed, over the network, by themselves. Users may add or reduce resources instantly according to the demand of their applications, and only pay for what they consume. Organizations, large or small, no longer have to carry the burden and cost of maintaining and updating their own hardware and software.

Several service models are available to meet various needs, including:

*Infrastructure as a Service (IaaS)*: IaaS clouds offer computers, physical or, more often, virtual machines, and can scale services up or down according to user requirements. Most of them also provide other value-added resources, such as VM disk images of various configurations, file or object storage, load balancers, firewalls, etc. Some research clouds allow researchers to select or create appropriate VM images of complex set-up for specific scientific applications (e.g., coupled climate models) that are time-consuming and need experts to configure.

*Platform as a Service (PaaS):*A PaaS cloud provides a computing platform, typically consisting of an operating system, programming language execution environment, often a web server and web service container, i.e., an environment for developers to create, host and manage applications without effort in managing the runtime system software, middleware, operating system, and other necessary tools required for development.

*Software as a Service (SaaS)*: SaaS is a software delivery model in which software applications are hosted centrally and accessed by users remotely. The cost of using SaaS is often based on subscriptions, e.g., monthly or annually. Various software technologies exist for users to remotely access SaaS services either as a desktop tool or via a web browser, such as Citrix ICA client, Remote Desktop, ThinLinc, and VNC (virtual network computing) software. More recently, the lightweight virtual container technology, represented by the likes of *Docker*[7], is becoming widely used by the research community to deploy and share scientific software.

[7] https://www.docker.com

The scientific community has embraced the cloud service models. While the classic HPC system and software environment support a variety of scientific computations, such as simulations, that can benefit from tightly coupled parallelism and the batch processing paradigm, the computational needs in the broader scientific community and vast application domains emphasize a high level of interactivity, availability of resources when needed (on-demand, albeit may run a little slower), and ease of use, all of which are promised by the cloud computing model. Research and academic institutions are adopting the IaaS model in making computing more affordable and accessible to researchers. The PaaS model allows researchers and engineers to focus on developing scientific software, rather than investing time and energy managing system software and tools. The SaaS model can deliver scientific applications and tools directly to the user, allowing researchers to access sophisticated software and computational resources transparently without the burden of installation and maintenance, thus significantly lowering the barrier for the broader user community of researchers, educators, and public as well. This model will, in particular, enable the sharing of software in a more effective and usable manner, overcoming many barriers researchers face when attempting to use software created by others.

The computational needs of the disaster research and mitigation domain encompass a wide range of applications, from modelling, data processing, data synthesis, data analysis, to visualization and decision support. Most of these can be well served by cloud computing service models.

## 5.5 Service-Oriented Architecture and Data Services

Service-oriented Architecture (SOA) refers to the computer software architecture pattern in which computational functions are provided as interoperable services among software applications, typically from different computers over a network. A service encapsulates a unit of functionality in a service interface for other applications to use. These services may operate on different platforms and frameworks from various vendors and providers. By using a standard communication protocol to describe its function, interface and data presentation, SOA makes it easy to compose complex software applications from a variety of functional building blocks implemented as services. For example, a decision support application may obtain time series data from a data repository and a map from a map server, and call up a geo-computation service to aggregate data based on geographical regions. Web services supports SOA implementations over the Internet. They make functions, or services, accessible over standard Internet protocols across different systems, frameworks and programming languages. Since the early web service standards such as SOAP/WSDL and many proprietary protocols by various commercial vendors, today the scientific community has fully embraced the REST (Representational State Transfer) Web service technology as it has become one of the most important technologies for web applications. Web services built on the REST architectural style for networked applications are called RESTful services. A RESTful service focuses on access to resources across the network and does so through standard ways of exchanging information between a client and the service, i.e., uniform interface for resource representation, discovery, information exchange and linking to other resources.

Web services are being adopted by the science community. Traditionally, researchers focus on experimentation and exploration in their own labs and settings. Research output was primarily in the form of publications of journal papers. Data sharing is now more widespread, especially through publicly available data repositories from government agencies and government funded research projects. However, many barriers still remain, preventing a high degree of utilization of shared data. Top on the list of challenges are the scientific understanding of the data and adapting the data in a usable manner for the study at hand. While existing Web service technologies such as REST can be adopted to help make scientific data more accessible and usable, common protocols of ***data services*** are needed to address the specific challenges associated with scientific data, especially in domains such as disaster research where data from multiple disciplines are common and necessary. Data services aim at making data easier to understand and use by using common protocols to communicate

the meaning and format of the data, in addition to access to the actual digital content of data. There are many on-going efforts in building data services. For example, the Data Observations Network for Earth (DataONE)[8] funded by the U.S. National Science Foundation federates earth science data repositories at distributed member data nodes through three coordinating nodes, and provides services and toolkits to be used by scientists to search and utilize remote datasets. U.S. Geological Survey (USGS) provides data web services for access to its data catalogs. Data services are also available at U.S. NOAA regional climate centres' ACIS system, providing access to metadata and various forms of the raw and processed climate data through public web services. Common data service protocols will be a critical part of a data infrastructure for the disaster research community.

### 5.6 Data science

Scientists have always conducted analysis on data, which they acquire through instruments, experiments and simulations, by employing mathematical and statistical methods. Today, the rapidly growing data size is changing the way we conduct scientific research, or simply, the way we approach the world around us. While the theoretical and fundamental understanding of natural phenomena remains important, progress in many fields is increasingly being driven by data analytics, visualization, mining and gaining insights from the data (e.g., genomics, biology, physics, cosmology, etc) (Halevy et al,2009). The term "data science" is often used to focus on the theories and techniques applicable to large volumes of data, ranging from mathematics, statistics, to information science and computer science. Such methods include probability models, machine learning, data mining, relational and non-relationship database, predictive analytics, uncertainty modelling, data visualization, and many others. The focus on handling "big data" is rapidly increasing our ability to understand and gain insights from data at a scale that was impossible before.

The growth of data science will benefit many fields, especially in risk and disaster management. This field is driven by ever increasing volumes of data in various forms, acquired by different instruments, historical and in real-time, structured and unstructured, from diverse sources including social media platforms. More and more, researchers in the disaster risk field use their data and analytical ability to find and interpret rich data sources, combine and synthesize datasets, build models, incorporate uncertainty in data, visualize data to aid in understanding, share data across domains leading to new insights, and communicate their findings and new insights from the data to the science community and general public. The new data science focus and emergence of new tools and software will significantly improve the productivity and efficiency of disaster research community. The next generation of the data infrastructure for the disaster research will need to facilitate and support data science capabilities.

## 6 Case Studies and Lessons learned on Linking Opened Data for Disaster Mitigation Around the World

In recent years, a number of international initiatives (listed below) were established to make data available for humanitarian and emergency response from international partners.

The International Charter for Space and Major Disasters, which was established in 1999, proposed the objective of utilizing space-based assets to contribute to the response to natural or technological disasters. The international Charter provided the strategy foundation for disaster data sharing.

In October of 2005, the ICSU 28th General Assembly agreed on the launch of Integrated Research on Disaster Risk Program (IRDR). IRDR aims to address the challenge of natural and human-induced environmental hazards. The DATA group of IRDR is committed to establishing

[8]DataONE project, https://www.dataone.org

disaster loss data infrastructure for stakeholders.

On December 16, 2006, the United Nations General Assembly agreed to establish the "United Nations Platform for Space-based Information for Disaster Management and Emergency Response (UN-SPIDER)". UN-SPIDER is the first to focus on the need for ensuring access to and use of Space-based Information throughout the disaster management cycle.

During China-US Roundtable Meeting of CODATA held on March 30, 2009, Prof. Li Guoqing proposed the idea of establishment of the Historical Disaster Data Grid (HDDG) to facilitate event-oriented data management and sharing for disaster scientific communities.

After the release of the report USGS Natural Hazards Response in 2012, USGS launched a project of Hazards Data Distribution System (HDDS). The HDDS site provides event-based download access to remotely sensed imagery and other datasets acquired for emergency response ;

In November 2012, European Space Agency made its contribution to GEO by releasing GEO Geohazards Supersites and Natural Laboratories (GSNL). It committed itself to data sharing of Geohazards, and formed the idea of building the Supersite.

China also published its National Spatial Data Acquisition and Emergency Data Sharing Platform in 2013, which aims to enable timely data coordination and sharing for emergency assistance(mainly for flood and earthquake).

In middle of 2014, CEOS (Committee on Earth Observation Satellite) proposed the idea of Recovery Observatory Infrastructure to further promote disaster data sharing.

In the sections below, we describe four of these initiatives in more detail to highlight the efforts and showcase the specific actions in making them successful.

## 6.1 SCU Disaster Loss Database

In the United States, many different federal and state agencies collect hazard data. For example, the U.S. Geological Survey collects geophysical and hydrological data, and the U.S. National Ocean and Atmospheric Administration focuses on atmospheric and hydrometerological data. Only some parameters of a hazard event are monitored consistently (e.g., magnitude, date of occurrence, location), while the document of others (e.g., deaths, injuries, or economic losses) in incomplete and inconsistent. In the absence of a national inventory of hazardlosses the Hazards and Vulnerability Research Institute at the Univeristy of South Carolina developed the Spatial Hazard Events and Losses Database for the United States (SHELDUS®, www.sheldus.org). In the early 2000s, this database was originally supported by grants from the National Science Foundation (Grant No. 99053252 and 0220712) and the University of South Carolina's Office of the Vice President for Research.Periodic support for the database has been provided by the South Carolina Emergency Management Division for South Carolina updates. Given the lack of consistent federal, or private sponsorship in subsequent years, SHELDUS® switched to a user-fee publicly accessible hazard loss database in 2014.

SHELDUS®is a U.S. county-level hazard data set covering 18 different natural hazard types and a time span from 1960 to present (Table 1). The data originatefrom several existing national data sources such as the National Climatic Data Centre's monthly Storm Data publications and the National Centers for Environmental Informations' (formerly National Geophysical Data Centre)Tsunami and Earthquake Event Databases. Event-related information includes location (state and county), deaths, injuries, property losses, crop losses, and beginning/ending dates.

Originally, SHELDUS® contained only those events that generated more than $50,000 in damage or at least one death. Since the release of SHELDUS®Version 13.1 these thresholds have been removed and SHELDUS®now includes every loss-causing hazard event since 1960. The annual database updates include new data releases (i.e., losses that occurred in the most recent calendar year),

data correction, data additions using supplementary data sources, and/or new data download and analytical features.

In addition to providing georeferenced loss information, SHELDUS®resolves historic boundary changes related to geography (e.g., new counties) as well as climatologoical forecast zones, the latter of which changes numerous times during a calendar year. The database also matches hazard event records with other identifiers (e.g., Presidential Disaster Declaration Numbers (PDDs), the Global Identifier Number (GLIDE) and the Billion Dollar Disasters product produced by NOAA's National Climatic Data Center.The GLIDE number (http://glidenumber.net)is an important feature because it is a globally-accepted identifier linking the data in SHELDUS®to international databases.

**Table 1 . Overview of loss information inSHELDUS®**

| SHELDUS® version 13 (1960-2013) | |
|---|---|
| Number of Records: | 831,182 |
| Direct Property Losses (in $2013) | $655.7 Billion |
| Direct Crop Losses (in $2013) | $148.5 Billion |
| Fatalities: | 31,495 |
| Injuries: | 235,739 |
| Costliest Year: | 2005 ($120 Billion) |
| Costliest Hazard Type: | Tropical Cyclones ($260 Billion) |
| State with Highest Losses: | Florida ($105 Billion) |
| State with Most Fatallities: | Texas (2,202,) |

*Source:Spatial Hazards Event Loss Database for the United States (http://sheldus.org)*

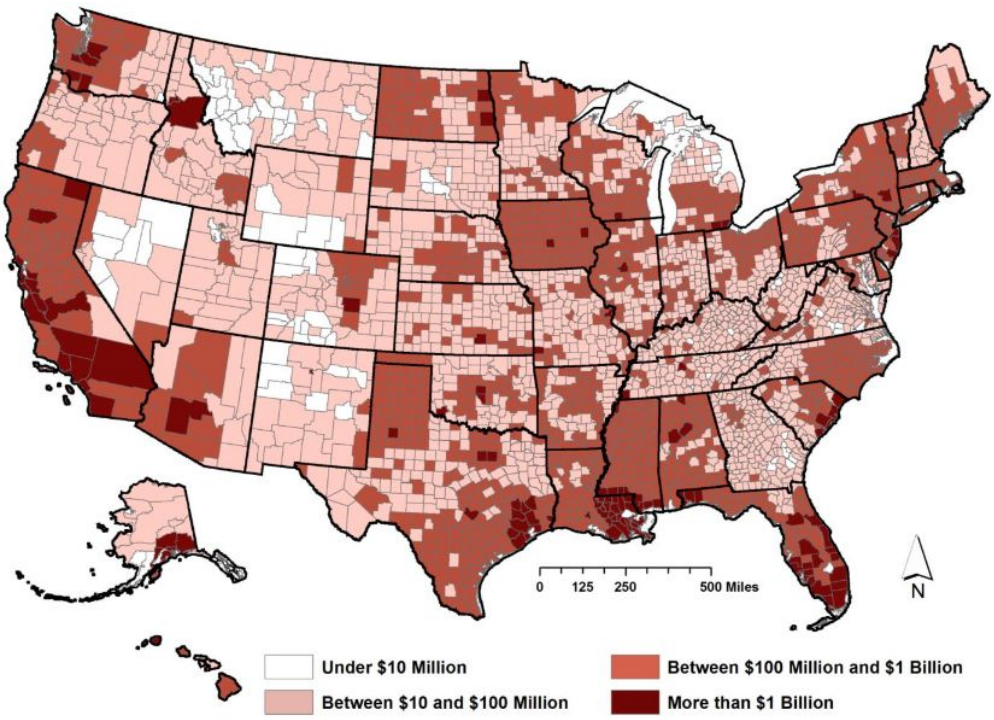



**Figure 1. Spatial distribution of direct, economiclosses (property and crop) from natural hazards between 1960 and 2013 (in billion,US$2013)**

*Source:Spatial Hazards Event Loss Database for the United States (http://www.sheldus.org)*

***Functionality***

The SHELDUS®website provides online data download functions along with maps, reports, references to SHELDUS®applications, and FAQs. There are four querying options: location, PDDs, GLIDE, or named major disasters. Under the *Location* option, users can download data either in aggregated form (e.g., by location, time, and/or hazard type) or individual, event-specific records (Figure 2). Furthermore, the *Location* option offers additional query options (e.g., time period, hazard type) to create user-defined data sets.*Named major disasters* allows users to query based on more readily identifiable events such as Hurricane Katrina, Superstorm Sandy, or the 1993 Mississippi River Floods without knowing the precise location of the event and its attributed losses. All the counties affected by the particular event are included in the query.

At its finest resolution (i.e., without any data aggregation), SHELDUS®outputs include the begin and end dates of the event, the hazard type, state, county, injuries, fatalities, property damage, and crop damage. Additionally, SHELDUS®offers the choice to retrieve economic loss data either in current U.S. dollar and/or as inflation-adjusted loss data.

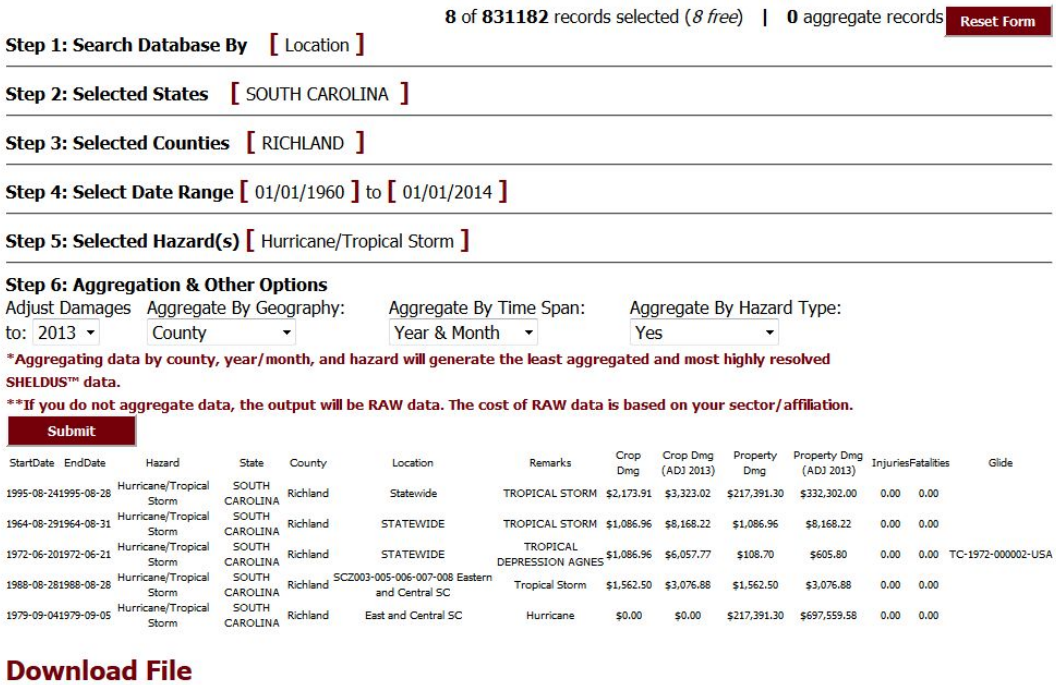



**Figure 2. Location-based search interface of SHELDUS®**

*Source:Spatial Hazards Event Loss Database for the United States (http://sheldus.org)*

***Applications***

The SHELDUS®database is widely used, particularly among researchers and planners (Figure 3). As of May 2015, there are more than 85 peer-reviewed publications and 50 thesis/dissertation that heavily utilize SHELDUS® data. The topics of research range from resilience models for the tourism and hospitality industry to statistical analyses on fat-tailed distributions. The most frequent users of SHELDUS® come from disciplines such as economics, geography, environmental studies, engineering, climatology, urban planning and public policy.

In the realm of planning, SHELDUS® data are extensively integrated into local and state hazard mitigation plans. Since 2000, local and state governments must maintain such plans in order to remain eligible for federal disaster dollars. The goal of these plans is to identify mitigation actions that reduce the impacts of natural hazards. A central component of a hazard mitigation plan is a risk assessment,

which details the type of hazards affecting a community as well as historic events and their losses. This is where SHELDUS® data play an important role because it allows planners to download historic event and loss information for the locality of their choice without any additional post-processing. Nearly half of all U.S. states and countless numbers of counties use SHELDUS® in ther risk assessment portion of hazard mitigation plans. There is also an increasing use of SHELDUS® in climate adaptation and resilience plans.

While research and planning represent the main usage areas of SHELDUS®, there are creative uses of the data elsewhere. For example, college instructors have incorporated SHELDUS® into their lesson plans and even museums' displays have drawn on SHELDUS® to communicate the socioeconomic impacts from natural hazards on the U.S.

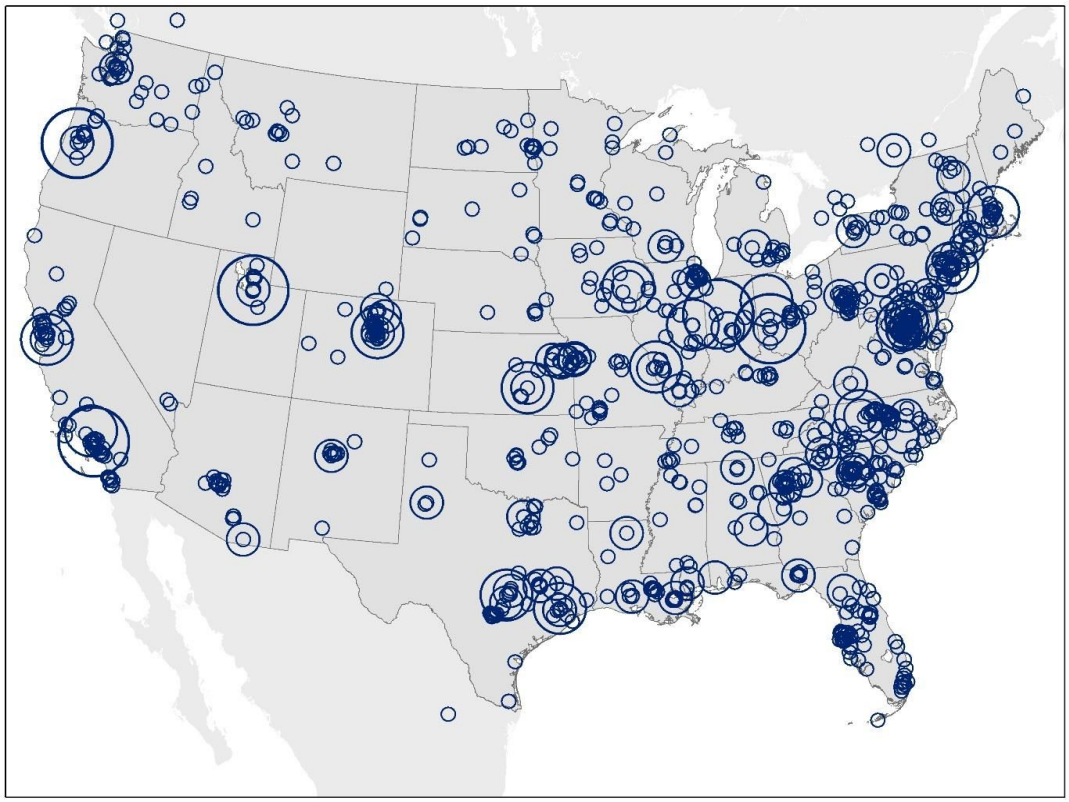


**Figure 3. Locations of SHELDUS downloads between July and December 2012. The size of the circle represents the number of downloaded SHELDUS records.**

## 6.2 USGS HDDS

Remotely sensed datasets such as satellite imagery and aerial photography can be invaluable resources to support the response to and recovery from many types of emergency events such as floods, earthquakes, landslides, wildfires, and other natural or human-induced disasters. When disasters strike there is often an urgent need and high demand for rapid acquisition and coordinated distribution of pre- and post-event geospatial products and remotely sensed imagery. These products and images are necessary to record change, analyze impacts of and facilitate response to the rapidly changing conditions on the ground.

The primary goal of U.S. Geological Survey (USGS) Emergency Response project is to ensure that the disaster response community has access to timely, accurate, and relevant geospatial products, imagery, and services during and after an emergency event. The USGS Hazards Data Distribution

System (HDDS) provides quick and easy access to the remotely sensed imagery and geospatial datasets that are essential for emergency response and recovery operations.

The USGS HDDS serves as a single, consolidated point-of-access for relevant satellite and aerial image datasets during an emergency event response. The coordinated and timely provision of relevant imagery and other datasets is an important component of the USGS support for domestic and international emergency response activities.

The disaster response liaison provides access for satellite tasking, imagery acquisition and distribution, image registration, creation and distribution of disaster-extent maps, and Web-based mapping services for products, and for pre– and post-disaster data storage and distribution. The USGS Earth Resources Observation and Science Center (EROS) staff also work on disaster preparedness, including providing basic pre-event data such as satellite images, vector data layers, and other pre-disaster data layers.

The Hazards Data Distribution System, as a disaster response system, incorporates satellite tasking and data acquisition, product development, Web applications, and data storage.

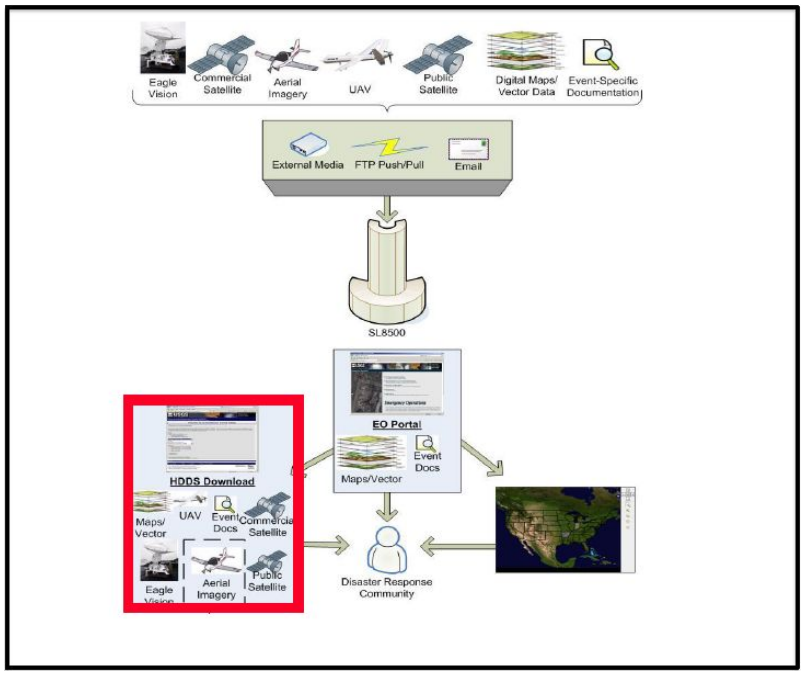



**Figure 4. Hazards Data Distribution System (HDDS)**

Data will be acquired for each event that meets specified criteria. The data will be obtained from the providers whose imagery meet the criteria needed for responding to the event. The data will be made available around the clock and acquired in a timely fashion by USGS staff. The data will be processed up to a standard level, preferably precision terrain corrected. If that level of processing is not possible, then the next highest level will be sought. The USGS EROS will also take on the responsibility of invoking the International Charter for Space and Major Disasters. The data provided by the Charter is free of charge but may have provider distribution restrictions that need to be enforced.

In the product development process, any value-added processing that is done can be made available via a web-based delivery system.

The USGS EROS can provide data sets from Landsat 1-5, 7, and 8, digital orthophoto quadrangles, digital raster graphics, digital line graph, and digital elevation models in accordance with applicable distribution policies and agreements. Other data sets that can be supplied include Advanced

Spaceborne Thermal Emmission and Reflection radiometer (ASTER), Moderate Resolution Imaging Spectroradiometer (MODIS), Advanced Very High Resolution Radiometer (AVHRR), Hyperion, Advanced Land Imager (ALI), IKONOS, Quickbird, WorldView, SPOT, Radarsat, aerial photography, and Disaster Monitoring Constellation (DMC).

