Gap analysis on open data interconnectivity for disaster risk research

Guoqing Li, Jing Zhao, Virginia Murray, Carol Song and Lianchong Zhang

ABSTRACT
Open data strategies are being adopted in disaster-related data particularly because of the need to provide information on global targets and indicators for implementation of the Sendai Framework for Disaster Risk Reduction 2015–2030. In all phases of disaster risk management including forecasting, emergency response and post-disaster reconstruction, the need for interconnected multidisciplinary open data for collaborative reporting as well as study and analysis are apparent, in order to determine disaster impact data in timely and reportable manner. The extraordinary progress in computing and information technology in the past decade, such as broad local and wide-area network connectivity (e.g. Internet), high-performance computing, service and cloud computing, big data methods and mobile devices, provides the technical foundation for connecting open data to support disaster risk research. A new generation of disaster data infrastructure based on interconnected open data is evolving rapidly. There are two levels in the conceptual model of Linked Open Data for Global Disaster Risk Research (LODG), which is the Committee on Data of the International Council for Science (ICRU): data characterization and data connection. In data characterization, the knowledge about disaster taxonomy and data dependency on disaster events requires specific scientific study as it aims to understand and present the correlation between specific disaster events and scientific data through the integration of literature analysis and semantic knowledge discovery. Data connection concepts deal with technical methods to connect distributed data resources identified by data characterization of disaster type. In the science community, interconnected open data for disaster risk impact assessment are beginning to influence how disaster data are shared, and this will need to extend data coverage and provide better ways of utilizing data across domains where innovation and integration are now necessarily needed.

1. Introduction
In February 2017, the UN General Assembly agreed on the definition of a disaster as

A serious disruption of the functioning of a community or a society at any scale due to hazardous events interacting with conditions of exposure, vulnerability and capacity, leading to one or more of the following human, material, economic and environmental losses and impacts. (UNISDR 2015)

This definition is qualified by the statements that the effect of the disaster can be immediate and localized, but is often widespread and could last for a long period of time; and that the effect may test or exceed the capacity of a community or a society to cope using its own resources, and therefore may require assistance from external sources, which could include neighboring jurisdictions, or those at the national or international levels.

Thus, disasters are serious events that bring damage, loss or destruction to populations and regions. Reports (Munich Re Group 2004) have indicated that more natural hazards have occurred in the past 60 years and that the economic and societal impact of disasters has increased by five times in the same period of time (Nichols et al. 2014). They are most often caused by hazards such as flood, hurricane, fire or earthquake, but as the Sendai Framework notes (UNISDR 2015), they can also be caused by man-made hazards. The Sendai Framework for Disaster Risk Reduction 2015–2030 aims "To strengthen technical and scientific capacity to capitalize on and consolidate existing knowledge and to develop and apply methodologies and models to assess disaster risks, vulnerabilities and exposure to all hazards" (UNISDR 2015), and achieve the outcome of “the substantial reduction of disaster risk and losses in lives, livelihoods and health and in the economic, physical, social, cultural and environmental assets of persons, businesses, communities and countries” (UNISDR 2015).

Research on disaster risk relies heavily on scientific data, including both observations and 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 and social media, are also beginning to play an important role in disaster risk studies. The need for open data and data interconnectivity, i.e. data which are accessible and usable by researchers, decision-makers and the public, is most critical in the area of disaster risk research, management and mitigation.

As stated in the Sendai Framework, disaster risk reduction requires a multidisciplinary approach to decision-making based on the open exchange and dissemination of disaggregated data. The work of the Linked Open Data for Global Disaster Risk Research (LODGD) is an increasingly important activity, endorsed by the Sendai Framework and the United Nations Office for Disaster Risk Reduction (UNISDR) Open-ended Intergovernmental Expert Working Group (OIEWG) on Indicators and Terminology Relating to Disaster Risk Reduction. On 2 February 2017, the UN General Assembly endorsed the 38 indicators proposed by OIEWG for monitoring the seven Global Targets of the Sendai Framework:

A. “Substantially reduce global disaster mortality by 2030, aiming to lower the average per 100,000 global mortality rate in the decade 2020–2030 compared to the period 2005–2015;

B. Substantially reduce the number of affected people globally by 2030, aiming to lower the average global figure per 100,000 in the decade 2020–2030 compared to the period 2005–2015;

C. Reduce direct disaster economic loss in relation to global gross domestic product (GDP) by 2030;

D. Substantially reduce disaster damage to critical infrastructure and disruption of basic services, among them health and educational facilities, including through developing their resilience by 2030;

E. Substantially increase the number of countries with national and local disaster risk reduction strategies by 2020;

F. Substantially enhance international cooperation to developing countries through adequate and sustainable support to complement their national actions for implementation of the present Framework by 2030;

G. Substantially increase the availability of and access to multi-hazard early warning systems and disaster risk information and assessments to people by 2030.” (UNISDR 2015)

To summarize, Targets A–D are outcome targets, and will require disaster loss and damage data; Targets E and G will require national self-assessment and Target F will relate to overseas development of system monitoring. These data are required for compliance with the national country reports driven by completion of the Sendai Framework Monitor. Such national voluntary reports will be submitted by all UN member states from 2018 onwards.

Disaster risk research demands open access to scientific data, because it is not possible to fully understand the cause and impact of a disaster event without collating multiple information resources. Nowadays, large amounts of disaster-related scientific data exist, such as data from monitoring equipment, base maps, evaluation, progress, socioeconomic statistics and so on. However, they are typically dispersed geographically and owned by various government agencies, research centers, private organizations and a wide variety of stakeholders including sometimes, individuals around the world. Researchers often find it difficult, if not impossible, to discover all relevant data needed for 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.

In-depth scientific analysis of disaster data requires management and acquisition patterns that support interconnection of dispersed scientific data related to disaster risk assessment and reduction. Disaster risk researchers face perhaps a greater challenge to find relevant datasets in a “sea” of distributed and disparate data resources. Therefore, gaps in data infrastructure, data sharing policies and data use governance should be addressed to unleash the potential of disaster risk research in helping regions, especially the developing countries, to improve risk assessment, reduction and management.

The analysis for this paper is generated from discussions and partnerships generated by LODGD Working Group of Committee on Data of the International Council for Science (CODATA).

Thus, this paper discusses two related areas: open data and data interconnectivity. It aims to identify the gaps in relevant policies and technology 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 some of the policies regarding data availability at various levels, and discusses potential
solutions and examples toward open data and data interconnectivity for disaster risk research.

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

2. Status of open data for disaster risk research

2.1. Issues related to open data

It is necessary to clearly understand the status of open data that is relevant to disaster risk research before we analyze the gaps toward a comprehensive and interconnected data infrastructure (Council 1995). Issues related to open data, the current state of open data in a number of scientific disciplines, mainly include data accessibility, data sharing and data interconnectivity (Haessler et al. 2014; Halevy, Norvig, and Pereira 2009).

Accessibility is the driving force toward the landscape of open science data. 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 “open-access” movement initially focused on research literature, and this concept has since been broadened to include scientific data, observational or derived, especially data that have been obtained through publicly funded research. In the past few years, the International Council for Science (ICSU) now the International Science Council (ISC) began to discuss access to data and information, and in particular, scientific data and information, as reflected in ICSU’s 2013–2017 Strategy Plan. At present, data accessibility always means timely access over computer networks and any implementation of open data access will have to support online, on-demand access methods, as well as access through web services and Application Programming Interfaces.

Data sharing is the practice of a data holder/owner making its data available to other users. There are two helpful examples:

(1) In the 2010 Beijing Declaration, the Group on Earth Observations (GEO) members committed themselves to implement the Global Earth Observation System of Systems (GEOSS) Data Sharing Principles by developing flexible policy that enabled a more open data environment (Rappaport 2011). This move has influenced national and regional data policies including INSPIRE directive (a European Union spatial data infrastructure), Copernicus (the European Union’s Earth Observation Programme) and the joint National Aeronautics and Space Administration (NASA) and United States Geological Survey (USGS) Landsat program which offers the longest continuous global record of the Earth’s surface in the United States (U.S.).

(2) The U.S. National Science Foundation began requiring a data management plan for all proposals submitted in 2011, and this has had a visible impact on investigators not only by their giving more thoughtful consideration to satisfy this requirement. Many now incorporate data sharing and accessibility into their proposals as a key component for dissemination of research outcome. This should accelerate in the coming years because cyberinfrastructure for data sharing becomes more mature and widely adopted in the U.S.