The data will be delivered in standard formats and map projections via the web, ftp, and media. The media deliveries will occur only if the web and ftp are not functioning as viable alternatives to a speedy response. All participants will have their own unique user code/password to access the HDDS. Access to the restricted/licensed imagery will be requested and granted on an event basis.All new data acquisitions will be distributed as licensed to promote the sharing of information by as many participating agencies as possible. USGS EROS recognizes the importance of agencies working together to respond effectively to a major terrorist incident or natural disaster. Therefore, participating agencies are encouraged to submit their value-added products or other data layers for distribution on the HDDS.

Currently, HDDS holds over 354 Tb of data from over 8.8 million files, specifically, corresponding to over 1300 baseline and disaster events. Public access data comprises about 248TB (7,761,504 files), which provides general public with data of unrestricted access; restricted data holdings account for approximately 106 TB (1,116,571 files),which provides data for designated emergency response agencies with password-protected access.

The functions of interactive HDDS interface include:

- Immediate download access to event-related imagery.
- Geographical data visualization with browse image and footprint area overlay.
- Extensive metadata available in multiple formats (for example, CSV, KML, SHP, XML, FGDC).
- XML-based Really Simple Syndication (RSS) feeds for newly ingested data.
- Registration service and log-in access for restricted datasets.
- Standing request services
- Custom web-mapping services
- Bulk data download capability.

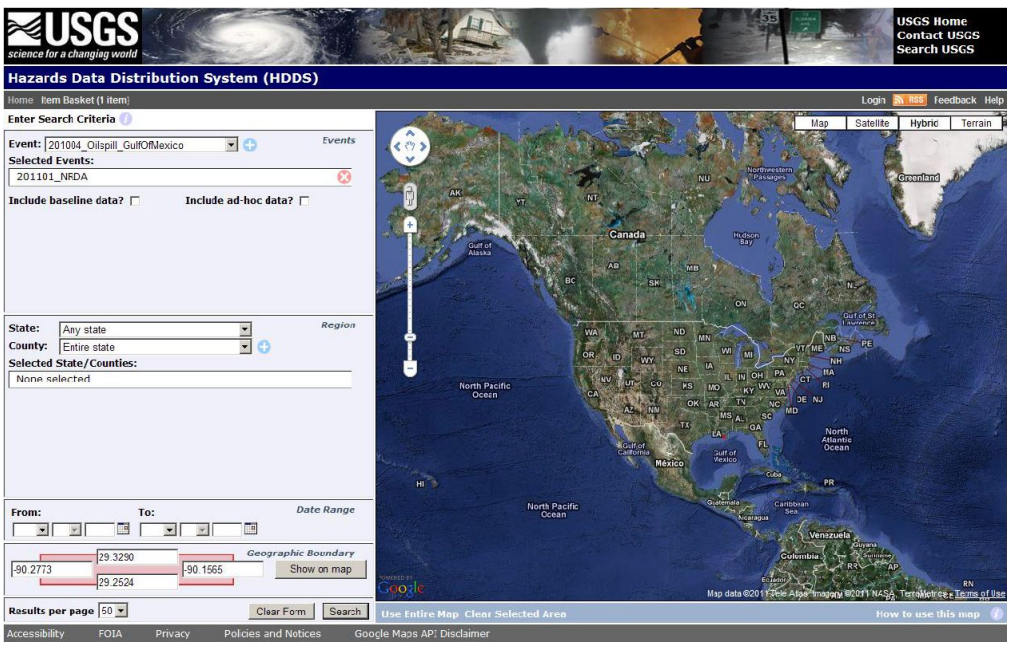



**Figure 5. HDDS2 - Graphical Interface of data retrieval**

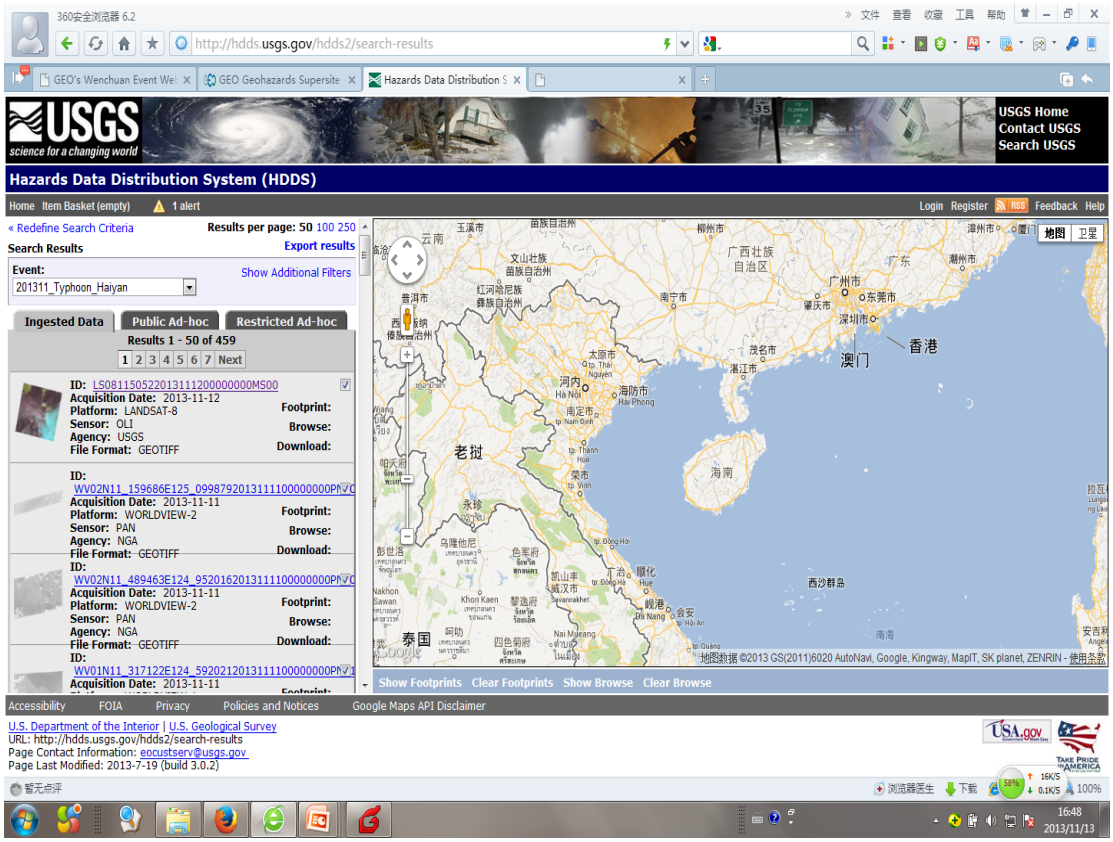



**Figure 6. Home Page of the USGS Hazards Data Distribution System (HDDS)The HDDS site provides event-based download access to remotely sensed imagery and other datasets acquired for emergency response.**

The data can be easily accessed in standard formats, and the response community will have the opportunity to share their value-added products with one another.

- Data sharing among agencies ensures that the same images are being used. This will allow for cooperation and sharing of value-added products, because standards are in place.
- Near-line and off-line archiving and retrieval ensure that the data is preserved for historical evaluation and reuse. This also saves dollars for future studies and for response to events that occur in the same location.

## 6.3 ESA SuperSite

This section summarizes the capabilities of the ESA SuperSite. For a more complete description, please refer to the Appendix B.

As far as the value-added products are concerned, ESA focuses effort on hosted processing capabilities offered by the ESA Geohazards Exploitation Platform (GEP), providing users with a range of data processing and dissemination services. In general, the processing capabilities are made available to users as-a-Service, where the users can define a set of input data and processing parameters, and trigger the execution of the algorithms.

In order to support the development of new algorithms, the ESA GEP also provides a Platform-as-a-Service capability (PaaS). In particular, Cloud Sandboxes enable developers and integrators to easily implement new algorithms. This solution utilizes the Virtual Machine technology and employs a middleware to provide a transparent interface to Cloud services. The cloud services are used to scale up the processing when the dimensions of the input dataset increases. GEP offers direct

access to dedicated Cloud Sandboxes to researchers in MARsite who are interested in experimenting with their own algorithms on ESA data.

The use of GEP removes the need of transferring huge amounts of input product data from the ESA archives to the users' machines, resulting in significant savings for the users. With GEP, the agency is providing partners with tools and infrastructure aimed at supporting geohazards researchers and practitioners with easy and open access to the ESA sensors data, community knowledge andexpertise,and collaborative research.

In the geohazard domain, science users require satellite EO to support mitigation activities designed to reduce risk. These activities are carried out before the earthquake (or other geological peril) occurs, and they are presently the only effective way to reduce the impact of earthquakes on society. Short-term earthquake prediction today offers little promise of concrete results. The assessment of seismic hazard requires gathering geo-information for several aspects: the parameterization of the seismic sources, knowledge of historical and instrumental rates of seismicity, the measurement of present deformation rates, the partitioning of strain among different faults, paleo-seismological data from faults, and the improvement of tectonic models in seismogenic areas. Operational users in charge of seismic risk management have needs for geo-information to support mitigation. Satellite EO can contribute by providing geo-information concerning crustal block boundaries to better map active faults, maps of strain to assess how rapidly faults are deforming, and geo-information concerning soil vulnerability to help estimate how the soil is behaving in reaction to seismic phenomena (read more from https://geohazards-tep.eo.esa.int/#!pages/initiative).

An "Exploitation Platform" refers to a virtual ICT environment, often cloud-based, providing users with very fast access to: (i) a large volume of data (EO/non-space data), (ii) computing resources (e.g. hybrid cloud/grid), and (iii) processing software (toolboxes, RTMs, retrieval schemes and visualization routines).

The idea underpinning exploitation platforms is to enable users to perform effectively data-intensive research by providing them with a virtual machine running dedicated processing software close to the data, thereby avoiding moving large volumes of data through the network and spending non-research time on developing ICT tools. The GEP portal is already accessible online for users (including public access level) at: http://geohazards-tep.eo.esa.int(Figure 7).

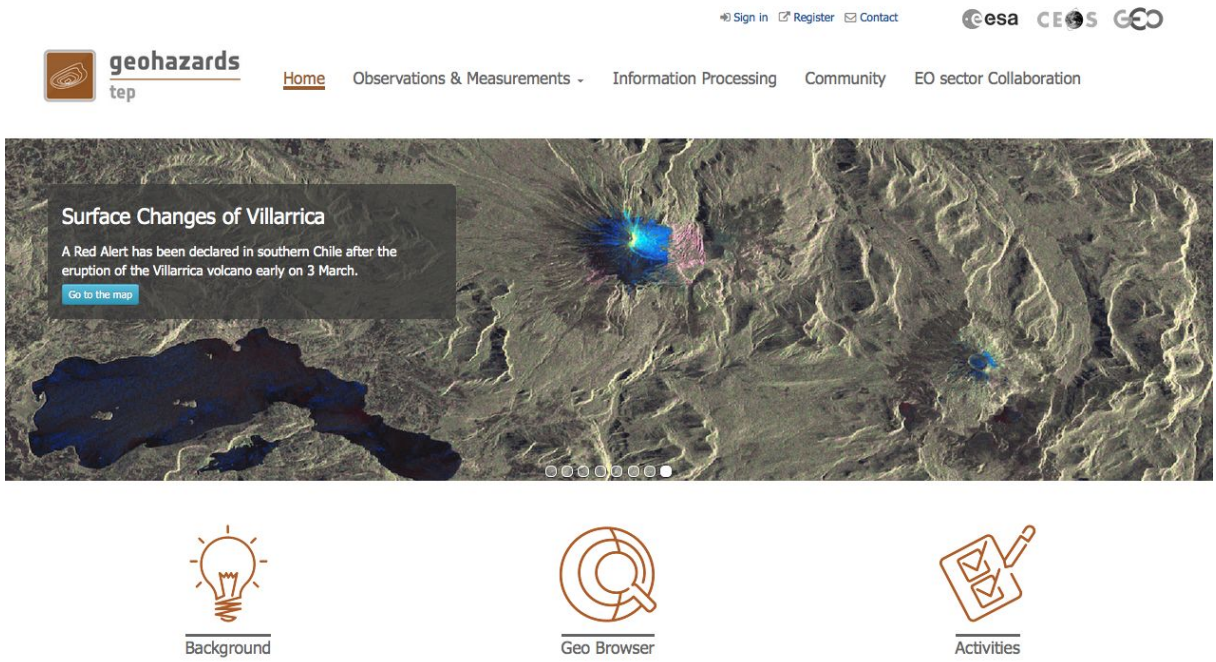



**Figure 7 - The Geohazards Exploitation Platform portal**

It is providing the following set of capabilities, which are made available to the MARsite partners.

**EO data discovery service** through a single point of access to visualize data collectionsin terms of acquisition footprints and sensor parameters, with resources available from ESA missions (especially the SAR missions from ENVISAT, ERS and SENTINEL-1) and third party missions (currently DLR TerraSAR-X and, upcoming for, ASI Cosmo-Skymed and CNES).

**EO data access service over distributed repositories** supporting the dissemination of imagery either stored in the GEP cloud platform environment or accessed through the GEP portal in other remote data repositories from the pool of contributing agencies. Data access is based on the authentication of registered users and the granting of data dissemination according to the user profile. For instance, EO data constrained by license terms and distribution restrictions can be accessed from the platform's geographic interface via active links to the repository of the data provider.

**Accounting service for EO data consumption** allowing the monitoring of the volumes of data use per EO source and according to the activity associated to the user profile. The accounting service can be used to support reporting concerning the exploitation of EO data, either from the Platform (e.g. hosted processing) or in the framework of application projects (e.g. CEOS pilots).

**EO processing services** for on-demand processing, exploiting software to transform EO data into measurements; the user may run an EO processor provided on the platform (ready to use software-as-a-service, SaaS), or integrate an application he/she has developed (platform-as-a-service capabilities, or PaaS).

The SaaS Processing can be invoked either interactively though a web browser, or through scripting using the OGC Web Processing Service (WPS) interface; The PaaS provides software development and integration tools, and enables users to perform their data exploitation activities with large flexibility and autonomy, by using one or several virtual hosts directly provisioned on the cloud platform and deployable on demand.

**Access to Value-Added products** generated on the GEP, or products contributed by third parties. The platform allows cataloguing and dissemination of products relevant to the geohazards community. It can be used to provide access to elaborated products in support of thematic exploitation goals.

The platform is currently in its Validation phase, which will explore up to October 2015 a growing set of functionalities. As of today, the Platform already allows public data search over a large archive of EO data from ESA sensors and from third-party missions (CEOS partners) like DLR's TerraSAR-X.

Practitioners can access the GEP infrastructue through the Geobrowser service, which is now made available with a set of baseline functionalities to all users (unregistered users / general public) for data search, data selection and data processing at: http://geohazards-tep.eo.esa.int/geobrowser (Figure 8).

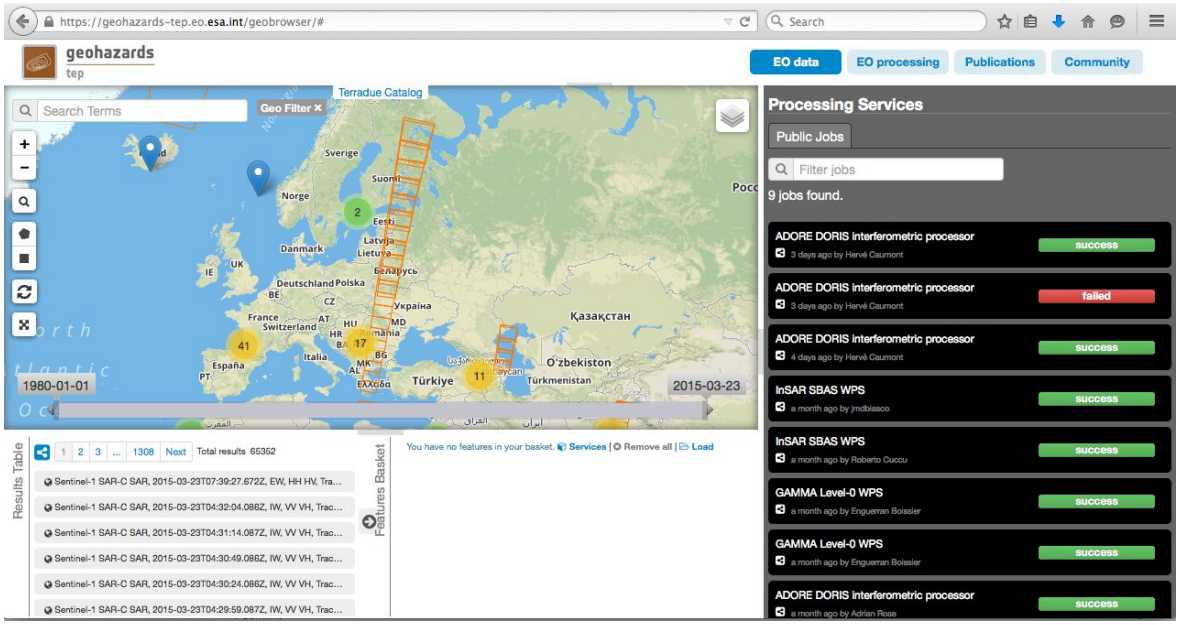



**Figure 8 - User access to the Geohazards Exploitation Platform**

The user documentation for GEP is also available online. It features an overview of the Platform concepts, a Community Portal User Guide, a Cloud Operations Administrator Guide and a growing set of data processing tutorials (SAR processing with ADORE DORIS, GMTSAR, ROI_PAC, and a set of G-POD services such as GAMMA-L0, SBAS). This user documentation will continue to evolve in the coming months (currently available at: http://terradue.github.io/doc-tep-geohazards/).

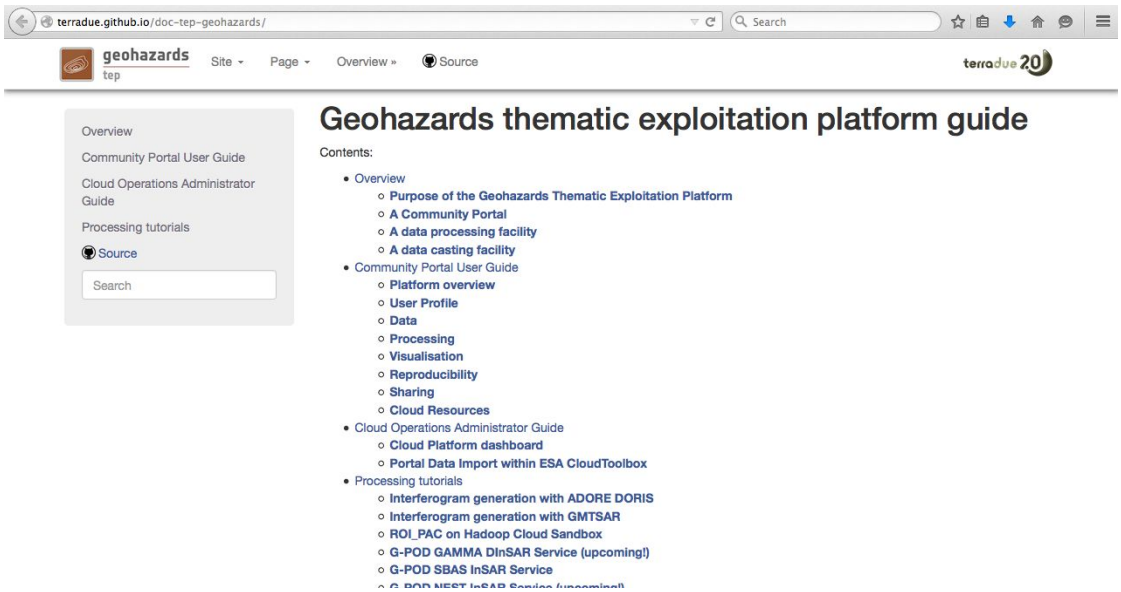



**Figure 9. User documentation for the Geohazards Exploitation Platform**

## 6.4 The "Disaster Reservoir" Project in China

When the Wenchuan earthquake occurred on May 12, 2008, transportation and communication infrastructure collapsed and the area was devastated. In relief efforts shortly after the earthquake,

various governmental departments urgently needed timely support of quality aerial remote sensing data. In the post-quake phrase, various departments needed to assess losses of the quake-hit areas in a timely fashion so as to appropriately guide the reconstruction. In order to provide real-time disaster information to the disaster relief command post, the earthquake relief campaign mobilized and assembled military and civilian satellite and other aerial remote sensing resources to obtain a number of satellite and aerial images, which provided powerful support to Wenchuan disaster relief. As evident through this effort, nonetheless, it was apparent that China lacks the technical system supporting coordinated emergency response to disaster, and the inadequate sharing service capability of spatial data severely impedes the data from playing its due role in disaster relief, in terms of access to and application of spatial data in disaster relief. In July 2008, the National Sci-tech Support Plan launched the project of "National Emergency Coordination System of Spatial data Acquisition and Application and Data Sharing Service Platform", to research spatial data acquisition and application, emergency coordination plan and dispatch technique, emergency data sharing technique, acquisition technology of aeronautical and spatial data, and spatial data application emergency cooperation mechanism. This endeavour aims to build and share the national spatial data acquisition and application emergency coordination system and data sharing service platform by incorporating both military and civilian strengths, enabling the use of spatial technology in nationwide disaster prevention and relief.

According to the Remote Sensing Pre-plan of National Important Public Emergencies, in case of national Level I or II emergency response, the spatial data acquisition and application emergency coordination platform shall plan the remote sensing data acquisition tasks and disseminate the data acquisition orders to corresponding military and civilian aeronautical and spatial data acquisition agencies which are required to acquire and process the data according to task requirements and send the assembled spatial data products to spatial data emergency sharing service centre that archives, manages and sends the data to military and civilian users. In this way, military and civilian aeronautical and spatial remote-sensing resources serve civilian purposes in peace and military purposes in war, substantially enhancing the application efficiency of the resources.

The emergency service system has activated the Ministry of Science and Technology of PRC – the PLA General Staff Department work mechanism on military-civilian remote sensing information emergency communication, deploying military satellites to get imagery of disaster-hit areas for civilian purposes, and designating the Ministry of Science and Technology as the general coordinator of data serving civilian purposes. After the natural disaster, the PLA General Staff Department deployed satellites to gain imagery of quake-hit areas and pre-quake information had been incorporated into the emergency platform of the disaster reservoir. Besides, the PLA Air Force General Intelligence Station (Aerial Remote Sensing Department I of National Remote Sensing Centre of the Ministry of Science and Technology, PRC) dispatched two aerial remote sensing planes to quake-hit areas within two hours after the earthquake. The aerial remote sensing images thus obtained served functions in earthquake relief.

The Ministry of Science and Technology of PRC has contacted satellite information acquisition units of UNESCAP, France, the United States via the international cooperation mechanism. China has agreement with approximately ten satellites operated by different countries that they will share their information acquired during the most recent two days. The Ministry of Science and Technology will step up communication and coordination with them, and timely disseminate data to related domestic units.

The relevant departments, bureaus and remote sensing centre of the Ministry of Science and Technology timely contacted the administrative organs and research institutions of Sichuan Province to establish remote sensing information service mechanism. In view of the needs of disaster-hit zones, they keep track of the disaster, coordinate remote sensing data supervision and evaluation, assist

locals in carrying out scientific disaster relief and, in some cases, dispatched personnel to organize many unmanned aerial vehicles (UAV) to engage in aerial remote sensing team's tasks.

The practice drill on emergency response to Qinghai Yushui Earthquake held on April 15, 2014, tested the capacity of systematic use of space technology to cope with public emergencies and achieved anticipated improvements. In the same day after the earthquake, the aerial remote sensing planes flew to quake-hit zones for 0.4-meter high resolution remote sensing images, the first-hand materials for seismic resistance and disaster mitigation in Yunshu. Meanwhile, the emergency coordination platform acquired, according to plan, archived historical data and programming data of military-civilian satellites, including data from meteorology satellites, resource satellites, a series of remote-sensing satellites, Beijing-1 Satellite, foreign commercial satellites, which provided substantial support for monitoring and evaluation of Yushu Earthquake as well as post-disaster reconstruction.

After Lushan Earthquake (2013), the Ministry of Science and Technology, PRC, immediately requested all related departments to launch the Ministry's Lushan Earthquake Space Remote-sensing Information Emergency Consultation Centre, coordinated the establishment among various departments the national spatial data acquisition system for emergency coordination and data sharing service platform, also called Disaster Emergency Data Reservoir (DEDR), opened emergency data sharing channel at 11:50 on April 20, 2013, and coordinated the shared use of data from international and domestic satellites and aerial remote sensing data to provide remote sensing monitoring data to relevant departments of the country and local governments. Through the emergency exchange and collection mechanism for remote sensing data deployed by the Ministry of Science and Technology after Wenchuan Earthquake, data from many sources was disseminated to where it was needed within 12 hours after the quake, including the resource satellites, environment satellites, ocean satellites, SPOT and LANDSAT as well as SJ-9A pre-quake data from China Centre for Resources Satellite Data and Application, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, National Satellite Ocean Application Service, 21st Century Aerospace Technology Co., Ltd., Beijing Remote Sensing Information Institute andAstrium-geo along with catalogued meta-data of military satellites are sent to various departments at their requests. Up to 10:00 of April 26, 2013, the DEDR has pooled satellite and aerial imagery of key disaster-hit areas at 112 GB, including pre-quake imagery at 61 GB and post-quake at 51 GB. The full set of information is timely provided to 40-plus units of 19 ministries and local organs; in total the remote sensing data about disaster-hit areas topped 2,000 GB.

After the earthquake, various units have significantly enhanced their capabilities in data acquisition, business coordination, resource sharing and large-scale application; they actively acquire and share pre-quake and post-quake remote sensing data and contribute to the operation of the DEDR. The participating organization include China Centre for Resources Satellite Data and Application, BeijingAstrium-geo Satellite Image Co., Ltd, Beijing Eastdawn Information Technology Co., Ltd, 21st Century Aerospace Technology Co., Ltd. and China Centre for Resources Satellite Data and Application.

Still in its initial phase, the DEDR project has established a coordinated emergency response framework and technical system of military-civilian earth observation resources. It has a long way to go. In particular, the co-development and dual nature of a shared mechanism for remote sensing resources and data deserve further study.

The "Disaster Reservoir" Project demonstrated the importance of governmental role in leading and building up the data exchange technology infrastructure to overcome boundaries and self-interests of departments and the limitations of delay in decision making during emergency response among various organizations. The standardized technical interface adopted enabled the interoperation of multiple data sources, hence accelerating the timeliness of disaster data response.

# 7 New-generation Disaster Data Infrastructure (DDI)

In this section, we discuss the main components of the next-generation disaster data infrastructure.

## 7.1 Main Characteristics of Disaster Data Instructure

Spatial Data Infrastructure (SDI) has been developed and has evolved in the past two decades. It is a framework and system that connect users to spatial data. SDIs are not merely data repositories where the spatial datasets are stored, they help users discover, understand, view, access and query geographic information of their choice from local to the global level, for a variety of uses. On many levels, a Disaster Data Infrastructure (DDI) is similar to SDIs, with a goal of connecting users to disaster related data and information and the relevant tools, often with a strong focus on certain events and timing, for various objectives ranging from natural disasters preparedness, monitoring, and response, to land use planning and impact assessment. At the technical level, disaster data infrastructure adopts a technical system similar to that of spatial data infrastructure. **Figure 10** illustrates the three layers in an interoperable SDI eco system and the functional services provided by each layer. The Physical SDI consists of data repositories that provide data storage, data organization and metadata, along with Application Programming Interfaces (API) and other types of service interfaces for data access. The Federated SDI enables interoperability among the various SDI repositories and databases, providing clearing house access to broadly distributed SDI resources. The Virtual SDI provides on-demand access, processing, analytics and visualization "in the cloud", enabling end users to access SDI resources anywhere any time around the globe.

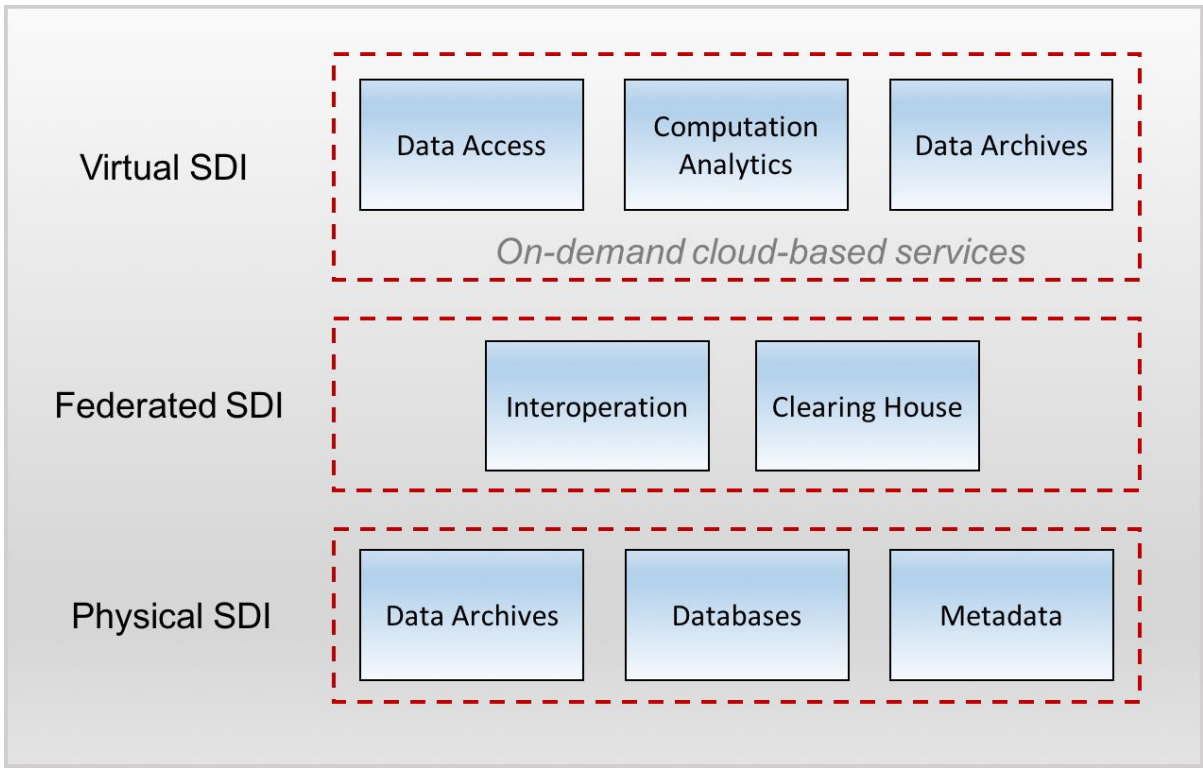



**Figure 10. An illustration of an interoperable SDI eco system**

A disaster data infrastructure has several characteristics due to the nature of disaster management. It must operate on-demand in real-time; it must be open; and it must be easy to access and use.

When a natural disaster occurs, relevant data and information need to be available to various groups in the shortest time possible. Being able to locate the data and dispatch to where it is needed in real-time is critical in preventing losses. Data usage may range from updating status, assessing the affected area, to assisting with decision making and directing relief resources and efforts. As past experiences have shown, it is necessary to pool relevant data from multiple sources, often from different organizations, regions and countries, in a short period of time, during the of emergency response phase. Data processing, analysis and visualization also need to be done in response to demands of different usage modes, from dynamic monitoring and evaluation of the disaster, conducting emergency relief effort, continuous monitoring, comprehensive evaluation, to aiding decision-making related to secondary disaster risk pre-warning. In addition, different governments and organizations usually have one or more corresponding SOPs (Standard Operating Procedure) directing the responders to take proper action. A Standard Operating Procedure (SOP) is a set of specific rules and actions need to be taken when a disaster occurs. Therefore, DDI must be integrated with SOPs so as to maximize its utilization. Figure 11 illustrates the relationship between DDI and SOP. Before a disaster, DDI must provide sufficient prediction data for different governments or organizations to issue an early warning. During a disaster, DDI should quickly parse collected data, and get ready to dispatch to relative units to respond rapidly to the disaster, such as issuing emergency alerts and grouping the first responders. After a disaster, data should be further processed, analyzed, visualized and dispatched to responders to a disaster, such as rescuing victims, setting up shelters, and providing emergency medical care. Furthermore, DDI needs to carefully verify crowd-sourced disaster information so that the rescue operations can be performed correctly and efficiently.

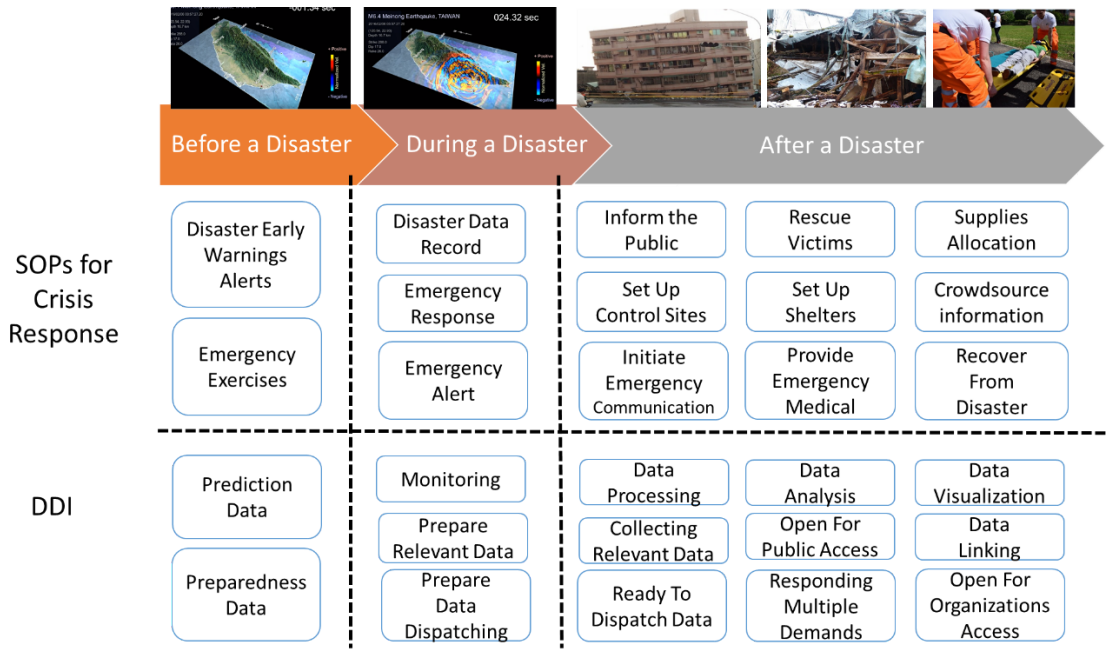



**Figure 11. An illustration of the relationship between DDI and SOPs**

As stated in the earlier chapters, disaster related data resources are often scattered across governmental departments and other organizations. The users of the data, e.g., organizations responding to disasters, are also distributed across organizational and geographical boundaries. Thus,

DDIs must be able to link users to the data they need, along with information (metadata) to facilitate use. Analysis and visualization tools should also be linked to data, ensuring that the tools are usable on the data set without additional barriers (e.g., software installation, compatibility, data conversion and format incompatibility).

A disaster data infrastructure must be open for multiple organizations to contribute data as well as consume data. Different departments with varying responsibilities should have full access to all the relevant data, as a disaster unfolds, to conduct emergency observation of the disaster-hit area and quickly acquire spatial data about the affected areas by coordinating all available spatial data sources. Therefore, DDI is required to have the capabilities of a unified and coordinated spatial data management and overall scheduling to deploy resources to meet the demands of rescue, recovery and relief efforts in a disaster situation. Meanwhile, the emergency observation tasks specified by various departments need to be executed in a unified and coordinated manner to avoid data redundancy, ensure the complete coverage of data in both space and time, and meet the basic demands of data acquisition at key time nodes throughout the disaster relief process.

DDI must also be open for public access. Disaster relevant information largely comes from the citizens in the impacted areas, therefore the acquisition of disaster data depends on public participation. At the same time, dissemination and use of disaster data also concerns the affected people and the public. Making the data and information available is an important part of government function. The construction of a DDI needs to take into account methods for public data dissemination. The public facing nature of DDI is a more urgent need than in traditional spatial data infrastructure which usually allows target consumers free access but charge fees for others. DDI development must also consider participation of the international community. International cooperation is often needed when one region is affected by a disaster, e.g., when remote sensing data is essentially the only effective real-time data source and the data has to come from another country’s satellites.

Currently, the international community attaches great importance to Disaster Data Infrastructure's role in disaster prevention and reduction. The Sendai Framework published by world conference on disaster reduction in March 2015 has set disaster information systems and infrastructure as one of the key tasks of future disaster reduction. The United Nation has established the international disaster reduction cooperation mechanism based on International Charters for Space and Major Disasters. This mechanism, via satellite resources contributed by its member organizations, offers free satellite remote-sensing data and information to countries hit by grave disasters to assist them in monitoring and evaluating the disaster.

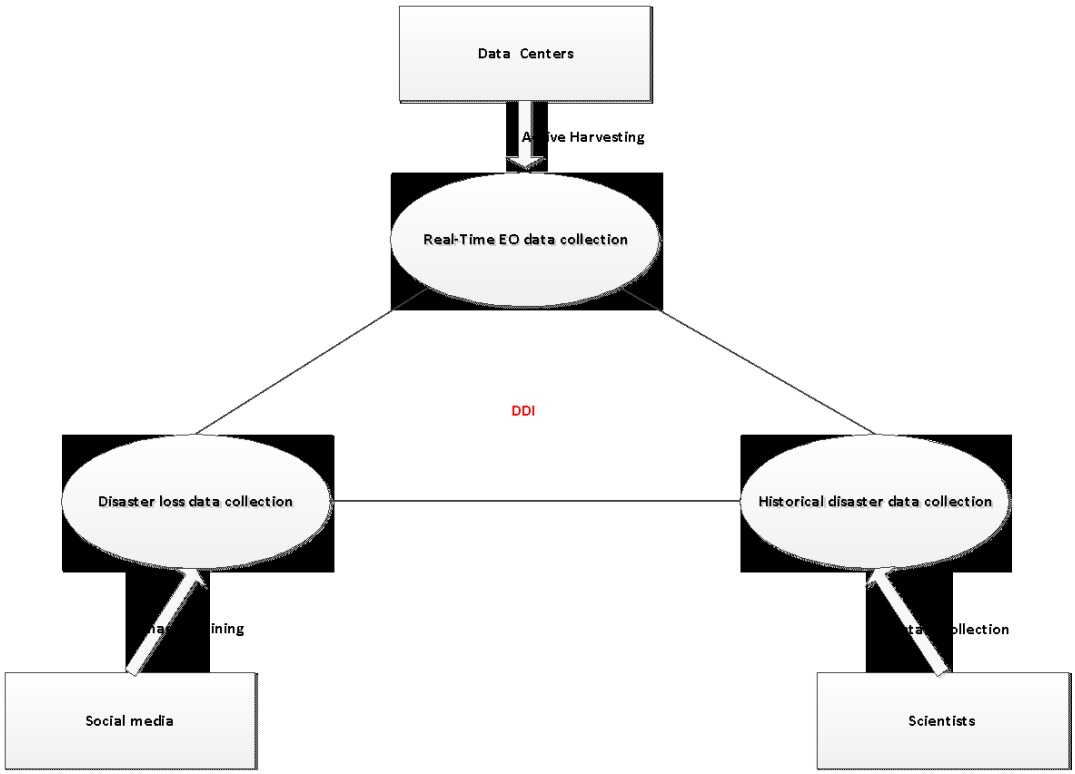



**Figure12. An architecture for Disaster Data Infrastructure as a framework linking disaster emergency data collection infrastructure, historical archived data infrastructure and disaster loss database.**

## 7.2 Disaster Emergency Data Infrastructure

Disaster Emergency Data Infrastructure (DEDI) realizes the integrated disaster data acquisition and transmission to users immediately and conveniently over the network of space borne-airborne-ground facilities to enable rapid response to disasters. Data supporting the disaster response community may come from multiple disciplines and sources, including hydrology data, meteorology data, basic geographical data, remote sensing images, geology and geophysical data and so on. For instance, effective rescue efforts depend on the immediate availability of lifeline engineering data and information about the terrain, transportation, population distribution and medical facilities of the impacted areas. The central role of a disaster emergency data infrastructure is to assemble past and real-time data rapidly and make them available effectively to disaster response and relief command units, as well as provide support and basis for the subsequent decision-making and planning of meteorology, hydrology and geology departments. *The DEDI expands the current SDIs by incorporating a variety of data source beyond remote sensing and airborne data and, in the future, supporting the interoperability of data, models and tools from multiple disciplines.*

DEDI adopts the grid-based technical framework of distributed spatial data infrastructure, taking the independently managed disaster data collections and federating them into unified grid-based data nodes through grid services. This infrastructure will connect the international rescue data into a unified emergency data collection system for accurate search and detection of constantly updated international rescue data. Local and domestic disaster management agencies, disaster relief organizations, scientists, and other stakeholders of the disaster emergency data would be able to access the data without carrying out special processing or adjustments in order to utilize the data for their immediate needs. This approach will not affect how the current FTP solutions work today, making the transition to next generation disaster data infrastructure feasible.

DEDI has three main functions: *data verification, data quality control and data storage*, as services to data contributors and users. Sub-systems (such as a user portal, service interfaces) will be needed to manage data import and export.

USGS's Hazards Data Distribution System (HDDS) is a prime example of a disaster emergency data infrastructure that receives and distributes earth observation data from a variety of sources rapidly in response to disaster events, including its role in helping the relief work after the 2010 Haiti earthquake. Within six weeks after the earthquake, over 600,000 files representing 54 terabytes of data were provided to the response community (Duda, Kenneth A.etal, 2011). Another example is the European Space Agency (ESA) data infrastructure The Supersites. application of the emergency data infrastructure, the typical case is the 2003 Algerian Earthquake. Within one day after a major earthquake in the region east of Algiers, Algeria, the European Space Agency quickly determined the size of the disaster-hit zones, estimated the number of collapsed building and the people affected, with the help of pre-quake and post-quake SPOT satellite images (Duda, Kenneth A.etal, 2011), assisting efforts of the rescue teams from multiple countries with accurate and useful scientific data.

## 7.3 Historical Archive Data Infrastructure

Historical Archive Data Infrastructure (HADI) for disaster management is a archival system of multiple data banks containing historical disaster events. It collects, curates, organizes and stores the data relevant to selected past disaster events in a uniform schema to facilitate easy, speedy and accurate retrieval of data for research.

The archived event data includes various structured data relevant to specific disaster events, with observation data (aerial remote sensing data and ground survey) accounting for the majority. Data providers are primarily geology departments, meteorology agencies and other competent departments in charge of disaster-related data. Data users include government agencies, scientific research community, as well as commercial entities (e.g., insurance companies) in supporting specific areas of research with their data collections. ESA's Supersite serves as a successful example of a historical archive data infrastructure. Its Geoharzard Supersites covers multiple geological disaster incidents of many countries. Its support for various data types are being continuously updated.

HADI employs multi-source data collection technology, and collects historical disaster data scattered in different places for real-time data ingest, management and sharing through submission of public resources. In addition to technical implementations, data ingest and curation processes must be developed to make the archived data understandable and easy to use. For instance, HADI needs to adopt the appropriate metadata standards, as well as provide easy to tools to obtain and extract metadata from the contributed datasets (e.g., if it is difficult or cumbersome to enter metadata into the system, data contributors will not provide metadata).

## 7.4 Disaster Loss Database Infrastructure

Disaster Loss Data Infrastructure (DLDI) gathers disaster loss information from various data resources and provides a uniform framework for storing, managing and accessing the data. DLDI can be viewed as a large database that records data of losses in historical disaster events. This type of data provides the basis for assessing disaster risks and vulnerabilities, tracking and evaluating disaster damages and linking to the causes. Quantitative approaches and use of disaster loss data may lead to measures and decisions to improve emergency preparedness and reduce future damages.

Different stakeholders are involved in various phases ranging from data collection to data distribution and then to shared use of disaster loss data. Disaster loss data can include a wide variety of information. General description of disaster damages may come from media reports, satellite images and case studies. Scientific data related to disaster events such as the magnitude and intensity of an earthquake, duration of precipitation and rainfall, are mainly acquired by relevant research institutes and administration departments of climate and geology. Data reflecting the disaster's impact

on people and their lives (e.g., casualties, missing and affected population) is mainly acquired through response agencies and population management departments. Economic data on losses (e.g., property losses and insurance pay-outs) is typically obtained from the government, statistics bureau and the insurance companies. other disaster related information including the disaster affected area and the affected population can be acquired from research institutes.

Today, disaster loss data systems exist at various scales based on their coverage of geographic scale, ranging from global, regional to national scale, as well as event-based and department-based disaster loss database. Table 2. lists several examples of such systems internationally:

**Table2. Examples of global disaster data systems**

| Kind of Data | Examples: Data Collectors | Comments |
|---|---|---|
| Global multi peril | EmDat, Munich Re, Swiss Re | |
| Regional multi peril | La Red,<br>DesInventar<br>EEA European Environmental Agency | |
| National multi peril | UNDP (country databases after TS 2004), Sheldus | |
| Event based | Dartmouth Flood Observatory<br>CEDIM Centre for Disaster Management and Risk Reduction Technology | Flood<br>Earthquakes, Landslides |
| Sector based | Ascend<br>USDA (US Dept. of Agriculture) | Aviation<br>Agriculture |

*© 2012 Münchener Rückversicherungs-Gesellschaft, Geo Risks Research, NatCatSERVICE– October 2012*

The user groups of disaster loss data include researchers and scholars, decision-makers and financial institutions. The data is intended for scientific research, damage evaluation and support for decision-making.

The building of disaster loss data infrastructure is premised on the collection and organization of related data. Disaster loss data include data on the direct, tangible monetary impact of the disaster based on damages to buildings, infrastructure, agriculture, and the natural environment, as well as monetary losses paid out by the insurance company. Disaster loss data also include quantifiable counts of direct human losses – the number of people who died, were physically injured, or missing due to the disaster event. Also included are the number of people needing immediate assistance for sheltering (homeless) and those who were temporarily forced to leave their homes (evacuated) or who were forced to leave for longer time periods (displaced).

The development of a disaster data loss infrastructure covers the whole process from data collection to database design, to data services to end users. One of the most important tasks in data collection is semantic information mining (artificial and automatic). This step gleans specific information about losses from various data resources and store the information in the database. The collection of disaster loss data is an arduous process yet to be standardized. See “Cutter et al. 2008, National Research Council 1999” for how to organize the resources and offer corresponding policy guidance.

SHELDUS is a representative disaster loss database infrastructure. Currently, the online service edition has been upgraded to 13.1 and will be launched into official operation in August 2014, with a more inclusive database on disaster loss. Its disaster data covers the years from 1960 to 2013 of

almost the whole of the American continent. Table3. lists some platforms of disaster loss data covering different geographic scales.

**Table3. Examples of the Global Disaster Loss Data Platforms**

| Organization | Examples | Comments |
|---|---|---|
| UNDP GRIP | National Loss Data Observatory | Aim:50-100 countries |
| UN ISDR | National databases / Desinventar | |
| GLIDE | Unique identifier | |
| Relief Web, OCHA | Information from various sources | Focus on human impact |
| ICSU / IRDR | Working Group on Disaster Loss Data and Impact Assessment | |

# 8 Conclusions and Recommendations

Disasters post a great threat and challenge to all human societies. It is the inevitable phenomenon resulting from interactions between natural evolution process of the earth and human activities, which make disaster recognition and disaster mitigation a complex task. The diversity in disasters requires a multi-disciplinary approach, integrating scientific research and application of the findings with pre-disaster prediction, decision making and assessment based on complete, scientific and reliable data.

As our societies entered the information age, the collection, storage and services of disaster related data have also entered the digital phase, enabling large scale analysis and processing of such data by the science community. Open data strategies have been adopted in more and more disaster related data service infrastructure, especially with the trend of global open data. In all phases of disaster management, including pre-disaster forecast, emergency rescue and post-disaster reconstruction, the need for interconnected multi-disciplinary open data for collaborative study, process and analysis around common issues is apparent, in order to recognize and discover the rules and discipline of disaster completely, scientifically and in time.

Progress has been made in the technology and infrastructure domain. Information technology, such as high-performance networking (e.g., Internet), high performance computing, service and cloud computing, and big data methods, provides the technical foundation for connecting open data to support disaster research. Organizations, especially in the earth observation community, have launched efforts to connect global disaster related data resources and achieved great successes in studies and real-life cases. A new generation disaster data infrastructure based on the interconnected open data begins to form. In the science community, interconnected open data for disaster is beginning to influence how disaster data is shared and will need to extend coverage of data and provide better ways of utilizing data across domains where innovation and integration are very much needed.

One way to heed the call of Sendai Framework for greater use of science is to build strong links between such National Platforms and leading networks of scientists, researchers and other academics. The Integrated Research on Disaster Risk Programme (IRDR), funded by the Chinese Academy of Science and co-sponsored by the International Council for Science(ICSU) , the United Nations Office for Disaster Risk Reduction (UNISDR), and the International Social Science Council (ISSC), aims to serve as that link to bring more science and data-driven approaches to disaster risk management.

What can the science community and governmental organizations do to advance the state of disaster relevant data for a better understanding of disasters and more effective ways of mitigating and reducing impact of disasters on lives and properties?

We hereby make the following recommendations:

1. Academia, disaster management agencies and international organizations should strengthen global collaboration on disaster data by coordinating the utilization of disaster data from multiple data repositories, and promoting the interconnection and use of multi-domain data as a high-priority scientific activity in disaster research.
2. Organizations at the United Nations (UNISDR), national governments and relevant agencies should mobilize resources and accelerate the effort of establishing common definitions and data standards for disaster data to ensure effective implementation of data interconnectivity at both technical and policy levels. Scientists and research institutions are urged to actively participate in such effort to ensure that these standards meet the needs of disaster research.
3. Disaster data acquisition, preservation and service organizations should improve the accessibility and usability of disaster related data and realize the CODATA principle of data sharing. Suggestions include:
    a. Help connect data users (researchers, disaster management agencies, policy makers, corporate managers, citizens) to data providers through IRDR/CODATA workshops and other professional meetings
    b. Provide best practice guidelines on implementing CODATA open data strategies and showcase example implementations
    c. Moving data repositories from research to service, reducing barriers for data access and use.
    d. Expand open data beyond the domain of earth observations to include other types of data, especially the economic, population, public health, infrastructure, and social media data, etc.
4. Consult with relevant agencies and communities, and establish disaster data copyright protection and acceptable use policy to ensure the legality and appropriate use of data during disaster mitigation.
5. Study, design and ultimately create the next-generation disaster data infrastructure to enable the discovery of and easy access to highly usable, distributed, multi-disciplinary datasets for disaster mitigation stakeholders and applications on the global scale. Developed countries should take the lead in enabling global data service capabilities.
6. The international community should focus on the urgent needs of developing countries in disaster management and help them by establishing appropriate mechanisms of global and regional cooperation and basic data infrastructure to utilize the open data resources from the international community over Internet.
7. Innovative ideas are needed to encourage the private sector to join this effort. The private sector and the public also have incentives to support efforts to open and link data for disaster research and reduction.
8. Establish one or more pilot projects on applications of cross-discipline approach and use of data for studying disaster risk reduction. This could involve academic experiments, datasets, infrastructure, testbeds, and institutions who are at the forefront of supplying and using data to assist with disaster management. The scale could be institutional, national, and regional.