2.2. Challenges of disaster data

A key challenge in disaster risk research, whether analyzing the causal relationship of various impacts or modeling to predict the impact of future disaster events, is to make use of multiple data sources, and to synthesize and discover the underlying relationships. The academic community (Buckholtz and Takagi 2002; Cohen and Walsh 2007) 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). Internationally, current efforts include those of the LODG of CODATA, a research group dedicated to promote the linkage of disaster data. Experts in this group have formed a knowledge network and are researching the technology framework that will serve the disaster risk research community by connecting relevant open data from repositories at national, institutional and research group levels. Additionally, Integrated Research on Disaster Risk (IRDR) has established the Disaster Loss Data (DATA) Working Group. IRDR-DATA focuses its efforts on the issues related to the collection, storage and dissemination of disaster loss data. We envision a new kind of data infrastructure for disaster risk research that will connect disaster-related datasets of observations, analysis, statistics, etc., from multiple scientific disciplines as well as through citizen participation.

While recognizing the existence of a large number of datasets, a representative set of data are included from various disciplines ranging from earth observation (EO)
calls for, in its guiding principles, stating that and country. Therefore, open access can be challenging data always has sensitive information about the region enable research on longer-term issues. Since this type of toward a comprehensive framework for assessment and model losses and exposure data, should contribute Linking them with other data sources, including simula-
tion, hence making it difficult to use across regions and scientific domains. Weak or nonexistent meteorological data services (Egeland, Wildish, and Huang 2010) tend to be in very large with diverse formats in many countries, especially in developing countries, and this can hamper data usage in early warning and other risk reduction efforts and present barriers for multidisciplinary researchers (Zhang et al. 2015) and the public. Some countries and regions restrict the dissemination of seismic data because of the sensitive nature of their potential usage, resulting in earthquake risk information not being fully utilized (Steinberg and Rabinowitz 2003; Kvaerna, Ringdal, and Baadshaug 2007; Richards 2016). Restricted access is often placed on the geographic data because the diversity of data formats, the lack of standardized formats and the necessary expertise in geospatial analysis (Granell, Fernández, and Diaz 2014) and processing are all barriers to users. Health disaster loss data (Fleming et al. 2014; Juarez et al. 2014; McMichael and Haines 1997) are potentially a very rich source of information, but the use of population health data raises important issues around data governance, ownership and privacy. The involvement of multiple disciplines and communities in disaster-health research can also be complex to coordinate, particularly because they may have different approaches to data synthesis, analysis and interpretation.

From this brief reflection, disaster loss data are highly diverse in formats and access mechanism, often lacking in global standards and unified metadata schema. Linking them with other data sources, including simulation, model losses and exposure data, should contribute toward a comprehensive framework for assessment and enable research on longer-term issues. Since this type of data always has sensitive information about the region and country. Therefore, open access can be challenging in some countries even though the Sendai Framework calls for, in its guiding principles, stating that

Disaster risk reduction requires a multi-hazard approach and inclusive risk-informed decision-making based on the open exchange and dissemination of disaggregated data, including by sex, age and disability, as well as on easily accessible, up-to-date, comprehensible, science-based, non-sensitive risk information, complemented by traditional knowledge. (UNISDR 2015) (As shown in Table 1).

2.3. Status of open data in developing countries
It is considered that with the Sendai Framework, developing countries should draw on the technological progress to effectively construct and increase their local disaster risk reduction capacities, but low-quality infrastructure, lack of effective mechanisms and policies to manage risks and a lower capacity for societal mobilization are often attributed to the greater loss of human lives and properties. Disaster data sharing capacity builds on a nation’s economic development and scientific strength (Moore and Rajsekar 2010). As the developing countries have limited independent remote sensing satellites and cannot always easily respond to disasters in a timely and effectively manner, equal access to data should be sought. For example, the United Nations Economic Commission for Africa (UN ECA) holds that timely acquisition of remote sensing data as a powerful instrument to promote regional sustainable development (Date Revolution Group 2014; Lim et al. 2016). Although remote sensing technology would provide the same benefits to all countries (i.e. quality and frequency of data), the high cost of the technology remains a barrier for developing countries (Davies, Farhan, and Alonso 2013). The international community may consider starting to facilitate transparent sharing and transferring advanced technology from developed countries to developing countries through “data democracy”, as well as possible resources and funding to assist with data analysis. Worryingly, there is also 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 type (Cutter 2010).