## Abbreviations

ADPC Asian Disaster Reduction Centre (ADRC)

ALI:Advanced Land Imager (ALI)
ALOS:Advanced Land Observation Satellite
ASI：Agenzia Spaziale Italiana
ASTER:Advanced Spaceborne Thermal Emmission and Reflection radiometer (ASTER),
AVHRR:Advanced Very High Resolution Radiometer (AVHRR), Hyperion,
BJ-1:Beijing-1 satellite
CBERS:China-Brazil Earth Resources Satellite
CEODE:Centre for Earth Observation and Digital Earth
CEOS:Committee on Earth Observation Satellites
CGMS:Coordination Group for Meteorological Satellites (CGMS)
CNES:Centre National d'Etudes Spatiales (CNES)
CODATA: Committee on Data for Science and Technology (CODATA)
CRED :Centre for Research on the Epidemiology of Disasters (CRED)
CRESDA : China Centre for Resources Satellite Data and Application (CRESDA)
DDI:Disaster Data Infrastructure (DDI)
DEDR:Disaster Emergency Data Reservoir (DEDR)
DLR：Deutsches Zentrum für Luft- und Raumfahrt
DMC:Disaster Monitoring Constellation (DMC)
EDC:Earth Resources Observation Systems (EROS) Data Center (EDC)
EPOS :European Plate Observatory System
ESA: European Space Agency
FEMA: Federal Emergency Management Agency (FEMA)
FY:Fengyun satellite
GCOS :Global Climate Observing System(GCOS)
GEO: Group on Earth Observations
Geonet :a geological hazards monitoring service in New Zealand run by GNS Science
GEOSS : Global Earth Observation System of Systems (GEOSS)
GPS:Global Positioning System
GRIP :Global Risk Identification Programme (GRIP)
GSNL: Geohazards Supersites and Natural Laboratories (GSNL)
HDDG :Historical Disaster Data Grid (HDDG)
HDDS:Hazards Data Distribution System (HDDS)
ICSU: The International Council for Science (ICSU)
INGV:Istituto Nazionale di Geofisica e Vulcanologia
InSAR :Interferometric Synthetic Aperture Radar (InSAR)
IOC:Intergovernmental Oceanographic Commission (IOC)
IRDR: Integrated Research on Disaster Risk
IRG :Integrated Risk Governance Project (IRG)
IRIS: Interface Region Imaging Spectrograph (IRIS)
JAXA: JapanAerospaceExplorationAgency
LA RED :Red de EstudiosSociales en Prevención de Desastres en América Latina (Network of Social Studies in the prevention of Disasters in Latin America - LA RED)
LODGD: linked open data for global disaster risk research.
MODIS:Moderate Resolution Imaging Spectroradiometer (MODIS),

NASA: National Aeronautics and Space Administration
NCDC:National Climatic Data Centre
NMA: National Meteorological Agency
NOAA: The National Oceanic and Atmospheric Administration
NSF: National Science Foundation
NWS: The U.S. National Weather Service (NWS)
OECD :the Organisation for Economic Co-operation and Development (OECD)
PRC: Peoples republic of China
SCI:Science Citation Index
SCU: University of South Carolina
SHELDUS: the Spatial Hazard Events and Losses Database (www.sheldus.org)
SJ-9A:Shijian-9A satellite
SPARC :Scholarly Publishing and Academic Resources Coalition (SPARC)
SPARC: Scholarly Publishing and Academic Resources Coalition (SPARC)ORPHEUS
SPOT: Systeme Probatoired'Observation dela Tarre
TWAS: The Third World Academy of Sciences
UAV: An unmanned aerial vehicle (UAV)
UN ECA:United Nations Economic Commission for Africa
UN-ESCAP :United Nations Economic and Social Commission for Asia and the Pacific (UN-ESCAP)
UN-ESCAP:United Nations Economic and Social Commission for Asia and the Pacific (UN-ESCAP)
UNAVCO: aNon-Profit University-Governed Consortium, Facilitates Geoscience Research And Education Using Geodesy.
UNDP :United Nations Development Programme (UNDP)
UNEP:United Nations Environment Programme (UNEP)
UNESCAP: United Nations Economic and Social Commission for Asia and the Pacific (ESCAP)
UNSPIDER: United Nations Platform for Space-based Information for Disaster Management and Emergency Response
USGS: UnitedStatesGeologicalSurvey
WDC: World Data Centre (WDC)
WDS: World Data System
WGDD :Working Group on Disaster Data (WGDD)
WMO :World Meteorological Organization (WMO)

## References


[1] Li Juan, Liu Dehong and Jiang Hong, Research on International Current Status on Sharing Science Data [J], Library Development, 2009.

[2] Committee on Scientific Accomplishments of Earth Observations from Space, National Research Council (2008). Earth Observations from Space: The First 50 Years of Scientific Achievements. The National Academies Press. p. 6. ISBN 0-309-11095-5. Retrieved 2010-11-24.

[3] On the Full and Open Exchange of Science data (1995) National Research Council, Washington, DC

[4] Open Data Centre Alliance. Defining a New Class of Data Centre and Cloud Infrastructure Solutions.

[5] http://www.opendatacentrealliance.org/

[6] SPARC. Open Data ,http://www.arl.org/sparc/opendata/.

[7] W3C. Linking Open Data［EB/OL］ 2011-03-15.

[8] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData#FAQ.

[9] http://en.wikipedia.org/wiki/Open_Data.

[10] Noor Huijboom, Tijs Van den Broek.Open data: an international comparison of strategies.European Journal of e-Practice.

[11] Bouwer L M, Crompton R P, Faust E, et al. Confronting disaster losses[J]. Science-New York then Washington-, 2007, 318(5851): 753.

[12] Kunreuther H. Mitigating disaster losses through insurance[J]. Journal of risk and Uncertainty, 1996, 12(2-3): 171-187.

[13] Dilley M. Natural disaster hotspots: a global risk analysis[M]. World Bank Publications, 2005.

[14] Gall M, Borden K A, Cutter S L. When do losses count? Six fallacies of natural hazards loss data[J]. Bulletin of the American Meteorological Society, 2009, 90(6): 799-809.

[15] Li Guoqing, Liu Dingsheng, Yu Wenyang, etc. Grid-based International Data Sharing Environment about Wenchuan Earthquake [J]. Scientific Research and Information Technology and Its Application, 2008, 1:66-75.

[16] Liu Ruifeng, Cai Jin'an, Peng Keyin, etc. Seismological Science Data Sharing Project [J]. Earthquake, 2007, 27 (2):9-16.

[17] China Aerospace Science and Technology Corporation, Improve Aerospace Infrastructure Construction and Enhance Capability of Disaster Prevention and Reduction [J]. 2008(17):http://www.qstheory.cn/zxdk/2008/200817/200906/t20090609_1428.htm

[18] Rochon, G. L., Quansah, J. E., Mohamed, M. A. et al Applicability of Near-Real-Time Satellite Data Acquisition and Analysis & Distribution of Geoinformation in Support of African Development UN ECA (2005)

[19] http://www.scidev.net/global/feature/zh-135036.html

[20] Atkins, Daniel. "Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure." (2003).

[21] Hey, Tony, and Anne E. Trefethen. "Cyberinfrastructure for e-Science." Science 308.5723 (2005): 817-821.

[22] Yang, Chaowei, et al. "Geospatial cyberinfrastructure: past, present and future." Computers, Environment and Urban Systems 34.4 (2010): 264-277.

[23] Foster, Ian. "Accelerating and democratizing science through cloud-based services." IEEE Internet Comput. 15.ANL/MCS/JA-69753 (2011).

[24] Moore, Reagan W., and ArcotRajsekar. "Irods: Data sharing technology integrating communities of practice." Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International. IEEE, 2010.

[25] Egeland, Ricky, Tony Wildish, and Chih-Hao Huang. "PhEDEx data service." Journal of Physics: Conference Series. Vol. 219. No. 6. IOP Publishing, 2010.

[26] Halevy, Alon, Peter Norvig, and Fernando Pereira. "The unreasonable effectiveness of data." Intelligent Systems, IEEE 24.2 (2009): 8-12.

[27] http://www.sheldus.org/

[28] http://webra.cas.sc.edu/hvriapps/sheldus_setup/sheldus_location.aspx

[29] Cutter S L. Social Science Perspectives on Hazards and Vulnerability Science [M]. Springer Netherlands, 2010.

[30] Nathan Paul, Multiple Natural Disaster Database and Mapping System Northwest Missouri State University, December 2009

[31] http://www.r-bloggers.com/towards-the-r-package-sheldus-part-1-natural-disaster-losses-in-the-us-in-2012/

[32] Lamb R M, Jones B K. United States Geological Survey (USGS) Natural Hazards Response[M]. 2012.

[33] Jones B K, Risty R R. The National Map Hazards Data Distribution System[J].

[34] Jones B K, Bewley R D. The Hazards Data Distribution System is Updated[M]. US Department of the Interior, US Geological Survey, 2010.

[35] Bewley R D. USGS Emergency Response Resources[J].

[36] Jones B K, Lamb R. USGS Emergency Response and the Hazards Data Distribution System (HDDS)[C]//AGU Fall Meeting Abstracts. 2013, 1: 1601.

[37] http://hddsexplorer.usgs.gov/

[38] https://twitter.com/USGS_HDDS

[39] The Geohazard Supersites Partnership White Paper and Implementation Plan,11 October 2011.

[40] Selection Process for GEO Geohazard Supersites, A Proposal of the Committee on Earth Observing Satellites. September 18, 2012

[41] Guidelines for Permanent Supersite Proposals Regarding In-situ Data, The Scientific Advisory Committee (*), December 2012.

[42] Geohazard Supersites and Natural Laboratories (GSNL) Initiative “Supersites Definitions” 2012-2015 GEO Work Plan – Disasters (DI-01) Task, September 6 2012.

[43] Lengert W, Popp H J, Gleyzes J P. GEO Supersites Data Exploitation Platform[C]//EGU General Assembly Conference Abstracts. 2012, 14: 11933.

[44] Rowan L, Baker S, Wier S, et al. Integrated Data Search and Access to Geophysical Data for Geohazards Supersites and Natural Laboratories[C] AGU Fall Meeting Abstracts. 2013, 1: 1621.

[45] Amelung F, Lengert W, Puglisi G. Supersites: An Initiative Towards Open Access for InSAR Data[C]//EGU General Assembly Conference Abstracts. 2010, 12: 11682.

[46] http://supersites.earthobservations.org/

[47] http://eo-virtual-archive4.esa.int/

[48] Cohen, Wesley and John Walsh (2008), "Real Impediments to Academic Research," inAdam B. Jaffe, Josh Lerner, and Scott Stern, eds.,*InnovationPolicy and the Economy,Volume 8*, Chicago: University of Chicago Press, 1-30.

[49] Haeussler, Carolin, Lin Jiang, Jerry Thursby, and Marie Thursby (2014), "Specific and general information sharing among competing academicresearchers" *Research Policy*, **43**, 465-475.

[50] Organisation for Economic Co-operation and Development, *Inquiries into Intellectual Property's Economic Impact*, OECD Publishing, Paris (2015a).

[51] Organisation for Economic Co-operation and Development, *Making Open Science a Reality*, OECD Publishing, Paris (2015b).

[52] Thursby, Marie, Jerry G. Thursby, Carolin Haeussler, and Lin Jiang (2009), "Do academic scientists share information with theircolleagues? Not necessarily," VOX – CEPR's Policy Portal, November 25.

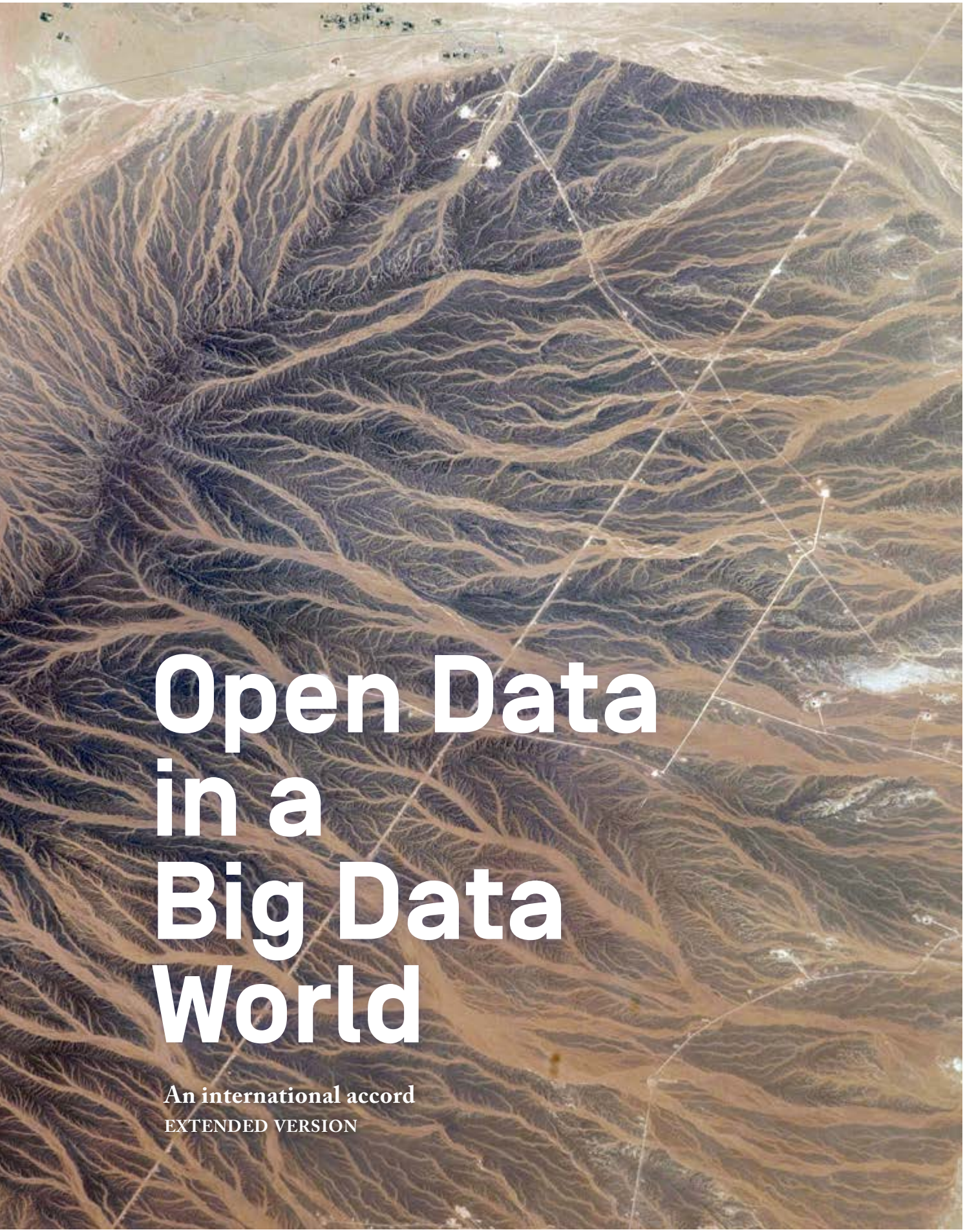
[53] United Nations Office for Disaster Risk Reduction, *Sendai Framework for Disaster Risk Reduction 2015 – 2030*, Geneva (2015).

## Revision History

| Date | Editors | Sections changed | Description |
|---|---|---|---|
| May 20 | | Version 16.0 | Release to external review |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

## Appendix A. Science International (2015): Open Data in a Big Data World

## Appendix B. The Geohazards Exploitation Platform



# Open Data in a Big Data World

An international accord

EXTENDED VERSION

ICSU International Council for Science
iap SCIENCE RESEARCH HEALTH the interacademy partnership
ISSC international social science council
twas

## Preface

Four major organisations representing global science, the International Council for Science (ICSU), the InterAcademy Partnership (IAP), The World Academy of Sciences (TWAS) and the International Social Science Council (ISSC), are collaborating in a series of action-oriented annual meetings, dubbed "Science International". These meetings are designed to articulate the views of the global scientific community on international matters of policy for science and to promote appropriate actions.

The following accord is the product of the first Science International meeting. The accord identifies the opportunities and challenges of the data revolution as one of today's predominant issues of global science policy. It sets out principles that are consistent with ones being carried out in practice in some national research systems and in some disciplinary fields. It adds the distinctive voice of the scientific community to those of governments and inter-governmental bodies that have made the case for open data as a fundamental pre-requisite in maintaining the rigour of scientific inquiry and maximising public benefit from the data revolution. It builds on ICSU's 2014 statement on open access by endorsing the need for an international framework of open data principles.

In the months ahead, Science International partners will promote discussion and adoption of these principles by their respective members and by other representative bodies of science at national and international levels. We will ask that these organizations review the accord and endorse it, and thereby provide further support in global policy venues for these constructive and vitally important principles.

An abbreviated version of this accord summarises the issues in section A of this document and presents the open data principles that it advocates.

# A. Opportunities in the Big Data World

## A world-historical event

**1.** The digital revolution of recent decades is a world-historical event as profound and more pervasive than the introduction of the printing press. It has created an unprecedented explosion in the capacity to acquire, store, manipulate and instantaneously transmit vast and complex data volumes[1]. The rate of change is formidable. In 2003 scientists declared the mapping of the human genome complete. It took over 10 years and cost $1billion – today it takes mere days and a fraction of the cost ($1000)[2]. Although this revolution has not yet run its course, it has already produced fundamental changes in economic and social behaviour and has profound implications for science[3], permitting patterns in phenomena to be identified that have hitherto lain beyond our horizon and to demonstrate hitherto unsuspected relationships. Researchers were amongst the first users of digital networks such that many areas of research across the humanities, natural and social sciences are being transformed, or have the potential to be transformed, by access to and analysis of such data.

**2.** The worldwide increase in digital connectivity, the global scale of highly personalized communications services, the use of the World Wide Web as a platform for numerous human transactions, the "internet of things" that permits any device with a power source to collect data from its environment together with advances in data analytics have coalesced to create a powerful platform for change. In this networked world, people, objects and connections are producing data at unprecedented rates, both actively and passively. This not only creates large data volumes, but also distinctive data streams that have been termed "big data", characterised by the four Vs[4]:

- the **volume** that systems must ingest, process and disseminate;
- the **variety** and complexity of datasets, originating from both individuals and institutions at multiple points in the data value chain;
- the **velocity** of data streaming in and out of systems in real time;
- the **veracity** of data (referring to the uncertainty due to bias, noise or abnormality in data), which is often included. This is a desirable characteristic, not an intrinsic feature of Big Data. The veracity and the peer review of results based on big data, however, pose severe problems for effective scrutiny, with a clear need to establish a "reproducibility standard."

**3.** A second pillar of the data revolution is formed by "linked data." Separate datasets that relate to a particular phenomenon and that are logically connected can be semantically linked in ways that permit a computer to identify deeper relationships between them. Semantic search links similar ideas together, permitting the World Wide Web to evolve from a web of documents into a Semantic Web in which meaning can be more readily deduced from linked data, connecting related data that were not necessarily designed for mutual integration. Such processes offer profound ways of understanding the structure and dynamics of systems where very diverse elements are coupled together to produce complex behaviour. They have the potential to yield an enormous dividend of understanding by breaking down the barriers that tend to separate disciplinary silos, although only if the data is openly available and free to be linked.

**4.** The great achievements of science in recent centuries lie primarily in understanding relatively simple, uncoupled or weakly coupled systems. Access to increasing computational power has permitted researchers to simulate the dynamic behaviour of highly coupled complex systems. But the advent and analysis of big and linked data now add to this the complementary capacity to characterise and describe complexity in great detail. Coupling these two approaches to the analysis of complexity has the potential to usher in a new era of scientific understanding of the complexity that underlies many of the major issues of current human concern. "Global challenges" such as infectious disease, energy depletion, migration, inequality, environmental change, sustainability and the operation of the global economy are highly coupled systems, inherently complex, and beyond the reach of the reductionist approaches and the individual efforts

1 We use the term ***data*** to refer to "representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship". C.L. Borgman, 2015. *Big Data, Little Data, No Data: Scholarship in the Networked World.* The MIT Press, p. 28.

2 Illumina announces landmark $1,000 human genome sequencing. *Science* 15 January 2014

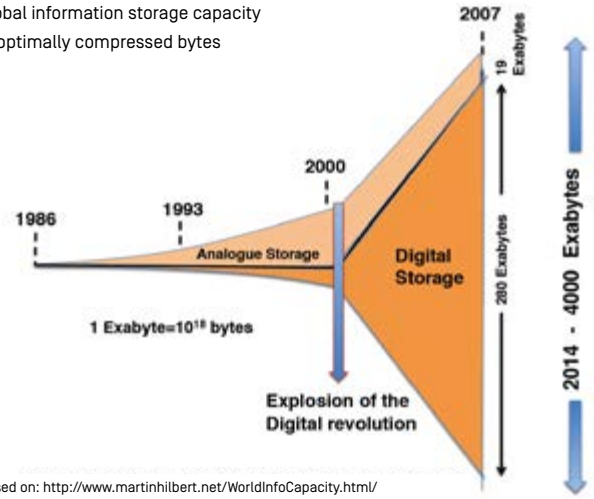
3 The word ***science*** is used to mean the systematic organisation of knowledge that can be rationally explained and reliably applied. It is used, as in most languages other than English, to include all domains, including humanities and social sciences as well as the STEM (science, technology, engineering, medicine) disciplines.

4 www.ibmbigdatahub.com/infographic/four-vs-big-data

BOX 1

### The beginning of the digital revolution

Global information storage capacity
In optimally compressed bytes




Based on: http://www.martinhilbert.net/WorldInfoCapacity.html/

BOX 2

### Linked Data and the Semantic Web

Linked Data use the techniques and concepts of the World Wide Web to describe the real world. They use web identifiers (Uniform Resource Identifier or URIs, often in the form of an http location) to identify facts, concepts, people, places and phenomena as well as documents that have common attributes. This allows connections to be discovered between different datasets, thereby increasing the value of each through the Network Effect, permitting a researcher to discover data important to their work. Programmes such as Resource Discovery for Extreme Scale Collaboration (http://rdesc.org) use these approaches to search for and discover data resources relevant to a particular scientific purpose. The approach is being increasingly applied in environmental fields. Operational examples relevant to business include OpenPHACTS, which uses the technology to provide easy access to more than 14 million facts about chemical and pharmacological data; the European Environment Agency's provision of reference datasets for species; and the Slovenian Supervizor portal which matches public spending to contracts to businesses, providing a powerful tool against corruption.

Linked Data is a subset of the wider Semantic Web, in which queries do not retrieve documents as in the standard web, but semantic responses that harvest information from datasets that are connected by logical links. This approach is being much exploited in genomics, one example being through a resource description framework (RDF) platform implemented through the European Molecular Biology Laboratory Elixir programme (see Box 6).

that nonetheless remain powerful tools in the armoury of science. The potential of big data in such cases is to permit analysis of complex system whilst still producing general explanations.

**5.** A further consequence of the increasing capacity to acquire data at relatively low cost, when coupled with great processing power, is to permit machines that sense data from their immediate environment to learn complex, adaptive behaviours by trial and error, with the disruptive potential to undertake what have hitherto been regarded as highly skilled, and necessarily human, tasks.

## B. Exploiting the Opportunities: the Open Data Imperative

### Maintaining "self-correction"

**6.** Openness and transparency have formed the bedrock on which the progress of science in the modern era has been based. They have permitted the logic connecting evidence (the data) and the claims derived from it to be scrutinised, and the reproducibility of observations or experiments to be tested, thereby supporting or invalidating those claims. This principle of "self-correction" has steered science away from the perpetuation of error. However, the current storm of data challenges this vital principle through the sheer complexity of making data available in a form that is readily subject to rigorous scrutiny. Ensuring that data are open, whether or not they are big data, is a vital priority if the integrity and credibility of science and its utility as a reliable means of acquiring knowledge are to be maintained.

**7.** It is therefore essential that data that provide the evidence for published claims, the related metadata that permit their re-analysis and the codes used in essential computer manipulation of datasets, not matter how complex, are made concurrently open to scrutiny if the vital process of self-correction is to be maintained. The onus not only lies on researchers but also on scientific publishers, the researchers who make up the editorial boards of scientific journals and those managing the diverse publication venues in the developing area of open access publishing, to ensure that the data (including the meta-data) on which a published scientific claim are based are concurrently available for scrutiny. To do otherwise should come to be regarded as scientific malpractice.

### The definition of open data

**8.** Simply making data accessible is not enough. Data must be **"intelligently open"**[5], meaning that they can be thoroughly scrutinised and appropriately re-used. The following criteria should be satisfied for open data, that it should be:

- **discoverable** – a web search can readily reveal their existence;
- **accessible** – the data can be electronically imported into or accessed by a computer;
- **intelligible** – there must be enough background information to make clear the relevance of the data to the specific issue under investigation;
- **assessable** – users must be able to assess issues such as the competence of the data producers or the extent to which they may have a pecuniary interest in a particular outcome;
- **usable** – there must be adequate metadata (the data about data that makes the data usable), and where computation has been used to create derived data, the relevant code, sometimes together with the characteristics of the computer, needs to be accessible.

Data should be of high quality wherever possible, reliable, authentic, and of scientific relevance. For longitudinal datasets, the metadata must be sufficient for users to be able to make a comparative analysis between timelines, and the sources must be valid and verifiable. It is important to be aware that the quality of some scientifically important datasets, such as those derived from unique experiments, may not be high in conventional terms, and may require very careful treatment and analysis.

### Non-Replicability

**9.** The replication of observations and experiments has a central role in science. It is the justification for the statement made by Galileo in Brecht's play[6] that "the aim of science is not to open the door to infinite wisdom, but to set a limit to infinite error." Recent attempts to replicate systematically the results of series of highly regarded published papers in, for example, pre-clinical oncology (53 papers)[7], social psychology (100 papers)[8] and economics (67 papers)[9], were successful in only 11 %, 39 % and 33 % of cases respectively. The reasons adduced for these failures included falsification of data, invalid statistical reasoning and absent or incompleteness of the data or metadata. Such failures were highlighted in *The Economist*[10] magazine under the headline: "Scientists like to think of science as self-correcting. To an alarming degree it is not." These failures will threaten the credibility of the scientific enterprise unless corrective action is taken. If data, meta-data and the code used in any manipulations are not available for scrutiny, published work, whether right or wrong, cannot be subject to an adequate test of replication.

**10.** An implication of the above results is that pre-publication peer review has failed in these cases in its primary purpose of checking whether the research has been performed to reasonable standards, and whether the findings and conclusions drawn from them are valid. Given the depth of analysis required to establish replicability, and the increasing pressure on reviewers because of the dramatic rise in the rate of publication[11], it is unsurprising that peer review fails in this regard. Under these circumstances, it is crucial that data and metadata are concurrently published in an intelligently open form so that it is also accessible to "post-publication" peer review, whereby the world decides the importance and place of a piece of research[12].

### Open data and "self correction"

**11.** The reputational and other rewards for scientific discovery can be considerable, with an inevitable temptation for misconduct involving the invention of data or intentional bias in their selection. In general we would expect open data to deter fraud, on the principle that "sunlight is the best disinfectant". In contrast, there are cases where the integration of datasets derived from different open sources could enable fraud by effectively hiding fraudulent components because of the difficulty of disentangling datasets. Without a standard of openness that permits, even in these cases, others to subject the related scientific claim to the test of reproducibility,

5 Science as an Open Enterprise. 2012. The Royal Society Policy Centre Report, 02/12. https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/

6 Bertolt Brecht, 1945. *The Life of Galileo.*

7 Begley, C.G. and Ellis, L.M. 2012. Nature, 483, p. 531 – 533.

8 Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. Doi: 10.1126/science.aac4716.

9 Chang, A. and Li, P. 2015. Finance and Economics Discussion Series 2015-083. Washington: Board of Governors of the Federal Reserve System

10 The Economist. 2013, October 19 – 25. pp. 21 – 23.

11 The term publishing and publisher simply refer to the act of making written or spoken work publically and permanently available. It is not restricted to conventional printed publication.

12 Smith, R. 2010. Classical peer review: an empty gun. *Breast Cancer Research* 2010, 12 (Suppl 4): S13 http://breast-cancer-research.com/supplements/12/S4/S13.

such claims may prove to be an irreducible barrier to scientific progress. The integrity of data is often of greater significance than the claim based on them. To quote Charles Darwin[13]: " false facts are highly injurious to the progress of science, for they often long endure; but false views, if supported by some evidence, do little harm, as everyone takes a salutary pleasure in proving their falseness"; leading to an outcome described by Arthur Koestle[14] as one in which "the progress of science is strewn, like an ancient desert trail, with the bleached skeletons of discarded theories that once seemed to possess eternal life".

## Valid reasoning

**12.** A major priority for data-intensive science must be greater analytical rigour and the establishment, discipline by discipline, of acceptable standards of replicability. Regression-based, classical statistics have long been the basic tools for establishing relationships in data. Many of the complex relationships that we now seek to capture through big or linked data lie far beyond the analytical power of these methods, such that we now need to supplement them in adapting topological and related methods to data analysis to ensure that inferences drawn from big or linked data are valid. Data-intensive machine-analysis and machine-learning are becoming ubiquitous, creating the possibility of improved, evidence-informed decision making in many fields. The creative potential of big data, of linking data from diverse sources and of machine learning not only have implications for discovery, but also for the world of work and for what it means to be a researcher in the 21st century. The potential disconnect between machines that learn from data and human cognitive processes poses profound issues for how we understand machine-analysed phenomena and their accessibility to human reasoning.

## Openness: the default for publicly funded research

**13.** We regard it as axiomatic that knowledge and understanding have been and will continue to be essential to human judgements, innovation and social and personal wellbeing. The fundamental role of the publicly funded scientific enterprise is to add to the stock of knowledge and understanding, such that high priority should be given to processes that most efficiently and creatively advance knowledge. The productivity of open knowledge, of having ideas and data made open by their originators, is illustrated by a comment attributed to the playwright George Bernard Shaw: *"if you have an apple and I have an apple and we exchange these apples, then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas".* The technologies and processes of the digital revolution as described above provide a powerful medium through which such multiplication of productivity and creativity can be achieved through rapid interchange and development of ideas by the networked interaction of many minds.

**14.** If this social revolution in science is to be achieved, it is not only a matter of making data that underpin a scientific claim intelligently open, but also of having a default position of openness for publicly funded data in general. In some disciplinary communities data are released into the public domain immediately after they have been produced, such as in the case of genome sequencing data since the agreement of the 1996 Bermuda Principles and the 2003 Fort Lauderdale Principles[15]. The circumstance and timescale of release are important. It many disciplines it is reasonable to expect that data need only be released upon the termination of the grant that funded their collection. Even then it may be appropriate for grant holders to have the first bite of the publication cherry before data release. Although it is tempting to suggest a embargo period, perhaps of the order of a year, it would be better for individual disciplines to develop procedures that are sympathetic to disciplinary exigencies, but without involving excessive delay.

## Boundaries of openness

**15.** Although open data should be the default position for publicly funded research data, not all data can or should be made available to all people in all circumstances. There are legitimate exceptions to openness on matters of personal privacy, safety and security, whilst further ethical concerns ought to constrain the way that data systems operate and data are used, as discussed in the next section. Given the increasing incidence of joint public/private funding for research, and with the premise that commercial exploitation of publicly funded research data can be in the broader public interest, legitimate exceptions to the default position for openness are also possible in these cases. These categories – which are largely discipline-dependent – should not however be used as the basis for blanket exceptions. Exceptions to the default should be made on a case-by-case basis, with the onus on a proponent to demonstrate specific reasons for an exception.

## Ethical issues

**16.** Open data and data sharing have important ethical dimensions that relate to researchers' responsibilities to the public, to those who provide personal data, and to fellow researchers. Although we advocate a normative view that publicly funded researchers have an obligation to make data that they have collected openly available as a public good in the interests of science and society, we recognise that this creates further dilemmas that require attention:

- Datasets containing personal information have the potential to infringe the right to privacy of data subjects and require governance practices that protect personal privacy.
- A substantial body of work in computer science has demonstrated that conventional anonymisation procedures cannot guarantee the

13 Darwin, C. 1871. *The Descent of Man.* John Murray, London. 2 vols.

14 Koestler, A. 1967. *The ghost in the machine.* Macmillan, London, 384 pp.

15 Human Genome Project (2003). Available at: http://www.genome.gov/10506376 and http://www.wellcome.ac.uk/About-us/Publications/Reports/Biomedical-science/WTD003208.htm

BOX 3

### Managing Ethical Risk

The Administrative Data Research Centre for England (ADRC-E), has examined attitudes to data handling of administrative data and developed a model for managing ethical risks that attempts to address public concerns. An IPSOS-Mori poll on behalf of the UK Economic and Social Research Council (ESRC) found that the public held few objections over the use of administrative data for research purposes, subject to certain important caveats: that there should be strong governance, effective de-identification and de-linkage, and a clear public — not commercial — benefit in using the data. The ADRC network model recognises three critical levels of scrutiny: of researchers, of project aims, and of the role of the ADRC itself. ADRC will accredit researchers, then, when an accredited researcher requests access to data, a panel will evaluate whether or not the proposed project will deliver a clear public benefit, is drawn on data that is essential to their research and is not available elsewhere. Once a request is approved, the ADRC assembles the requested data sets and takes responsibility for linkage and de-identification. Importantly, in the language of data protection, the ADRC acts only as a 'data processor', not as a 'data controller'. This approach accommodates public concerns, and creates an acceptable synergy between researchers, the nature of data supplied and where the data are located. The model has proved financially sustainable following significant start-up funding.

In, Science Europe Social Sciences Committee (September 2015), 'Workshop Report: Ethical Protocols and Standards for Research in Social Sciences Today': D/2015/13.324/7

security of personal records[16], such that stronger, more secure practices may be required[17].

- Researchers have a moral obligation to honour relationships that they have developed with those who have entrusted them with personal information. Data sharing threatens these relationships because it entails a loss of control over future users and future usage of data. In the humanities and social sciences, data are often co-constructed by researchers and respondents, and also contain much sensitive information relating to both respondents and researchers.
- Open data can override the individual interests of the researchers who generate the data, such that novel ways of recognizing and rewarding their contribution must be developed (see Section D). Junior researchers, PhD students and/or technicians may be particularly vulnerable to lack of recognition, and with limited say in data reuse.
- In international projects, data sharing may become a form of scientific neo-colonialism, as researchers from well-funded research systems may stand to gain more than those from poorly funded systems. This could happen because of differences in infrastructure investment, or different levels of granularity.

## Open Global Participation

**17.** The ways in which big data, linked data and open data can be used for data-driven development and can be leveraged to positively impact the lives of the most vulnerable are becoming clearer[18]. There is great potential for data-driven development because of its detail, timeliness, ability to be utilized for multiple purposes at scale and in making large portions of low-income populations visible. Although many well-funded national science systems are adapting rapidly to seize the data challenge, the great promise of big data remains remote for many less affluent countries, and especially for the least developed countries (LDCs), where the costs of adaptation referred to in the next section pose particular problems.

**18.** LDCs typically have poorly resourced national systems. If they cannot participate in research based on big and open data, the gap could grow exponentially in coming years. They will be unable to collect, store and share data, unable to participate in the global research enterprise, unable to contribute as full partners to global efforts on climate change, health care, and resource protection, and unable fully to benefit from such efforts, where global solutions will only be achieved if there is global participation. Thus, both emerging and developed nations have a clear, direct interest in helping to fully mobilize LDC science potential and thereby to contribute to achievement of the UN Sustainable Development Goals. It is vital that processes that deliver local benefit are developed based on effective governance frameworks and the legal, cultural, technological and economic infrastructures necessary to balance competing interests.[19]

## Changing the dynamic

**19.** Creative and productive exploitation of this technologically-enabled revolution will also depend upon the creation of supporting "soft" and "hard" infrastructure and changes in the social dynamics of science, involving not only a willingness to share and to release data for re-use and re-purposing by others but the recognition of a responsibility to do so.

**20.** Although science is an international enterprise, it is largely done within national and disciplinary systems that are organised, funded and motivated by national and disciplinary norms and practices. Effective open data in a data-intensive age can only be realised if there is systemic action at disciplinary, national and international levels. At the national level there is need for government to recognise the value to be gained from open data, for national science agencies to adopt a coordinating role, for science policy makers to set incentives for openness from universities and research institutes, for these institutions to support open data processes by their researchers and for the learned societies that articulate the priorities and practices of their disciplines to advocate and facilitate open data processes as important priorities.

**21.** The rationale for a national open data policy lies in ensuring the rigour of national science based on its reproducibility and the accessibility of its results, in capturing the value of open data[20] for national benefit and as the basis for efficient collaboration in international science. New partnerships, infrastructures and resources are needed to ensure that researchers and research institutions work with government and private-sector big data companies and programmes to maximize data availability for research and for its effective exploitation both for public policy and direct economic benefit.

**22.** Soft and hard enabling infrastructures are required to support open data systems. Soft infrastructure comprises the principles that establish behavioural norms, incentives that encourage their widespread adoption

BOX 4

### Open research data in South America

The Latin American region is one with a strong tradition of cooperation in building regional information and publishing systems. Today, an estimated 80% of active journals are open access, complemented by repositories (regional subject repositories, and more recently institutional repositories) which are gaining momentum promoted by national open access legislation approved in Peru, Argentina, Mexico, and in discussion in Brazil and Venezuela. These require publicly-funded research results to be deposited in open access repositories, in some cases explicitly including research data.

The issue of open research data is starting to take off in the region, with activities to build awareness and consensus on good practices, sponsored by national research agencies (e.g. national systems for data–climate, biological, sea, genomics-coordinated by the Ministry of Science, Technology and Innovation of Argentina);

the initiative datoscientíficos.cl promoted by the National Commission of Scientific and Technological Research in Chile to seek opinions for a proposed policy for open research data; and a national meeting of open data organized by the Brazilian Institute of Information in Science and Technology. These national actions provide context and guidance for new institutional and national open research data initiatives within the region, which also look at other existing open research data programmes (e. g. UN Economic Commission for Latin America and open research data at the National Autonomous University of Mexico-UNAM].

In parallel, there is a movement in Latin America towards open government data, open knowledge and open data in general, as part of international movements and initiatives. Governments and civil society organize open data events and projects, open data schools, unconferences and data hackathons that build awareness about the need and opportunities to open government data, which also benefits research. To facilitate regional research cooperation and exchange of big research data, the National Research and Education Networks (NREN) are members of the Latin American Cooperation of Advanced Networks (RedCLARA) which provides advanced Internet networking facilities to countries and institutions of the region.

16 For example: Denning D (1980). *A fast procedure for finding a tracker in a statistical database.* ACM Transactions on Database Systems (TODS), 5, 1. *Differential Privacy*. International Colloquium on Automata, Languages and Programming (ICALP), 1–12; Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007).

17 For example: Thomas R & Walport M (2008). Data Sharing Review. Available at: http://www.justice.gov.uk/reviews/docs/data-sharing-review-report.pdf

18 http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html#

19 Linnet Taylor & Ralph Schroeder (2015) 'Is bigger better? The emergence of big data as a tool for international development policy, *GeoJournal* 80(4), pp. 503 – 518. 10.1007/s10708-014-9603-5

20 The economic value of open data has been estimated as $ 3 – 5 trillion per annum across seven commercial sectors. McKinsey Global Institute: Open Data, 2013.

and practices that ensure efficient operation of a national open data system that is also consistent with international standards. This part of the soft infrastructure is not financially costly, but depends upon effective management of the relationships summarised in the preceding paragraph and effective international links. The costly component is the need for time-intensive data management both by research institutions and researchers. By contrast, the physical or hard infrastructure required to sustain data storage, analysis, broadband transmission and long-term preservation is not separable from that required to support a strong national science base. Both soft and hard infrastructures are essential enabling elements for producing and using scientific data, though, as commented above, they pose especially difficult challenges for doing research in low- and middle-income countries.

**23.** Responsibilities also fall on international bodies, such as the International Council for Science's (ICSU) Committee on Data for Science and Technology (CODATA)[21] and World Data System (WDS)[22], and the Research Data Alliance (RDA)[23], to promote and support developments of the systems and procedures that will ensure international data access, interoperability and sustainability. Members of these bodies represent a wide range of countries, and both through them and through other national contacts, international norms should aim to be compatible with national procedures as far as possible. In establishing where change is required, it is important to distinguish between those habits that have arisen because they were well adapted to a passing technology but which may now be inimical to realisation of the benefits of a new one, and those habits that reflect essential, technology-independent priorities and values. In this regard, it is a priority to establish new ways of recognising, rewarding and therefore incentivising efforts in data management, preservation and curation. It involves questioning ingrained assumptions about the primacy of "high-impact" publications as a measure of scientific excellence, and finding ways to acknowledge communication of science, such as the development and dissemination of "open software", and participation in international programmes of data donation and curation.

**24.** Although the articulation by international representative bodies of the ethical and practical benefits of open data processes is important, it is the actions of practising scientists and scientific communities that will determine the adoption, extent and impact of these processes. These are fundamental issues for science, society and the economy and depend on the willingness of scientists to open up their data for sharing, re-use and re-purposing, even if there are personal, technical, organizational and political barriers to doing so. New solutions for making data open are required that demand collective efforts from all stakeholders involved in the production of knowledge, including individual researchers, the institutions in which they work, and the myriad organizations which influence their work. It is of course recognised that the gap between aspiration and practical implementation is a large one, both in terms of the willingness of individuals and institutions to change mindsets, and the capacity to adapt behaviour because of the availability of tools, management systems and hard infrastructure.

**25.** Major bottom-up changes are however happening at the level of disciplinary and multi-disciplinary communities. Strong processes of open data sharing have developed in areas such as linguistics[24], bioinformatics[25] and chemical crystallography[26]. In human palaeogenetics, it appears that open data sharing is almost universal (>97%), not as a consequence of top-down requirements, but because of awareness of its value by the relevant research community.[27] Moreover, a growing number of researchers share their data from the start of their research projects, both to receive comments from peers and to engage in open collaboration. These developments are sensitive to the needs of the disciplines involved, they provide

21 http://www.codata.org/

22 https://www.icsu-wds.org/

23 https://rd-alliance.org/node

24 http://www.linguistic-lod.org/llod-cloud

25 https://www.elixir-europe.org/

26 http://www.crystallography.net/

27 Anagnostou, P., Capocasa, M., Milia, N., Sanna, E., Battaggia, C., Luzi, D. and Destro Bisol, G. 2015. *When Data Sharing Gets Close to 100%: What Human Paleogenetics Can Teach the Open Science Movement.* PLOS ONE · March 2015, DOI: 10.1371/journal.pone.0121409

BOX 5

## Open Data Initiatives in Africa

Many African countries are energetically developing their own capacities to exploit the data revolution for the benefit of their public policies and economies.

### *United Nations perspective*

The UN report, A World that Counts (2015 – www.undatarevolution.org), sets out the public policy imperative to improve data gathering and to make data open for maximum impact and reuse: "Data are the lifeblood of decision-making. Without data, we cannot know how many people are born and at what age they die; how many men, women and children still live in poverty; how many children need educating; how many doctors to train or schools to build; how public money is being spent and to what effect; whether greenhouse gas emissions are increasing or the fish stocks in the ocean are dangerously low; how many people are in what kinds of work, what companies are trading and whether economic activity is expanding".

### *Open Data for Africa portal* (*www.opendataforafrica.org*)

This includes such data on food prices, GDP per capita, energy statistics, demographics, water, energy and energy forecasts, food, education, government debt, healthcare infrastructure, malaria, migration, mortality, urbanization etc.

### *National initiatives*

The Kenyan Data Forum (http://www.dataforum.or.ke/) emphasizes the need for the domestication of the data revolution as a key step in accelerating implementation of the national development agenda, which is aligned with regional and global goals. It convenes stakeholder communities from government, private sector, academia, civil society, local communities and development partners who engage on the informational aspects of development decision-making.

***Agriculture:*** Agriculture accounts for 65% of Africa's workforce and 32% of the continent's GDP. In some of Africa's poorest countries, including Chad and Sierra Leone, it accounts for more than 50% of GDP. The Global Open Data for Agriculture and Nutrition initiative (http://www.godan.info) recently published a report which asks 'How can we improve agriculture, food and nutrition with Open Data'. The report presents numerous case studies of precisely how Open Data can advance research and practice in these areas with numerous positive outcomes.

AgTrials is an example of improving crop varieties with open data about breeding trials. Scientists have used 250 open AgTrials datasets to build crop models specific to the West Africa region. The models are used to project the local impacts of climate change, addressing issues such as drought tolerance, heat stress, and soil management and defining breeding programmes for adaptation.

### Mobilising Science Capacity

To accompany this open data accord, Science International will promote a collaborative initiative involving the South African Government's Department of Science and Technology, other national science bodies in sub-Saharan Africa and CODATA and its international partners (RDA and WDS) in mobilising the African research community in developing big data/open data capacities.

BOX 6

## Open Data Platforms

### *A national data platform*

The National Science and Technology Infrastructure (NSTI) of the Peoples Republic of China is the networked, ITC based system that provides shared service for technology innovation and economic and social development. The NSTI programme supports 10 scientific data centres and 3 scientific data sharing networks. It integrates more than 50,000 science and technology databases in 32 categories and 10 technical fields, including agriculture, meteorology, seismicity, population health, materials, energy, geology, etc. It has established a managed service for scientific and technology data and information sharing, based on a series of standard specifications. Under this programme, a number of high-profile data and information sharing services have been set up. In 2011, NSTI supported the creation of 6 scientific data platforms, which facilitate standardised management of data resources and offer a quality-controlled service. By 2014, the NSTI platform website had received more than 50 million visits and provided 60 terabytes of information. The platform currently provides a service for nearly 3000 national key science and technology projects and plays an important role in innovation and public service. The NSTI is demand-driven: in specific instances it responds with comprehensive, systematic, special services, and creates scientific data products.

### *A disciplinary platform: ELIXIR — an integrated data support system for the life sciences.*

ELIXIR is the European life-science infrastructure for biological information. It is a unique and unprecedented initiative that consolidates Europe's national centres, services, and core bioinformatics resources into a single, coordinated infrastructure. It brings together Europe's major life-science data archives and, for the first time, connects them with national bioinformatics infrastructures throughout ELIXIR's member states. By coordinating local, national and international resources the ELIXIR infrastructure is designed to serve the data-related needs of Europe's 500,000 life-scientists. Open access to bioinformatics resources provides a valuable path to discovery. National nodes develop national strategies and are the sources of support for national communities and the route through which ELIXIR resources, including data, analytic software and other tools are accessed. There is a strong ethos of data sharing in many life science communities, but even here practices vary. In structural biology and genomics it is established practice to deposit sequence data as soon as it is acquired. In many fields it is a requirement to deposit data for publishing. In other areas, such as biomedical research, practice is varied, though there is strong pressure from funders for openness.

an open corpus of information for their communities that is far greater than any single researcher could acquire, offer support and advice, and animate creative collaboration between their members. It is important that top-down processes do not prescribe mechanisms that inhibit the development of such initiatives, but are able to learn from their success and be supportive of and adaptive to their needs through the provision of appropriate soft and hard infrastructures that are sensitive to local possibilities and resources.

## Open Science and Open Data

**26.** The idea of "open science" has developed in recognition of the need for stronger dialogue and engagement of the science community with wider society in addressing many current problems through reciprocal framing of issues and the collaborative design, execution and application of research. "Open data" (as a set of practices and a resource) is an essential part of that process. In an era of diminished deference and ubiquitous communication it is no longer adequate to announce scientific conclusions on matters of public interest and concern without providing the evidence (the data) that supports them, and which can therefore be subject to intense and rigorous scrutiny. The growth of citizen science, which involves many participants without formal research training, and the increasing participation of social actors other than scholars in co-creation of knowledge, are enriching local and global conversations on issues that affect us all and are eroding the boundary between professional and amateur scientists. At the same time, the apparent increase in fraudulent behaviour, much of which includes invention or spurious manipulation of data, risks undermining public trust in science, for which openness to scrutiny must be an important part of the necessary corrective action.

## Public Knowledge or Private Knowledge?

**27.** Open scientific data and the resulting knowledge have generally been regarded as public goods and a fundamental basis for human judgement, innovation and the wellbeing of society. Many governments now recognise the benefits of being open with their own data holdings in order to provide opportunities for creative commercial re-use of a public resource, to achieve specific public policy objectives, to increase government accountability and to be more responsive to citizens' needs. Access to such data can also be of considerable scientific value, particularly in the social sciences for evaluating social and economic trends, and in the medical sciences for evaluating optimal public health strategies from population health records. There are inter-governmental initiatives to promote openness, such as the *Open Government Partnership*[28], which now involves 66 participating countries worldwide, the G8 Open Data Charter[29] and the report to the UN Secretary-General from his Independent Advisory Group on *the Data Revolution for Sustainable Development*[30].

**28.** It is tempting to think that the boundary of open data is the boundary between the publicly funded and the commercially held, but this is not necessarily the case. Different business sectors take different approaches, with some benefitting from openness. For example, it is in the interests of manufacturers of environmental data acquisition systems for the data to be open in ways that stimulate new businesses based on novel ways of using them, thereby increasing demand for the hardware. The massive data volumes that are daily captured by retail and service industries offer great research potential if made available to social science researchers. Thus, policy makers have a responsibility to consider new ways of incentivising

BOX 7

## Opening up government data: the Indian strategy

The Indian National Data Sharing and Accessibility Policy, passed in February 2012, is designed to promote data sharing and enable access to Government of India-owned data for national planning and development. The Indian government recognises the need for open data in order to: maximise use, avoid duplication, maximise integration, spread ownership of information, and increase better decision-making and equity of access. Access will be through data.gov.in. As with other data.gov initiatives, the portal is designed to be user-friendly and web- based without any process of registration or authorisation. The accompanying metadata will be standardised and contain information on proper citation, access, contact information and discovery. The policy applies to all non-sensitive data available either in digital or analogue forms having been generated using public funds from within all Ministries, Departments and agencies of the Government of India.

28 www.opengovpartnership.org

29 https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex; see also the related G8 Science Ministers Statement, London, 12 June 2013: https://www.gov.uk/government/publications/g8-science-ministers-statement-london-12-june-2013

30 www.undatarevolution.org

private companies to make their data open. New forms of university-industry engagement around public and private data could generate important insights and benefits for science, society and the economy.

**29.** There is currently an important international debate about whether to make public data freely available and usable by everyone, or just the not-for-profit sector. Should the private, for-profit sector pay for access and use of publicly funded data? This is a complex issue, but as long as the original data remain openly available on the same terms to all, it does not seem sensible, appropriate or productive to discriminate between not-for-profit and for-profit users. Robust evidence is accumulating of the diverse benefits and broader economic and societal value derived from the open sharing of research data.[31]

**30.** It is however important to recognise that there is a countervailing trend to openness, of business models built on the capture and privatisation of socially produced knowledge through the monopoly and protection of data. Such trends towards privatisation of a public resource or uncontrolled and unconsented access to personal information are at odds with the ethos of scientific inquiry and the basic need of humanity to use ideas freely. If the scientific enterprise is not to founder under such pressures, an assertive commitment to open data, open information and open knowledge is required from the scientific community.

## C. Principles of Open Data

**31.** Such is the importance and magnitude of the challenges to the practice of science from the data revolution that Science International believes it appropriate to promote the following statement of principles of responsibility and of enabling practice for data-intensive science. Science International partners will advocate them for adoption by scientific unions, national representative science bodies and others that influence the operation of national and international science systems. The principles are an evolution of – but consistent with – priorities and recommendations set out in earlier reports on data-intensive science by Science International partners, by governmental and Inter-governmental bodies and by academic groups.[32] These principles recognise not only the benefits of open data and open science, but also the complexity of the international research landscape, with sometimes overlapping and sometimes competing needs and interests between different stakeholders. Section D sets out further rationale for the principles and practical options for their implementation.

### Responsibilities

#### *Scientists*

***i.*** Publicly funded scientists have a responsibility to contribute to the public good through the creation and communication of new knowledge, of which associated data are intrinsic parts. They should make such data openly available to others as soon as possible after their production in ways that permit them to be re-used and re-purposed.

***ii.*** The data that provide evidence for published scientific claims should be made concurrently and publicly available in an intelligently open form. This should permit the logic of the link between data and claim to be rigorously scrutinised and the validity of the data to be tested by replication of experiments or observations. To the extent possible, data should be deposited in well-managed and trusted repositories with low access barriers.

#### *iii. Research institutions and universities*

have a responsibility to create a supportive environment for open data. This includes the provision of training in data management, preservation and analysis and of relevant technical support, library and data management services. Institutions that employ scientists and bodies that fund them should develop incentives and criteria for career advancement for those involved in open data processes. Consensus on such criteria is necessary nationally, and ideally internationally, to facilitate desirable patterns of researcher mobility. In the current spirit of internationalisation, universities and other science institutions in developed countries should collaborate with their counterparts in developing countries to mobilise data-intensive capacities.

#### *iv. Publishers*

have a responsibility to make data available to reviewers during the review process, to require intelligently open access to the data concurrently with the publication which uses them and to require the full referencing and citation of these data. Publishers also have a responsibility to make the scientific record available for subsequent analysis through the open provision of metadata and open access for text and data mining.

#### *vi. Funding agencies*

should regard the costs of open data processes in a research project to be an intrinsic part of the cost of doing the research, and should provide adequate resources and policies for long term sustainability of infrastructure and repositories. Assessment of research impact, particularly any involving citation metrics, should take due account of the contribution of data creators.

#### *vii. Professional associations, scholarly societies and academies*

should develop guidelines and policies for open data and promote the opportunities they offer in ways that reflect the epistemic norms and practices of their members.

#### *viii. Libraries, archives and repositories*

have a responsibility for the development and provision of services and technical standards for data to ensure that data are available to those who wish to use them and that data are accessible over the long term.

### Boundaries of openness

***viii.*** Open data should be the default position for publicly funded science. Exceptions should be limited to issues of privacy, safety, security and to commercial use that is in the public interest. Exceptions should be justified on a case-by-case and not blanket basis.

### Enabling practices

#### *ix. Citation and provenance*

When, in scholarly publications, researchers use data created by others, those data should be cited with reference to their originator, their provenance and to a permanent digital identifier.

#### *x. Interoperability*

Both research data, and the metadata which allows them to be assessed and reused, should be interoperable to the greatest degree possible.

#### *xi. Non-restrictive reuse*

If research data are not already in the public domain, they should be labelled as reusable by means of a rights waiver or non-restrictive licence that makes it clear that the data may be reused with no more arduous requirement than that of acknowledging the prior producer(s).

31 An brief yet comprehensive survey of current evidence is provided in Paul Uhlir for CODATA (2015) *The Value of Open Data Sharing: A White Paper for the Group on Earth Observations* http://dx.doi.org/10.5281/zenodo.33830

32 Reports by Science International partners include: ICSU-CODATA 2000; IAP 2003; CODATA 2014. Governmental or inter-governmental statements include: Bromley 1991; WMO 1995; OECD 2007 and 2008; and G8 2013. Academic statements include: the Bermuda Principles 1996; Berlin Declaration 2003; The Royal Society 2012; Bouchout Declaration 2014; Hague Declaration 2014; and RECODE Project 2015. A compendium of many national and international policy documents for Open Data may be found at: Sunlight Foundation 2015 or Open Access Directory 2015. Further statements are referenced in appendix 2.

*xii. Linkability*

Open data should, as often as possible, be linked with other data based on their content and context in order to maximise their semantic value.

## D. The Practice of Open Data

**32.** This section expands on the rationale for the above principles and consequential issues of practice that should be addressed.

### Responsibilities

*Normative values*

**33.** The accord makes the normative assertion that publicly funded research should be undertaken in a way that creates maximum public benefit. It argues that the open release of data is the optimal route by which this is achieved.

**34.** The argument that such openness should be openness to the world and not merely contained within national boundaries is part of both the utilitarian and normative arguments for open publication:

- that no one country dominates the international scientific effort and that maximum national benefit is gained if all openly publish their results and all are able to utilise them;
- that the acquisition of knowledge is an essential human enterprise and should be open to all.

Statements and reports that emphasise these priorities are referenced in appendix 2.

*Data used as evidence for a scientific claim*

**35.** The data that provide evidence for a published scientific claim must be concurrently published in a way that permits the logic of the link between data and claim to be rigorously scrutinised and the validity of the data to be tested by replication of experiments or observations. To do otherwise should be regarded as scientific malpractice. The intelligent openness criteria of principle ii should be applied to the data. It is generally impracticable for large data volumes to be included in a conventional scientific publication, but such data should be referenced by means of a citation including a permanent digital identifier and should be curated in and accessible from a trusted repository.

**36.** The main responsibility for upholding this important principle of science lies with researchers themselves. However, given the onerous nature of this task in areas of data-intensive science, it is important that institutions create support processes that minimise the burden on individual scientists. It is a false dichotomy to argue that there is a choice to be made between funding provision for open data and funding more research. The practice of open data is a fundamental part of the process of doing science properly, and cannot be separated from it.

**37.** Responsibilities for ensuring that this principle is upheld also lie with the funders of research, who should mandate open data by researchers that they fund,[33] and by publishers of scientific work, who should require, as a condition of publication, deposition of open data that provides the evidence for a claim that is submitted for publication. Funders should also accept that the cost of curation of open data is part of the cost of doing research and should expect to fund it.[34]

*National responsibilities*

**38.** The capacities required to efficiently implement and to maximise benefit from the application of the principles set out in this accord and the responsibility to do so are not exclusively those of researchers and their institutions. They depend upon mutually supporting, systemic responsibilities and relationships that need to be embedded at every level of both national and international science systems, operating as parts of a dynamic ecology. It is also important to recognise that individual and institutional interests are not necessarily identical to the interests of the scientific process or to national interests in stimulating and benefiting from open data. These issues of motivation need to be identified and addressed. Box 9 shows relationships between the two key elements of national infrastructure for open data, the hard technologies and the soft relationships and responsibilities (based on Deetjen, U., E. T. Meyer and R. Schroeder (2015), "Big Data for Advancing Dementia Research: An Evaluation of Data Sharing Practices in Research on Age-related Neurodegenerative Diseases", OECD Digital Economy Papers, No. 246, OECD Publishing. http://dx.doi.org/10.1787/5js4sbddf7jk-en).

**39.** We characterise responsibilities and relationships as follows:

**Publicly funded scientists** should recognise that the essential contribution to society of publicly funded research is to generate and communicate knowledge, and that open data practices are essential to its credibility and utility. This latter requirement poses two problems of motivation:

- preparing data and metadata in a way that would satisfy the criteria of "intelligent openness" is costly in time and effort;
- data are regarded by many as "their" data, and as a resource which they are able to draw on for successive publications that are conventional indices of personal productivity, sources of recognition and grist for promotion.

**Universities and Research Institutes** have a responsibility to address the above motivational issues by:

- providing support that minimises the burden of compliance for individual researchers and allows them to focus less on process and more on research;
- developing processes of advancement and recognition that recognise and reward open data activities, with the need to ensure broad commonality at international level so as not to inhibit researcher mobility.

They also need to provide a managed environment to train researchers in big data and linked data analytics and in open data management, to provide expert support in these areas, and to manage open data processes.

**Institutional Libraries** have a continuing role to collect, to organize, to preserve knowledge, and to make it accessible. Many are now adapting to the technological change from paper to digital formats and to the open data management issues highlighted by this accord, but it is a major and difficult transition that requires sustained effort.

**Funders of Research and Research Institutions** have a responsibility to promote and enable open data processes by funding relevant hard and soft infrastructure; by stimulating research on fundamentals of data science; and by creating incentives for research performing institutions that help them to exercise their responsibilities and accepting that the cost of open data is an inseparable cost of doing research.

**Governments** hold data that are of great value to the scientific enterprise if made open, particularly in the social sciences, in addition to the broader societal value that they may create. Governments should also express broad national policies and objectives that are important in providing a frame for national efforts in developing an open data environment and system priorities, though they should not prescribe how they should be delivered.

**National Academies and Learned Societies** are distinctive in speaking to scientists directly without institutional intermediaries and influencing "bottom-up" initiatives by expressing the principles and priorities of research in their specific fields. They should develop guidelines and policies for open data and promote the opportunities they offer in ways that reflect the epistemic norms and practices of their members.

33 See the comprehensive survey of funder data policies: Hodson and Molloy (2014) *Current Best Practice for Research Data Management Policies* http://dx.doi.org/10.5281/zenodo.27872

34 See for example the RCUK Common Principles on Data Policy http://www.rcuk.ac.uk/research/datapolicy/; see also the discussion of policy positions on the costs of RDM in Hodson and Molloy (2014) http://dx.doi.org/10.5281/zenodo.27872 pp. 11 – 12.

**40.** Ensuring a sustainable data infrastructure (including the management systems, standards, procedures and analysis tools for what is often called 'live' or 'active' data and the infrastructure of 'Trusted Digital Repositories' –TDRs– for long term curation of valuable data) is a core responsibility of research funders and research performing organisations (see below, para. 62–64). As emphasised above, it is a false dichotomy to argue that there is a choice to be made between funding provision for open data and funding more research. The practice of open data is a fundamental part of the process of doing science properly, and cannot be separated from it. Data infrastructure forms an essential tool for science, as necessary as networked and high performance computers, access to high quality scientific literature, in vitro labs and organic or inorganic samples.

### *International responsibilities*

**41.** International science organisations play an important role in establishing principles and encouraging practices to ensure the worldwide adoption of "open data" and "open science" regimes to maintain the rigour of scientific processes and take advantage of the data revolution. Many have already developed their own data principles or protocols, as noted above. They can also help ensure that some of the most influential stakeholders are mobilised. The most effective examples of open data transformations have occurred when individual research communities, including funders, learned societies or international scientific unions, journals and major research performing organisations have endorsed community principles for open data sharing. Those established for the international genomics community are the most well known and successful, but there are others.[35]

**42.** It is a responsibility of the international science community to ensure that as far as possible, the capacities and the means to take up the big data and open data challenges are developed in all countries, irrespective of national income. It is for this reason that Science International and its parent bodies collaborate with low- and middle-income countries in capacity building programmes. In order to minimise such a knowledge divide, and resulting fragmentation, CODATA in collaboration with the RDA has organised relevant training workshops,[36] and Science International is currently discussing the possibility of launching a major big data/open data capacity mobilisation exercise for low- and middle-income countries, starting with an initiative in Africa. The rationale for this initiative is the danger that if a low income country has little capacity in modern data handling, its own data resources are likely either to be kept behind closed doors to protect it from foreign exploitation or, if open, to be exploited by such groups without reciprocal benefit to the host. If national capacities are mobilised, not only is a country able to exploit its own national data resources but also those that are available internationally.

**43.** Transformative initiatives, however resoundingly endorsed in principle, will be ineffective without investment in education and skills. The need to inculcate the ethos of Open Science outlined above and to develop data science and data handling skills for researchers is widely recognised.[37] Additionally, there are well-documented calls to develop skills and career paths for the various data-related professions that are essential to research institutions in a data-intensive age: these include data analysts, data managers, data curators and data librarians.[38]

## Scientific publishers

**44.** Publishers of research papers that present scientific claims should require the evidential data to be concurrently made intelligently open in a trusted data repository. It is a fundamental principle of transparency and reproducibility in research that the data underlying a claim should be accessible for testing[39]. A model for good practice can be found in the Joint Data Archiving Policy that underpins the role of the Dryad Data Repository[40]. Journal editors, editorial boards, learned societies and journal publishers share responsibility to ensure such principles are adopted and implemented. Data infrastructure, comprising specialist, generic data archives and institutional data repositories which support these practices are now emerging in national jurisdictions and some international programmes[41]. The international science community should promote worldwide capability in these areas. Furthermore, journal publishers and

35 See the summary of genomics data sharing agreements at http://www.genome.gov/page.cfm?pageID=10506537; there is longstanding but far from comprehensive data sharing in the astronomical and geophysical sciences as well as in the social sciences; crystallographers successfully publish final, 'science ready' data using the CIF standard http://www.iucr.org/resources/cif

36 See the CODATA-RDA Research Data Science 'Summer Schools' or short courses http://www.codata.org/working-groups/research-data-science-summer-schools

37 The CODATA-RDA Research Data Science courses start from the premise that 'Contemporary research — particularly when addressing the most significant, transdisciplinary research challenges — cannot effectively be done without a range of skills relating to data. This includes the principles and practice of Open Science and research data management and curation, the use of a range of data platforms and infrastructures, large scale analysis, statistics, visualisation and modelling techniques, software development and annotation, etc, etc. The ensemble of these skills, we define as 'Research Data Science'.'

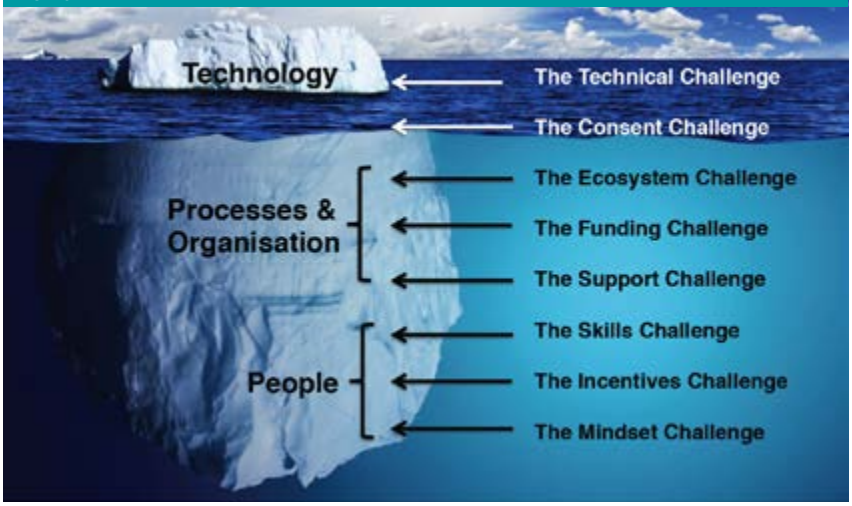
38 See for example the ANDS page on 'Data Librarians' http://ands.org.au/guides/dmframework/dmskills-information.html and the Harvard 'Data Science Training for Librarians' http://altbibl.io/dst4l/

39 The Royal Society's 'Science as an Open Enterprise' report stated: 'As a first step towards this intelligent openness, data that underpin a journal article should be made concurrently available in an accessible database. We are now on the brink of an achievable aim: for all science literature to be online, for all of the data to be online and for the two to be interoperable.' Royal Society 2012, p. 7.

40 Joint Data Archiving Policy (JDAP): 'This journal requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as GenBank, TreeBASE, Dryad, or the Knowledge Network for Biocomplexity.' http://datadryad.org/pages/jdaphttp://datadryad.org/pages/jdap

41 For example, the Pangaea data archive has bidirectional linking between datasets and articles in Elsevier journals. Dryad, FigShare and now Mendeley provide repositories for data underlying journal articles. In addition to specialist, discipline specific repositories, the generic repositories like FigShare and Zenodo provides places where researchers can deposit datasets. An increasing number of research institutions are providing repositories for data outputs of research conducted in the institution.

**BOX 8**




The infrastructure requirements for an efficient open data environment. Technology is only a part. The vital, submerged elements relate to processes, organisation and personal skills, motivation and ethos.

editors have increasingly realised that providing direct access to the data, sometimes with visualisation, increases the appeal of the journal[42]. It is not however sufficient for data to be accessible only as poorly described 'supplementary materials' provided in formats that hamper reuse. Data that directly support research articles should not lie behind a paywall. As the OECD Principles and Guidelines on Access to Research Data from Public Funding make clear, it is not legitimate for purely commercial reasons to close access to those data which have been gathered with the support of public funds and those which support published research findings.[43] However, it can be legitimate for repositories to monetise data products for which there has been considerable value-adding investment in order, for example, to present useful and reliable reference data for researchers.

## The boundaries of openness

**45.** Openness as defined above should be the default position for scientific data although there are proportional exceptions for cases of legitimate commercial exploitation, privacy and confidentiality, and safety and security. Not all data should be made available and there are well-recognised reasons when this is the case. However, it should be recognised that open release of data is the default, such that the exceptions listed must not be used to justify blanket exceptions to openness. Rather, as it is difficult to draw sharp, general boundaries for each of these cases, they should be applied with discrimination on a case-by-case basis. Important considerations at these boundaries include:

### *Commercial interests*

**46.** There can be a public interest in the commercialisation of scientific discovery where that is the route to the greatest public benefit in the national jurisdiction in which the discovery is made. The case for long-term suppression of data release on commercial grounds is weak however. Patenting is a means of protecting intellectual property whilst permitting release of important scientific data. Demands for confidentiality from commercial partners may exercise a chilling effect on swathes of research activity and the openness that should characterise it. There have been many major discoveries where suppression of data release or the privatisation of knowledge would have been highly retrograde, such as the discovery of electricity, the human genetic code, the internet etc. Difficult and potentially contentious issues include: where there has been a public/private partnership in investing in a scientific discovery; where the contribution of a private contributor should not be automatically assumed to negate openness; where commercial activities carry externalities that influence societal individual wellbeing; and where the data supporting a risk analysis should be made public.

### *Privacy and confidentiality*

**47.** The sharing of datasets containing personal information is of critical importance for research in many areas of the medical and social sciences, but poses challenges for information governance and the protection of confidentiality. There can be a strong public interest in managed openness in many such cases provided it is performed under an appropriate governance framework. This framework must adapt to the fact that other than in cases where the range of data is very limited, complete anonymisation of personal records in databases is impossible. In some cases, consent for data release can be appropriate. Where this is not possible, an effective way of dealing with such issues is through what are sometimes called "safe havens", where data are kept physically secure, and only made available to bona fide researchers, with legal sanctions against unauthorised release.[44]

### *Safety and security*

**48.** Careful scrutiny of the boundaries of openness is important where research could in principle be misused to threaten security, public safety or health. It is important in such cases to take a balanced and proportionate approach rather than a blanket prohibition. Scientific discoveries often have potential dual uses — for benefit or for harm. However, cases where national security concerns are sufficient to warrant wholesale refusal to publish datasets are rare.[45] and cultural choice whether to encourage or obstruct its pursuit.

## Enabling practices

### *Timeliness of data release*

**49.** Data should be released into the public domain as soon as possible after their creation. Data that underpin a scientific claim should be released into the public domain concurrently with the publication of the claim. Where research projects have created datasets with significant reuse value, and particularly when such projects are publicly funded, the data outputs should also be released as soon as possible.[46] Recognising the effort involved in data creation and the intellectual capital invested, the policies of some funders allow public release to be delayed for precisely limited periods, allowing data creators privileged access to exploit the asset. In contrast, however, the genomics community has demonstrated the benefits of immediate data release.[47] It is important to evaluate the benefits of immediate release in other research domains.

### *Non-restrictive re-use*

**50.** Research data should be dedicated to the public domain by legal means that provide certainty to the users of the right of their re-use, re-dissemination and, for cases where research is conducted over multiple datasets, their "legal interoperability".[48] This can be accomplished by a variety of means, either broadly, as a governmental agreement, statute or policy, or as a narrow waiver of rights or a non-restrictive license that applies to a specific database or data product on a voluntary basis. The RDA-CODATA Interest Group on Legal Interoperability of Research Data has produced Principles and Implementation Guidelines that are currently in review.[49]

42 Both FigShare http://figshare.com/blog/figshare_partners_with_Open_Access_mega_journal_publisher_PLOS/68 and Dryad now provide 'widgets' which allow simple visualisations of data associated with a given article. Nevertheless, the so-called 'article of the future' is taking quite a long time to become a reality in the present ... [e. g. see http://scholarlykitchen.sspnet.org/2009/07/21/the-article-of-the-future-lipstick-on-a-pig/]

43 See OECD Principles and Guidelines for Access to Research Data from Public Funding http://www.oecd.org/sti/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm and other statements of principle like the RCUK Common Principles on Data Policy http://www.rcuk.ac.uk/research/datapolicy/; Uhlir, Paul for CODATA (2015) marshals evidence to demonstrate that greater economic benefits and return on public investment are achieved through open data that through charging regimes designed to recover costs of data distribution.

44 See, e.g. the workshop and report on data safe havens from the Academy of Medical Sciences http://www.acmedsci.ac.uk/policy/policy-projects/data-in-safe-havens/; see also the UKDA Secure Data Service https://www.ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab; and the restricted use data held by ICPSR http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/restricted/

45 See Royal Society, 2006. *Report of the RS-ISP-ICSU international workshop on science and technology developments relevant to the Biological and Toxin Weapons Convention.*

46 These categories of research data to be shared are identified, for example, in the EC's Horizon 2020 Data Policy, see Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, p. 10; http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot- guide_en.pdf

47 See the summary of data release policies in the genomics field at http://www.genome.gov/page.cfm?pageID=10506537 and the more general discussion and summary of period of privileged access in Hodson and Molloy (2014) Current Best Practice for Research Data Management Policies, p. 18 http://dx.doi.org/10.5281/zenodo.27872

48 "Legal interoperability occurs among multiple datasets when: i) use conditions are clearly and readily determinable for each of the datasets; ii) the legal use conditions imposed on each dataset allow creation and use of combined or derivative products; and, iii) users may legally access and use each dataset without seeking authorization from data rights holders on a case-by-case basis, assuming that the accumulated conditions of use for each and all of the datasets are met." Definition provided in GEO (2014) *White Paper: Mechanisms to Share Data as Part of the GEOSS Data-CORE. Data Sharing Working Group.* Available at: https://www.earthobservations.org/documents/dswg/Annex%20VI%20-%20%20Mechanisms%20to%20share%20data%20as%20part%20of%20GEOSS%20Data_CORE.pdf

49 The final Implementation Guidelines for the Principles on the Legal Interoperability of Research Data developed by the CODATA-RDA Group will be released in March 2016 following community review.

**51.** The broadest approach to placing research data in the public domain is to develop and use a convention or executive agreement at the international level, or legislation or executive policies at the national level. For example, the U.S. federal government excludes all information produced within its ambit from copyright protection under the 1977 Copyright Law. Different ministries or research agencies may adopt a policy that allows research data produced through their funding to be placed in the public domain. Because it is more difficult to agree to such far-reaching exemptions from intellectual property protection, the rights holder also may expressly state on a voluntary basis that the data are in the public domain.

**52.** In the absence of a broad law that enables the re-use, re-dissemination and legal interoperability of data, a voluntary rights waiver or a non-restrictive, "common-use" licence can be used by the rights holder (see: www.creativecommons.org). If a non-restrictive license is used, it should make it clear that the data may be reused with no more arduous requirement than that of acknowledging the original producer of the data. It is good practice to use a public domain waiver of rights (e.g. CC0) or non-restrictive licence (such as CC-BY). The license requires nothing more than that the producer of the data is acknowledged. Imposing further restrictions against commercial use defeats the objectives of open data and the dedication of those data to the public.[50]

**53.** Although the use of an attribution-only (CC-BY) license may be appropriate in some circumstances, the challenges associated with providing recognition to the generators of datasets integrated into complex data products, a phenomenon of data-intensive research, means that many authorities argue that licences such as CC-BY that require attribution are not sustainable or appropriate in a Big Data age.[51]

### *Citation and provenance*

**54.** When used in scholarly communication, research data must be cited with reference to specific information and a permanent digital identifier[52]. The information attached to the citation and the identifier must allow the provenance of the data to be assessed. The practice of citing data in scholarly discourse is important for two reasons. First, citing sources is essential to the practice of evidence-based reasoning and distinguishes scientific texts from other writing. Second, 'citations' are one of the metrics by which research contributions are assessed. Although not without flaws and subject to possible gaming, article-level citation metrics are the "least bad" means of measuring research contribution and are without doubt an improvement on journal level impact factors.[53]

**55.** It would be naïve to pretend that citation is not an important component of the system of academic recognition and reward. Therefore, integrating the practice of data citation must be seen as an important step in providing incentives for 'data sharing'.

**56.** Citations also provide essential information – metadata – that allow the data to be retrieved. A permanent digital identifier (for example, a Digital Object Identifier issued by the DataCite organisation)[54] allows other researchers to determine without ambiguity that the data in question were indeed those which underpin the scientific claim at issue. This is particularly important when dynamically created subsets or specific versions of time-series datasets may be at issue.[55]

**57.** Additional metadata is necessary to determine the provenance of the data and to understand the circumstances in which they were created and in what way they may be reused. Standards exist in most research disciplines for the way in which data should be described and the circumstances of their creation reported.[56]

### *Text and data mining*

**58.** The historical record of scientific discovery and analysis published in scientific journals should be accessible to text and data mining (TDM). At the very least, this should be at no additional cost by scientists from journals to which their institution already subscribes, though there is a case for broader access to the corpus of scientific literature for TDM. The importance for science lies in the unprecedented capacity offered by text and data mining to harvest the cumulative scientific knowledge of a phenomenon from already published work. TDM has the potential to greatly enhance innovation. It can lead to an exponential increase in the progress of the rate of discovery, such as when facilitating the discovery of cures for serious diseases.

**59.** The Hague Declaration on Knowledge Discovery in the Digital Age[57], lays out the scientific and ethical rationale for the untrammelled freedom to deploy TDM in order to analyse scientific literature at scale. The Hague Declaration asserts that 'Intellectual property was not designed to regulate the free flow of facts, data and ideas, but has as a key objective the promotion of research activity'. In the digital age, the benefits of TDM are vast and necessary in order to support systematic review of the literature through machine analysis. Publisher resistance to TDM on the grounds of defending intellectual property are weak in the light of a skewed business model in which scientists sign copyright transfer agreements, make up journals' editorial boards and reviewer cohorts at no cost to the publisher, whilst scientists then pay to publish, and institutions pay for electronic copies of journals. There has been strong academic criticism of commercial publishers of research for claimed restrictive business practices and excessive profits[58].

### *Interoperability*

**60.** Research data, and the metadata which allow them to be assessed and reused, should be interoperable to the greatest degree possible. Interoperability may be defined as the 'property of a product or system … to work with other products or systems, present or future, without any restricted access or implementation.'[59] Interoperability is an attribute that greatly facilitates usability of research data. For example, semantic interoperability depends on shared and unambiguous properties and vocabulary, to which data refer, allowing comparison or integration at scale.

**61.** In relation to data, interoperability implies a number of attributes. These include the following:

- The encodings should be open and non-proprietary and there should be ready sources of reference, of a high quality, that allow the data to be ingested to other systems.
- The values which the data represent should use units describing properties for which there are standardised definitions.
- Standardised ontologies that are a key to interoperability.
- Metadata, particularly those reporting how the data were created

50 The DCC Guide 'How to Licence Research Data' is a very useful resource on this issue http://www.dcc.ac.uk/resources/how-guides/license-research-data

51 Carroll MW (2015) Sharing Research Data and Intellectual Property Law: A Primer. PLoS Biol 13(8): e1002235. doi:10.1371/journal.pbio.1002235

52 See the Joint Declaration of Data Citation Principles https://www.force11.org/group/joint-declaration-data-citation-principles-final.

53 For example: Arnold, Douglas N. and Kristine K. Fowler. "Nefarious Numbers." Notices of the American Mathematical Society v. 58, no. 3 (March 2011): 434 – 437. http://www.ams.org/notices/201103/rtx110300434p.pdf
Carlton M. Caves, "High-impact-factor Syndrome", APSNEWS November 2014 · Vol. 23, No. 10, http://aps.org/publications/apsnews/201411/backpage.cfm

54 https://www.datacite.org/

55 Research Data Alliance Working Group on Data Citation: https://rd-alliance.org/filedepot/folder/262?fid=667

56 See the RDA Metadata Standards Directory http://rd-alliance.github.io/metadata-directory/ building on work by the UK's Digital Curation Centre http://www.dcc.ac.uk/resources/metadata-standards; and the BioSharing catalogue of standards https://www.biosharing.org/standards/

57 The Hague Declaration on Knowledge Discovery in the Digital Age http://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age/

58 Harvie, D., Lightfoot, G., Lilley, S. and Weir, K. 2014. Publisher be damned! From price gouging to the open road. Prometheus: Critical Studies in Innovation. Vol. 31, No. 3, 229–239, http://dx.doi.org/10.1080/08109028.2014.891710

59 See http://interoperability-definition.info/en

and the characteristics of the properties should use, where possible, accepted standards.

### *Sustainable data deposition*

**62.** To ensure long-term stewardship in a sustainable data infrastructure, research data should be deposited in trusted digital repositories (TDR).[60] A TDR has the following attributes:

- an explicit mission to provide access to data and to preserve them in a defined area of competency;
- expertise and practices that conform to the principles laid out above;
- responsibility for long-term preservation and management of this function in a planned and documented way;
- an appropriate business model and funding streams to ensure sustainability in foreseeable circumstances;
- a continuity plan to ensure ongoing access to and preservation of its holdings in the case of wind-down.

**63.** Most trusted digital repositories cater for well-defined research disciplines, providing an appropriate and efficient focus of effort. However, the scale of the challenges and opportunities are such that multi-disciplinary repositories are emerging and research-performing institutions need also to provide TDRs to manage their research data outputs.

**64.** Research funders and national infrastructure providers have an obligation to ensure that an ecology of TDRs functions on a sustainable footing. This involves some serious rethinking of business and funding models for these essential but often undervalued elements of the research infrastructure.

### *Incentives*

**65.** Actions that encourage appropriate open data practices fall into three categories – those that encourage researchers to make data open, those that encourage the use of open data, and those that discourage closed data practices. The potential roles of four key actors need to be considered – research funders, institutions, publishers and researchers themselves. These actors are the key elements of the research community. They need to work together to ensure that data are considered legitimate, citable products of research; with data citations being accorded the same importance in the scholarly record as citations of other research objects, such as publications[61].

**66.** A developing method for researchers to gain credit for their data activities is through the formal publication and then citation of datasets, often via the route of a peer-reviewed data paper. There are a growing number of journals which either focus on publishing data papers, or have data papers as one of the article types within the journal.[62] These published datasets can then be formally cited within a research paper that makes use of the data, allowing the use and impact of the datasets to be tracked and rewarded in the same way as research papers. Many specialised data repositories – as well as the new multi-disciplinary data repository infrastructures, such as Dryad,[63] Figshare[64] and Zenodo,[65] which place particular emphasis on this feature – provide digital object identifiers (DOIs) for datasets they hold, which can then be referenced when the data are reused, providing credit for the data provider.

**67.** Institutions, especially funders, can reward data sharing by refining their research assessment analyses and other impact assessments, including those related to tenure and promotion, to include recognition of the considerable contribution to research of making data available for reuse.

**68.** By providing dedicated funding lines to support the reuse of open data, funders can start to encourage researchers to begin to unlock the value within open data. For example, the UK's Economic and Social Research Council is supporting a Secondary Data Analysis Initiative[66] which aims to deliver high-quality, high-impact research through the deeper exploitation of major data resources created by the ESRC and other agencies. Such dedicated funding can help facilitate the development of a re-use culture within research communities.

**69.** Journals have a key role in ensuring that researchers make their data open, by requiring that the data that underpin the research are openly available for others, and that research papers include statements on access to the underlying research materials. Major publishers, such as PLoS and Nature now have formal data policies in place, and many publishers are actively considering how to ensure that data availability becomes a mandatory part of the publication workflow.[67]

**70.** It is now common for research funders to have policies that require data arising from the research they fund to be made openly available where practical.[68] What is currently less common is for funders to monitor the adherence to their policies and to sanction researchers who do not comply. However, some funders are now starting to address this issue.[69]

60 See the foundational work done by OCLC on 'Attributes of Trusted Digital Repositories' http://www.oclc.org/research/activities/trustedrep.html. The Data Seal of Approval http://datasealofapproval.org/en/ and the ICSU World Data System's certification procedure https://www.icsu-wds.org/services/certification each offer lightweight and basic approaches to assessment of trusted digital repositories. More in-depth accreditation is offered by DIN 31644 — Criteria for trustworthy digital archives http://www.din.de/en/getting-involved/standards-committees/nabd/standards/wdc-beuth:din21:147058907 and ISO 16363 — Audit and certification of trustworthy digital repositories http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510

61 See the Joint Declaration of Data Citation Principles (ref Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/datacitation]).

62 Examples include: Nature Scientific Data, CODATA Data Science Journal, Wiley — Geoscience Data Journal, Ubiquity Press Metajournals like the Journal of Open Archaeology Data http://openarchaeologydata.metajnl.com/ and the Journal of Open Research Software http://openresearchsoftware.metajnl.com/

63 http://datadryad.org/

64 http://figshare.com/

65 https://zenodo.org/

66 http://www.esrc.ac.uk/research/our-research/secondary-data-analysis-initiative/

67 See: http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/ and http://www.nature.com/authors/policies/availability.html

68 For example, in the UK see http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies

69 For example EPSRC dipstick testing — https://www.jisc.ac.uk/guides/meeting-the-requirements-of-the-EPSRC-research-data-policy

## Appendix 1: Working Group members

**Geoffrey Boulton**, Regius Professor of Geology Emeritus in the University of Edinburgh and President of the Committee on Data for Science and Technology. Working Group Chair.

**Dr. Dominique Babini**, Coordinator of the Latin American Council of Social Sciences Open Access Program (ISSC representative).

**Dr. Simon Hodson,** Executive Director of the Committee on Data for Science and Technology (ICSU representative).

**Dr. Jianhui LI,** Assistant Director General of the Computer Network Information Centre, Chinese Academy of Sciences (IAP representative).

**Professor Tshilidzi Marwala**, Deputy Vice Chancellor for Research, University of Johannesburg (TWAS representative).

**Professor Maria G. N. Musoke,** University Librarian of Makerere University, Uganda, and Professor of Information Sciences (IAP representative).

**Dr. Paul F. Uhlir, Scholar,** US National Academy of Sciences, and Consultant, Data Policy and Management (IAP representative).

**Professor Sally Wyatt,** Professor of Digital Cultures in Development, Maastricht University, & Programme Leader of the eHumanities Group, Royal Netherlands Academy of Arts and Sciences (ISSC representative).

## Appendix 2: Statements and reports on open data


Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003).
Available at: http://openaccess.mpg.de/Berlin-Declaration.

Bermuda Principles. *Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing.* Human Genome Organization (1996). Available at: http://www.casimir.org.uk/storyfiles/64.0.summary_of_bermuda_principles.pdf.

Bouchout Declaration. *Bouchout Declaration for Open Biodiversity Knowledge Management.* Plazi (2014).
Available at: http://bouchoutdeclaration.org/.

Bromley, D. Allen. Principles on Full and Open Access to "Global Change" Data, Policy Statements on Data Management for Global Change Research. Office of Science and Technology Policy (1991).

Carroll, MW. Sharing Research Data and Intellectual Property Law: A Primer. PLoS Biol 13(8) (2015.

CODATA. *Nairobi Principles on Data Sharing for Science and Development in Developing Countries.* CODATA (2014). Available at: https://rd-alliance.org/sites/default/files/attachment/NairobiDataSharingPrinciples.pdf.

G8. *Open Data Charter* (2013). Available at:
https://www.gov.uk/government/publications/open-data-charter.

Group on Earth Observations. *Implementation Guidelines for the GEOSS Data Sharing Principles.* GEO VI, Document 7, Rev. 2 (17 – 18 November 2009). Available at: http://www.earthobservations.org/documents/geo_vi/07_Implementation%20Guidelines%20for%20the%20GEOSS%20Data%20Sharing%20Principles%20Rev2.pdf
*GEOSS Data Sharing Principles Post-2015.* Data Sharing Working Group (2014). Available at: https://www.earthobservations.org/documents/dswg/Annex%20III%20-%20GEOSS%20Data%20Sharing%20Principles%20Post-2015.pdf.

The Hague Declaration on Knowledge Discovery in the Digital Age. LIBER (2014). Available at: http://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digital-age/.

Hodson, Simon and Molloy, Laura, for CODATA (2014) *Current Best Practice for Research Data Management Policies* (commissioned from CODATA by the Danish e-Infrastructure Cooperation and the Danish Digital Library) http://dx.doi.org/10.5281/zenodo.27872

ICSU-CODATA Ad Hoc Group on Data and Information, Access to databases: A set of principles for science in the internet era (June 2000). Available at: http://www.icsu.org/publications/icsu-position-statements/access-to-databases/.

Interacademies Panel, IAP Statement on Access to Scientific Information (2002). Available at: http://www.interacademies.net/10878/13916.aspx.

Organisation for Economic Co-operation and Development (OECD). *Principles and Guidelines for Access to Research Data from Public Funding.* OECD (2007). Available at:
http://www.oecd-ilibrary.org/content/book/9789264034020-en-fr.

*Recommendation on Public Sector Information.* OECD (2008).
Available at: http://www.oecd.org/sti/44384673.pdf.

Open Access Directory (2015). Available at:
http://oad.simmons.edu/oadwiki/Declarations_in_support_of_OA.

RECODE Project. *Policy Guidelines for Open Access and Data Dissemination and Preservation.* European Commission (2015). Available at: http://recodeproject.eu/wp-content/uploads/2015/02/RECODE-D5.1-POLICY-RECOMMENDATIONS-_FINAL.pdf.

The Royal Society (2012). *Science as an Open Enterprise.* The Royal Society Policy Centre Report, 02/12. https://royalsociety.org/topics.../science...enterprise/report/

Uhlir, Paul for CODATA (2015) *The Value of Open Data Sharing: A White Paper for the Group on Earth Observations*
http://dx.doi.org/10.5281/zenodo.33830

A summary version of this accord can be found at
http://www.science-international.org

# Imprint

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

Science International
www.science-international.org

Cover photo: NASA

Design: Curie Kure, Hamburg
www.curiekure.de

ICSU
International Council for Science

iap
SCIENCE
RESEARCH
HEALTH
the interacademy partnership

ISSC
international social science council

twas

www.icsu.org
www.interacademies.net
www.worldsocialscience.org
www.twas.org

# The Geohazards Exploitation Platform (GEP)

## 1. An approach for integration of EO data products and toolboxes

As far as the value-added products are concerned, ESA focuses effort on hosted processing capabilities offered by the ESA Geohazards Exploitation Platform (GEP), providing users with a range of data processing and dissemination services. In general, the processing services are made available to users as-a-Service, where the users will be enabled to define the set of input data, the processing parameters, and trigger the execution of the algorithms.

In order to support the development of new algorithms, the ESA GEP also provides a Platform-as-a-Service capability (PaaS). In particular, Cloud Sandboxes enable developers and integrators to easily implement new algorithms. This solution makes use of the Virtual Machines technology and includes a middleware providing transparent interfaces to Cloud services, used for scaling-up the processing when increasing the dimensions of the input dataset. In case researchers in MARsite are interested in experimenting with their own algorithms on ESA data, GEP offers them direct access to dedicated Cloud Sandboxes.

The use of GEP removes the need of transferring huge amounts of input product data from the ESA archives to the users' machines, resulting in significant savings for the users.

### 1.1. The Open Science model

With GEP, the agency is providing partners with tools and infrastructure aimed at supporting geohazards researchers and practitioners with easy and open access to the ESA sensors data, community knowledge and expertise, and collaborative research.

The e-Science principle of open, reproducible, verifiable research experiments is a critical goal of Open science, tackling key scientific challenges such as the transparency of experimental approaches, traceability of the collection of observations and the public availability and reusability of scientific data.

Especially with regard to the exploitation of Earth observation (EO) data, there is a particular need to support researchers in setting up computer-intensive experiments, and for practitioners to better select and apply research outcomes in their exploitation tasks. In this perspective, GEP supports researchers in transitioning from siloed computational science activities addressing the simulation of complex phenomena, to e-Science scenarios expanding current capabilities to data exploration approaches: more integrated data access and processing, more support for documentation and sharing of processing experiments, and enhanced traceability across data and scientific literature.

In the geohazards domain, science users require satellite EO to support mitigation activities designed to reduce risk. These activities are carried out before the earthquake (or other geological peril) occurs, and they are presently the only effective way to reduce the impact of earthquakes on society. Short-term earthquake prediction today offers little promise of concrete results. The assessment of seismic hazard requires gathering geo-information for several aspects: the parameterization of the seismic sources, knowledge of historical and instrumental rates of seismicity, the measurement of present deformation rates, the partitioning of strain among different faults, paleo-seismological data from faults, and the improvement of tectonic models in seismogenic areas. Operational users in charge of seismic risk management have needs for geo-information to support mitigation. Satellite EO

can contribute by providing geo-information concerning crustal block boundaries to better map active faults, maps of strain to assess how rapidly faults are deforming, and geo-information concerning soil vulnerability to help estimate how the soil is behaving in reaction to seismic phenomena (read more from https://geohazards-tep.eo.esa.int/#!pages/initiative).

The ESA GEP exploits different types of cloud appliances. With GEP, it is possible to run computer-intensive workflows, enabled by specific cloud appliances that are built on "Hadoop Cloud Sandbox" components. Such Cloud Sandbox components were initiated by ESA funding as part of its G-POD Cloud evolutions (Cloud Interoperability Operational Pilot: An EO Sandbox Service) and then further developed through other EC FP7 projects, e.g. GEOWOW for the interface with the GEO Data Access and Broker (DAB).

For years now, ESA has been supporting the development and adoption of open source toolboxes (cf. https://earth.esa.int/web/guest/pi-community/toolboxes) for reading, processing, analysing and visualising ESA (ERS-1/2, Envisat, Sentinel-1) and other spaceborne sensor data, in SAR, Altimetry or Raster modes. The geohazards community have also been developing SAR data processors. For the first time, geohazards researchers and practitioners can join a shared platform where such toolboxes can be integrated, connected to large volumes of data, and exploited as-a-Service by end-users.

This approach for the integration of earth observation data, products, and toolboxes is central to the newly-launched ESA Thematic Exploitation Platforms (TEPs). The TEPs architecture is developed from a cloud-based e-Infrastructure approach focusing on the virtualization and federation of satellite EO applications. The TEPs are e-infrastructures providing scientists with a "testbed" to do their research without wasting time on ICT. Users are supported in exploiting resources made available as-a-service, in developing and testing their own computer-intensive processing chains, and overall in sharing data, knowledge and processing services across communities.

In this context, the GEP is providing a set of capabilities, that support Open Science activities through services for EO Data discovery, EO data access over distributed repositories, EO data consumption accounting, scalable on-demand EO processing, and sharing of value-added products.

In this approach, the GEP is a contribution to the global interoperability of open access data e-infrastructures, and provides a linkage with other global initiatives under GEO and Committee on Earth Observation Satellites (CEOS) umbrellas. This will allow the MARsite partners to deliver physical access to research resources, including EO data access and value-added products dissemination.

Considering drivers for change, open is better for the strengthening of links between science and society.

## 1.2. Links with GEOSS

We discussed how GEOWOW promoted new concepts and related tools as a European contribution to the GEOSS Common Infrastructure (GCI) and the GEOSS Data-CORE.

The PaaS capability of GEP inherited from these efforts, and provides today an environment for scientists to prepare data and scalable processing chains (including development and

testing of new algorithms), designed to automate the deployment of the resulting environment to a Cloud computing facility, that allows running compute-intensive tasks.

This approach towards the GEOSS community is being currently evolved and coordinated by ESA, through the following activities around GEP.

- First, by coordinating the GEP developments with the CEOS. GEP follows the SuperSites Exploitation Platform (SSEP), originally developed in the context of the Geohazards Supersites and Natural Laboratories initiative (GSNL). The geohazards platform has been expanded to address broader objectives of the geohazards community. In particular it is a contribution to the CEOS WG Disasters to support its Seismic Hazards Pilot and the terrain deformation applications of its Volcano Pilot.
- Second, by supporting users from the EC Supersites FP7 projects (MED-SUV, MARsite and FutureVOLC). These three projects are the EC contribution to the GEOSS GSNL. ESA in developing the GEP has accommodated a Validation phase starting in Q1 2015 for a set of Early Adopters, with a formalized relationship between a user and the GEP following its acceptance, to activate exploitation scenarios with specific guidance and support.

Support to the GEOSS researchers is also provided through a level of automation enabling users to reference work done on the Platform with the use of Digital Object Identifiers (DOI), where each application gets a DOI to track the service's impact through citations. This functionality is completed in GEP with job processing reproducibility (users can save and share the parameters of a processing job so that partners or reviewers can re-execute a job) and sharing functions (users can advertise their activity and results on major social media platforms).

As planned in May 2012, when ESA and the GEO Secretariat convened the International Forum on Satellite EO for Geohazards (the "Santorini Conference"), the seismic community has set out a vision of EO's contributions to an operational global seismic risk program. In 5 to 10 years' time, EO could provide fundamental new observations of the seismic belts - around 15% of the land surface and improved understanding of seismic events through the work of the GSNL. At that same Santorini Conference, the volcanic community identified priorities for satellite support to geohazards. In the long-term, the community aims to monitor all 1,500 Holocene era volcanoes on a global basis, a dramatic increase from the roughly 10% that are monitored now using both satellites and terrestrial sensors.

One of the core user communities for the GEP is the group of users and practitioners working on the CEOS Seismic Hazards Pilot, a three-year demonstration project of CEOS to showcase how satellite EO can be applied to seismic hazard research.

## 1.3. Links with the EPOS (ENVRI) infrastructure

EPOS (the European Plate Observing System) is a long-term integration plan of Research Infrastructures for solid Earth science in Europe. EPOS is promoting and making possible innovative approaches for a better understanding of processes controlling earthquakes, volcanic eruptions, tsunamis, surface dynamics and tectonics. EPOS is developing new concepts and tools to better address the grand challenges in solid Earth science. In particular, EPOS is promoting open access to geophysical and geological data as well as

modelling/processing tools, enabling a step change in multidisciplinary scientific research for Earth sciences.

From the EPOS Preparatory Phase, Working Group 8 (Satellite Information Data) addressed the goals of involving European data providers, i.e. the space agencies, defining a strategy for the acquisition of satellite data available to the EPOS Community, and exploring solutions for data repositories. One contribution of MARsite to this EPOS endeavour is to link EPOS Core Services to ESA Data Repositories.

In the light of the EPOS integration plan, the EC Supersites projects are part of the EPOS Community Layer - Thematic Services, while the GEP is part of the EPOS Integration Layer - Integrated Services (access to high performance computing, processing tools, access to data products). Having MARsite and GEP working together in the frame of the MARsite project provides a direct contribution to the EPOS vision and integration plan, especially when considering the EPOS principle of organising the Community Layer as a set of distributed data archives and services.

Other key EPOS concepts well supported by the GEP contribution in support of MARsite activities are the concept of “Product incubation chamber” (where scientists develop, test and improve new products before they are established as standard products). As a component of EC infrastructure projects, MARsite will benefit from the GEP added value to facilitate data discovery, enable repeatable workflows, facilitate effective large-scale data management, facilitate multidisciplinary analysis, and enable adequate data source identification. This will empower the MARsite project partners for their community involvement in setting the requirements for the EPOS integrating core services, and ultimately in building it.

## 2. ESA return of experience in the TEP Geohazards

In 2012, the European Space Agency started the SSEP flagship application within the Helix Nebula, Science Cloud initiative.

SSEP promoted a synthetic aperture radar (SAR) data exploitation platform focused on a few 'natural laboratories' around the world as defined within the GEO GSNL (http://www.earthobservations.org/gsnl.php) effort.

The SSEP brought together existing SAR interferometry EO toolboxes and EO data in a workspace that allowed researchers to run algorithms over data in a Cloud Computing environment. A paradigm shift was implemented that no longer relied on downloading data locally and running desktop software. Users were empowered with a set of tools hosted on the platform, using cloud and grid resources to enable compute-intensive tasks over massive data repositories.

The SSEP effort presented the first view over the assembled knowledge and experience from a large number of ESA assets. First, it was tied to the development and maintenance of the Agency's Grid Processing on Demand (G-POD, https://gpod.eo.esa.int/) system, and the related activities for integration of scientific applications. Second, it was empowered by new assets brought by the developments of Developer Cloud Sandbox services, matured through the Cloud Computing Interoperability Pilots (CIOP) project (http://wiki.services.eoportal.org/tiki-index.php?page=CIOP ), and evolved through the EC FP7 GEOWOW project (http://www.geowow.eu/) that concluded in September 2014.

Today, building on SSEP insights, the TEP Geohazards 'QuickWin' platform is a progression and enhancement towards a phase of GEP pre-operations. It brings the SSEP developments, mainly focused on proof-of-concept efforts, to a next level of federated architecture, with more power, additional tools (including pre-negotiated software licences with pay-per-use conditions, software renting,…), strengthened Web and Cloud Application Programming Interfaces (APIs) and a dedicated governance process to on-board users with a phased and coordinated approach (including user support services). At the current stage of these developments, outputs of the ESA's 'QuickWin' project are empowering a first community of 'early adopters' (from leading organizations like BGS, INGV, NOA,…), providing them with a Cloud platform having large data hosting capabilities, applications and services (e.g. InSAR processors and EO toolboxes) running on scalable Cloud resources without vendor lock-in (interoperability with several Cloud providers), and being exploitable through a new geobrowser service to access data, trigger processing tasks, and share the availability of EO-based products with other users on the platform as well as towards the larger community.

### 2.1. Leveraging the GEP to integrate Cloud Appliances as Processing Services

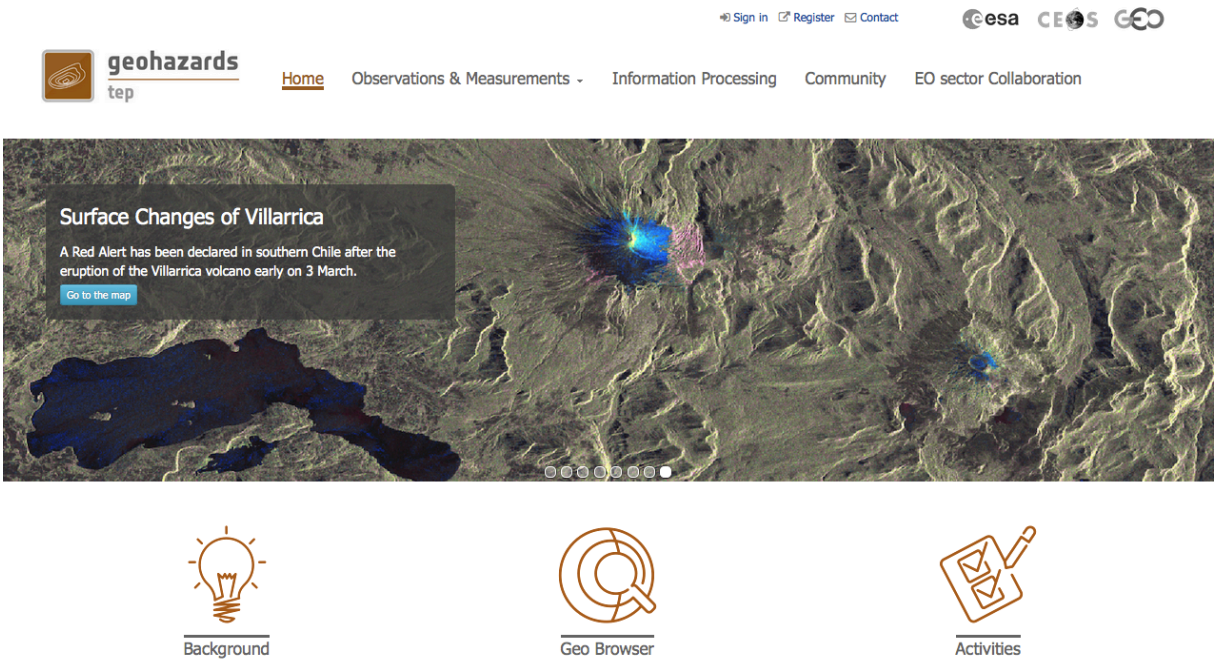
An "Exploitation Platform" refers to a virtual ICT environment, often cloud-based, providing users with very fast access to: (i) a large volume of data (EO/non-space data), (ii) computing resources (e.g. hybrid cloud/grid), and (iii) processing software (toolboxes, RTMs, retrieval schemes and visualization routines).

The idea underpinning exploitation platforms is to enable users to perform effectively data-intensive research by providing them with a virtual machine running dedicated processing

software close to the data, thereby avoiding moving large volumes of data through the network and spending non-research time on developing ICT tools.

The GEP portal is already accessible online for users (including public access level) at: http://geohazards-tep.eo.esa.int (cf. Figure 1).

**Figure 1 - The Geohazards Exploitation Platform portal**




It is providing the following set of capabilities, which are made available to the MARsite partners.

- **EO data discovery service** through a single point of access to visualize data collections in terms of acquisition footprints and sensor parameters, with resources available from ESA missions (especially the SAR missions from ENVISAT, ERS and SENTINEL-1) and third party missions (currently DLR TerraSAR-X and, upcoming for, ASI Cosmo-Skymed and CNES).

- **EO data access service over distributed repositories** supporting the dissemination of imagery either stored in the GEP cloud platform environment or accessed through the GEP portal in other remote data repositories from the pool of contributing agencies. Data access is based on the authentication of registered users and the granting of data dissemination according to the user profile. For instance, EO data constrained by license terms and distribution restrictions can be accessed from the platform's geographic interface via active links to the repository of the data provider.

- **Accounting service for EO data consumption** allowing the monitoring of the volumes of data use per EO source and according to the activity associated to the user profile. The accounting service can be used to support reporting concerning the exploitation of EO data, either from the Platform (e.g. hosted processing) or in the framework of application projects (e.g. CEOS pilots).

- **EO processing services** for on-demand processing, exploiting software to transform EO data into measurements; the user may run an EO processor provided on the platform (ready to use software-as-a-service, SaaS), or integrate an application he/she has developed (platform-as-a-service capabilities, or PaaS).

The SaaS Processing can be invoked either interactively though a web browser, or through scripting using the OGC Web Processing Service (WPS) interface; The PaaS provides software development and integration tools, and enables users to perform their data exploitation activities with large flexibility and autonomy, by using one or several virtual hosts directly provisioned on the cloud platform and deployable on demand.

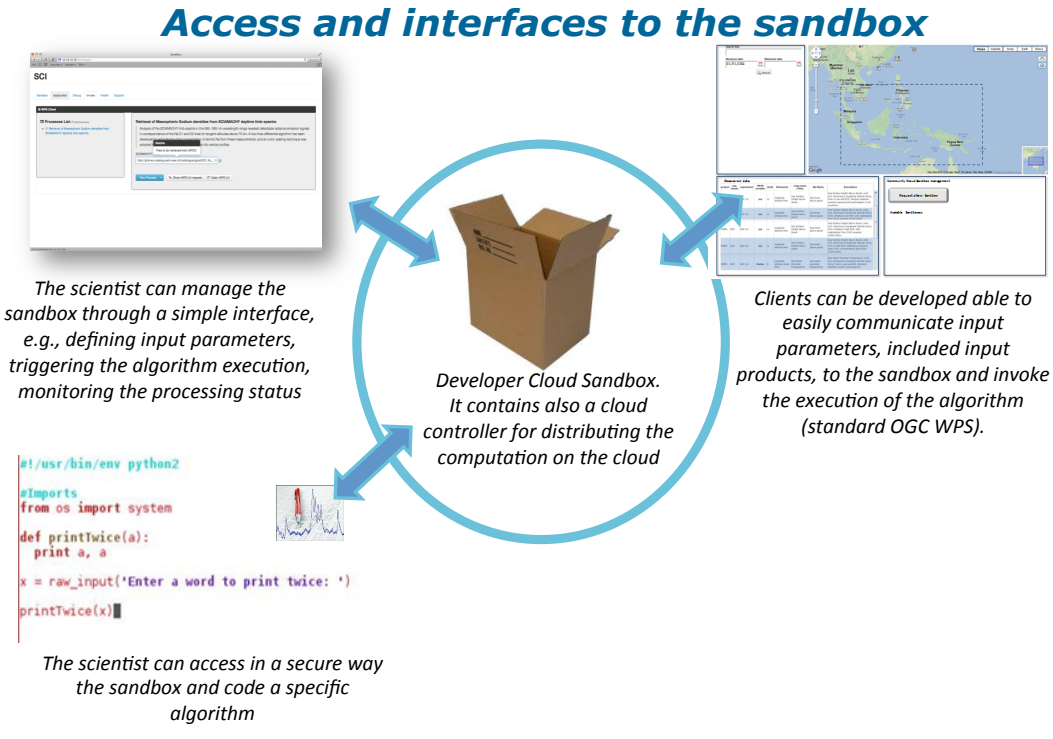
- **Access to Value-Added products** generated on the GEP, or products contributed by third parties. The platform allows cataloguing and dissemination of products relevant to the geohazards community. It can be used to provide access to elaborated products in support of thematic exploitation goals.

## 2.2. The Developer Cloud Sandbox computing model

The Developer Cloud Sandbox (Figure 2) is a computing model that allows scientists to develop and test EO services with:

- Directed Acyclic workflows;
- Parallel computing;
- OGC WPS submission interface, that makes it easy to:
  - Invoke the service from the MARsite portal;
  - Develop any custom web-client able to interface the WPS;
- Tools to query and access EO data;
- Output data published and if needed registered to GEOSS.

**Figure 2 - Access and interfaces to a Cloud Sandbox**

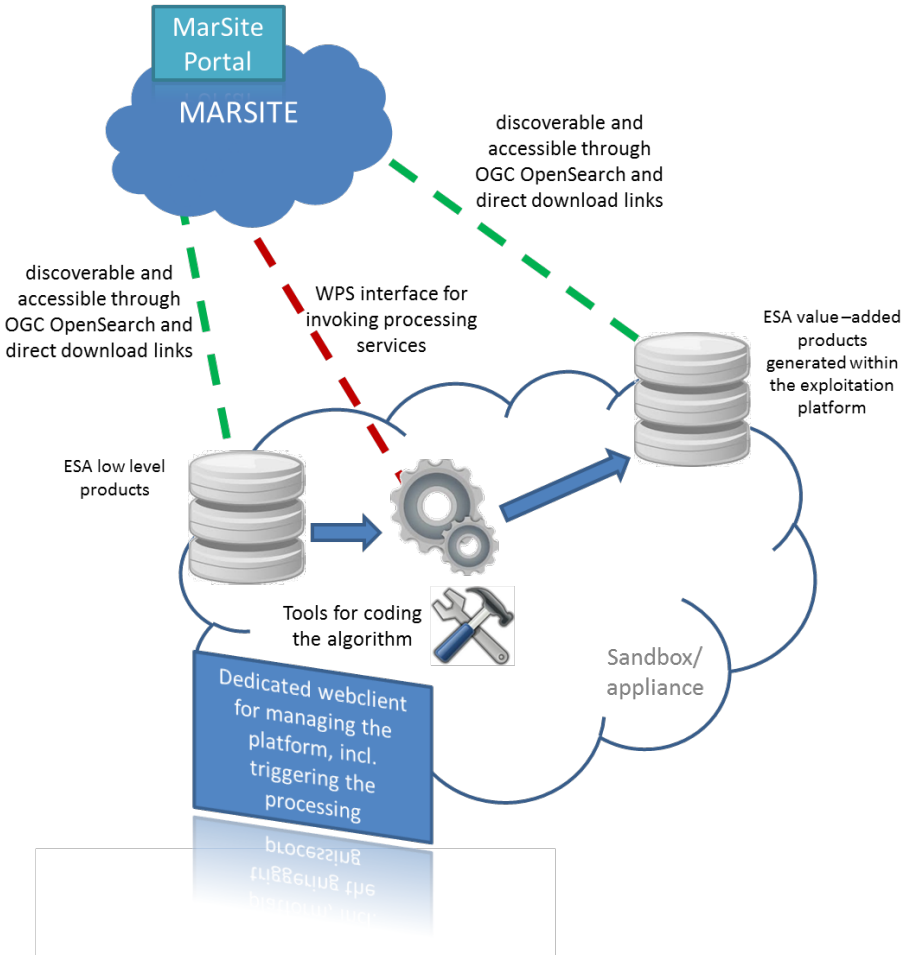


The Developer Cloud Sandbox is a computing model that allow scientists to develop and test EO services with Directed Acyclic Graph workflows, OGC WPS submission interface that makes it easy to develop any custom web-client able to interface the WPS, and tools to query and access EO data.

With this technology embedded, the GEP targets the evaluation of the different opportunities to move beyond EO data discovery bottlenecks and focuses on the computational science for data intensive analysis (Figure 3).

- Use of Infrastructure as a Service (IaaS) enables data and resource sharing, provides optimized costs and allows for a massively scalable ICT infrastructure.
- Adoption of pay-per-use models gives access to resources that users would not be able to afford on their own.
- Evaluation of innovative "EO application stores" and EO Software-as-a-Service (SaaS) concepts, along with options for sourcing of content (data) and applications (processors) from both open stores and commercial providers.

**Figure 3 - How GEP Cloud Sandbox fits in the bigger MARsite picture**



## 2.3. Interoperability: OGC OpenSearch and OGC WPS interfaces

GEP is making available best practices for search services, using the OGC OpenSearch interface standard with Geo, Time and EO extensions, as defined by CEOS, that allows standardized and harmonized access to EO catalogues from satellite EO data providers worldwide.

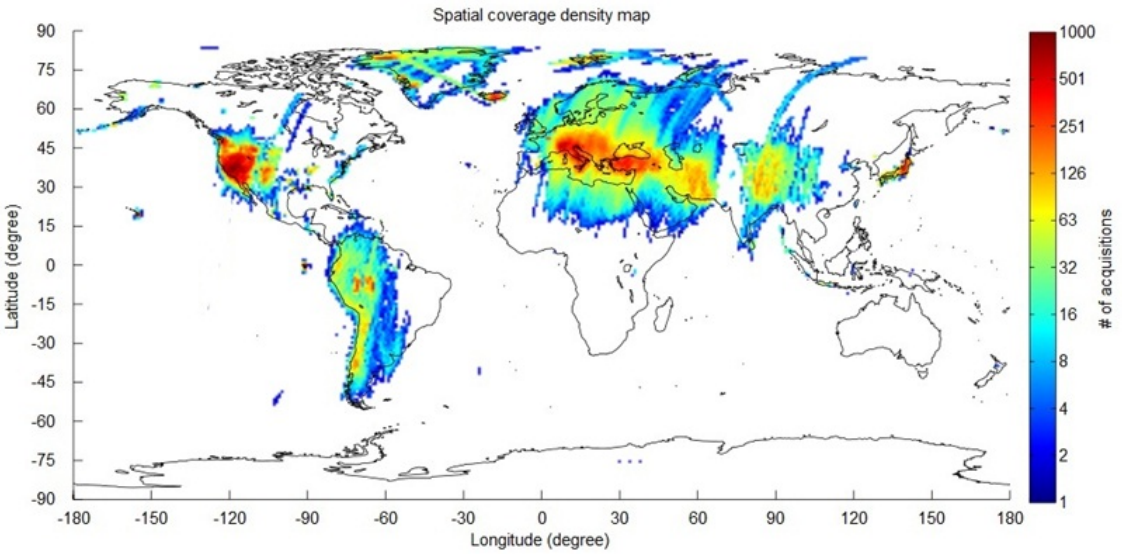
Starting in 2008, Terradue has pushed the OpenSearch approach towards its international standardization by proposing and editing the specifications from OGC. This interface was recognized and adopted in 2014 by the OGC as the preferred baseline for the catalogue services. ESA selected this interface for the implementation of the Agency's Next Generation User Services for EO (ngEO), NASA applied it to the newly deployed system of Earth Observing System (EOS) Clearing House (ECHO) and their interface to the Federation of Earth Science Information Partners (ESIP) that enables the discovery and access to the totality of NASA archives. The CEOS WGISS Integrated Catalog (CWIC) recently established a common interoperability best practice of OpenSearch in order to allow for standardized and harmonized access to metadata and data of CEOS agencies.

This track record of OGC OpenSearch implementation in the EO domain is strengthening interoperability for the whole geohazards community.

GEP is also putting into practice the OGC WPS interface standard, a web interface allowing for the dynamic invocation of user processes in a distributed computing environment. The OGC WPS interface handles the description of processing offerings, the triggering of processors with dynamic input parameters, and retrieval of processing outputs. Combined with encoding standards for processing results like OpenSearch description documents and/or IETF Metalink specification (for products download / retrieval), OGC WPS provides a simple and efficient solution to bind geoprocessing services with web applications.

Other notable interoperability protocols or encoding formats available from GEP are OGC WMS (with upcoming functionality of automatic layer publication after a processing job) and OGC OWS Context (for the exchange of data packages and other contextual information).

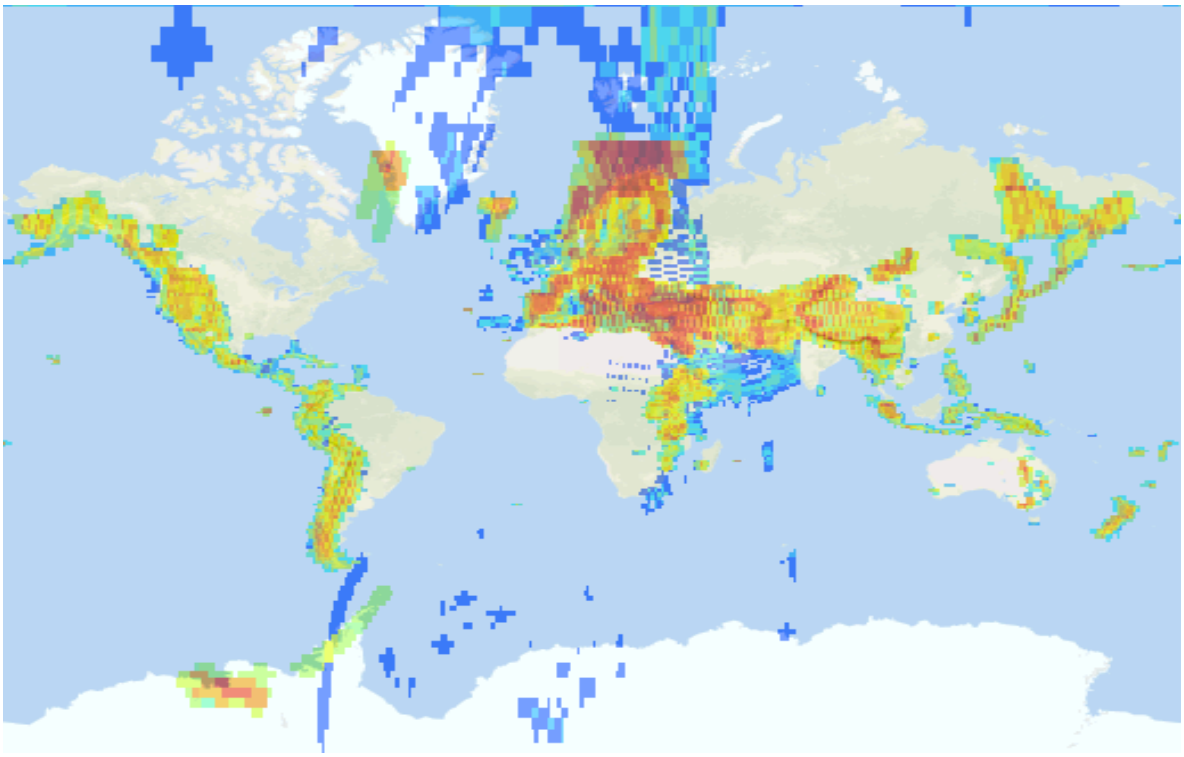
**Figure 4 - ERS SAR & ENVISAT ASAR data archive in GEP (as of March 2015)**




The GEP is also used to help users gradually access Sentinel-1 data as it flows from the Sentinel-1 SciHub (http://scihub.esa.int).

The following SAR data archives from SENTINEL-1 are already available via the GEP (in terms of number of acquisitions, see Figure 5).

**Figure 5 - SAR data archive in GEP from S-1 SciHub**



Regarding Sentinel-1 data, the assessment of data collections is done via the 'geobrowser' service of the GEP (http://geohazards-tep.eo.esa.int/geobrowser).

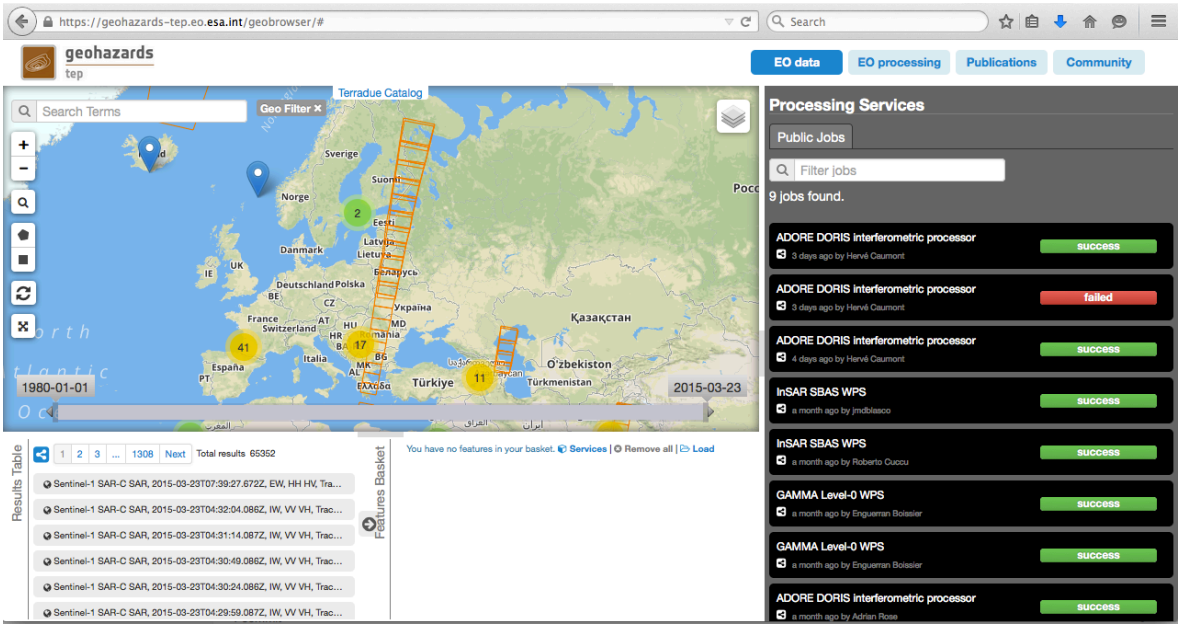
The data provisioning activity in GEP also intends to support access to other EO missions. Currently the DLR TerraSAR-X mission is accessible from the GEP geobrowser (the source catalogue being https://supersites.eoc.dlr.de/); ASI CosmoSkymed mission is upcoming.

The first step for the MARsite partners willing to work with GEP and to create new products, is to search, access and select from GEP the ESA low-level products of interest, and consider using the publicly-shared processing services as their primary production tools.

The platform is currently in its Validation phase, which will explore up to October 2015 a growing set of functionalities. As of today, the Platform already allows public data search over a large archive of EO data from ESA sensors and from third-party missions (CEOS partners) like DLR's TerraSAR-X.

Practitioners can access the GEP infrastructue through the Geobrowser service, which is now made available with a set of baseline functionalities to all users (unregistered users / general public) for data search, data selection and data processing at: http://geohazards-tep.eo.esa.int/geobrowser (cf. Figure 6).

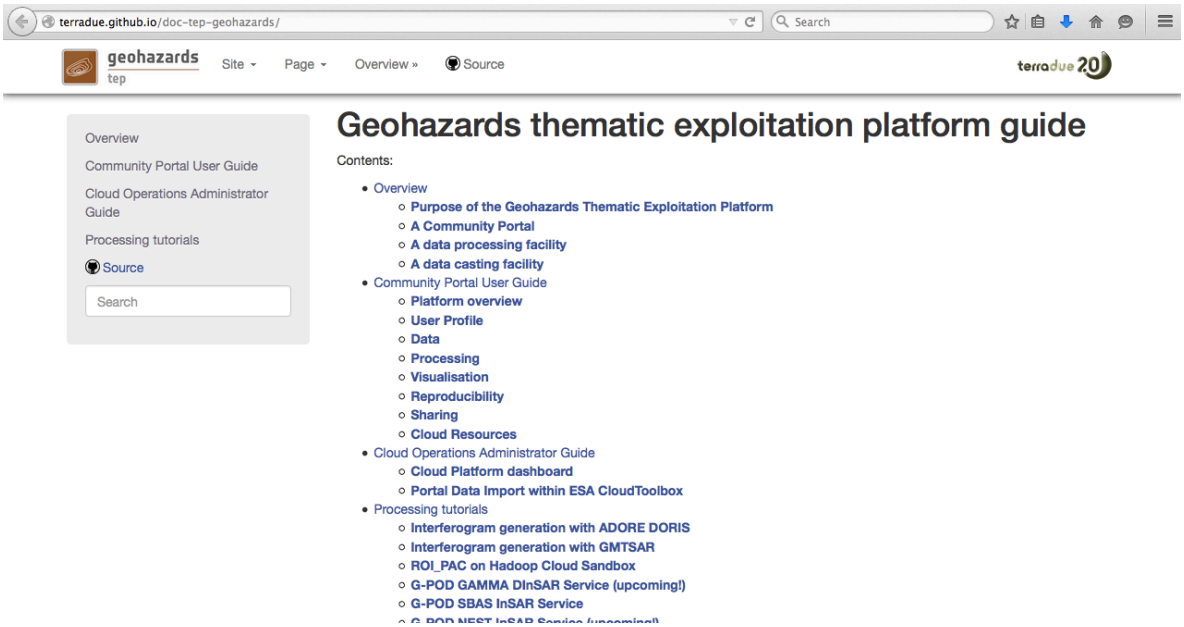
**Figure 6 - User access to the Geohazards Exploitation Platform**




The user documentation for GEP is also available online. It features an overview of the Platform concepts, a Community Portal User Guide, a Cloud Operations Administrator Guide and a growing set of data processing tutorials (SAR processing with ADORE DORIS, GMTSAR, ROI_PAC, and a set of G-POD services such as GAMMA-L0, SBAS).

This user documentation will continue to evolve in the coming months and it is available at:http://terradue.github.io/doc-tep-geohazards/ (cf. Figure 7).

**Figure 7 - User documentation for the Geohazards Exploitation Platform**




## 2.4. SAR processing with Hadoop Cloud Sandboxes as Cloud Appliances

As introduced before, registered users (currently, the early adopters that have applied for the GEP Validation phase activities) can access Cloud resources from GEP. One type of GEP cloud resources is the Developer Cloud Sandbox configured with the Hadoop framework enabling a massively parallel computation model.

The Developer Cloud Sandbox is a Virtual Machine (VM) that provides scientific developers with an Exploitation Platform-as-a-Service (PaaS). It consists of a development environment for processor integration and testing, and a framework for Cloud provisioning. The Developer Cloud Sandbox PaaS allows you to plug scientific applications written in a variety of languages (e.g. Java, C++, IDL, Python, R), then deploy, automate, manage and scale them in a very modular way. The algorithm integration is performed from within a dedicated Virtual Machine, running initially as a simulation environment (sandbox mode) that can readily scale to production (cluster mode). Accessed from a harmonized Shell environment, support tools also facilitate the data access and workflow management tasks.

Current outcomes from using the GEP Hadoop Cloud Sandbox (including for MARsite users) are the availability ‘as-a-Service’ of several SAR processors on the platform: ROI_PAC for co-seismic interforogram generation, ADORE DORIS for interferograms generation, PF-ASAR Level 0 to Level 1 Instrument Processing Facility, GMTSAR for interferogram stack generation. More processors are planned for being integrated in the coming month (DIAPASON, StaMPS, DLR’s InSAR-QL and so forth).

Once integrated and validated, the resulting Cloud Appliance is a computing model that allows scientists to execute EO services with the processing power of parallel computing (leveraging the Hadoop Framework), an OGC WPS submission interface that makes it easy to invoke the service from an Exploitation platform or from a custom web-client application, and to manage output data to be published externally and if needed registered into the GEOSS Common Infrastructure.

As indicated previously, the MARsite partners willing to integrate their own processing chain and make them available on GEP as-a-Service can also apply as early adopters, as part of the ongoing GEP validation phase.

## 2.5. EO software toolboxes integration as Cloud Appliances

The ESA CloudToolbox is a Virtual Machine that offers a flexible amount of CPUs, RAM and dedicated storage, tailored to a user needs and type of machine required. This provisioning is flexible, so that users can request upgrades of the configuration if needed (for example, asking more processing power), in compliance with the Cloud infrastructure constraints of resource types that are made available by Cloud providers.

A pre-built Virtual Machine template offers ready-to-use machines for SAR Interferometric processing or generic EO data processing. Moreover, besides the free and/or licensed software tools (e.g. Sentinel-1 toolbox, NEST, GAMMA, Matlab, etc.) that can be installed on the machines, users may also request installation of additional tools. A data access tool to ingest a 'data package' (the result of a search and select process via the GEP geobrowser) is also available from that environment.

To create a CloudToolbox, registered users simply access the GEP cloud dashboard (a user interface for cloud resources management), click to create a new Virtual Machine, set a name (e.g 'my InSAR toolbox'), select the "ESA CloudToolbox" template, click on create for the VM to be deployed, and then get the <ESA CloudToolbox IP> address. From there, users connected to the GEP Virtual Private Network can access the Virtual Machine through a VNC client.

As indicated previously, the MARsite partners willing to work and experiment with such cloud-based EO processing software and data access, can also apply as early adopters, as part of the ongoing GEP validation phase.