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: (1) national coordination mechanism to fully use the internal data resources; (2) regional cooperation on disaster data sharing; and (3) international assistance mechanism data for disaster from NASA, USGS, JAXA, ESA, etc. (Guo et al. 2012), and a rapid response tool On-Site Operations Coordination Centre (OSICC) (https://vosocc.unocha.org/). This model of three-legged coordination and data sharing practice could be offered in partnership with other developing countries.

3. Scientific issues, key enabling technologies and case studies
3.1. Scientific issues in linking open disaster data
Hazards cause disasters include floods, droughts, heavy rain, snowstorms, earthquakes, typhoons, landslides, wildfires, infestations, etc., and cause loss of
| Types                  | Social and economic data                                                                 | Space based data                                                                 | Hydrological data                                                                 | Meteorological data                                                                 | Seismic data                                                                 | Geographic data                                                                 | Health data                                                                 | Disaster loss data                                                                 |
|------------------------|-------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|---------------------------------------------------------------------------------|---------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| Types                  | Economic development/Social structure/Culture/Ethics/Population/Socioeconomic statistics/Social media, etc. | Place-based (positioned) disaster loss data/Earth observation data. | Water level/Discharge/Sediment/Precipitation/Evaporation/Water quality. | Temperature/Air pressure/Moisture/Wind/Radiation, etc. | Comes from two areas: seismology and earthquake engineering. | Earth's measuring points of control/Residents and facilities/Transportation/Geomorphology, etc. | Diseases/Injuries/Risk Factors. | Casualties/Damaged buildings/GDP/Other economic losses. |
| Uses                   | Important background knowledge.                                                             | Provide accurate and near real-time Earth surface information.                  | Vital to disaster prevention and reduction, as well as management of water resources; Basis for evaluation. | Helpful for disaster prevention and reduction. | Build national earthquake early warning and reduction system; Could be used for sensitive purposes. | Critical to the design, management and risk mitigation of urban infrastructure such as municipal facilities, transportation, etc. | A useful source recognizing the Sendai framework Global Targets on mortality and morbidity; Provide important information. | Key for measuring the implementation of the Sendai Framework. |
| Relevant Policies/Attempts | Open Data Barometer: Global Report 2013 (involving more than 40 countries and regions). | The GEOSS 10-year Implementation Plan (endorsed by all GEO Members). | World Meteorological Organization (WMO) Resolution 40 (Kg-XII); Global Climate Observing System (GCOS). | China's National Seismic Science Data Sharing Centre. | Free access to the United States Geological Survey (USGS); A topographic database (NZTopo) from the Land Information New Zealand (LINZ). | The Medical and Environmental Data Mash-up Infrastructure project; An "Exposomics" project funded by the European Commission, etc. | Europe has worked to align efforts (e.g. Joint Research Centre); Develop standards; Open-ended intergovernmental expert working group (OIEWG); Other groups such as LODGD. |
| Current situation       | Not pen/Unstructured.                                                                      | Not widely available.                                                            | Long-term hydrological data series are very rare; Varied formats and data representation. | Weak or nonexistent data services; Inadequate data collection in developing countries; large and diverse formats. | Not fully utilized. Certain countries and regions restrict the dissemination of seismic data because of sensitive nature of their potential usage. | Diversity format and lack of standard format; Necessary expertise in geospatial analysis and processing; Restricted access. | The use of population health data raises important issues around data governance, ownership and privacy. Complex to coordinate in multiple disciplines and communities. | Diverse in formats and access mechanisms; Lacking a global standard and unified metadata schema; A comprehensive framework; Always has sensitive information. |
| Cross-domain/Unique Challenges | Cross-domain.                                                                            | Cross-domain.                                                                   | Cross-domain and Unique.                                                          | Cross-domain and Unique.                                                          | Cross-domain and Unique.                                                          | Cross-domain and Unique.                                                          | Cross-domain and Unique.                                                          | Cross-domain and Unique.                                                          |
life, damage and hardship. Multiple hazards may contribute to a same disaster event, e.g. landslides and flooding may occur at the time of an earthquake. Obviously, disaster risk research is multidisciplinary by nature and highly dependent on scientific data. 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 multisource data. For example, Dankers and Feyen utilized data on climate, geography, plantation and land cover in their simulation of flood disasters (Dankers and Feyen 2009; Dankers et al. 2014). Ulbrich utilized socioeconomic data, hydrologic data and observation station data to study the formation of precipitation and flood (Ulbrich et al. 2003a, 2003b). So, researchers studying disaster events often face a significant technical challenge of finding and accessing the relevant data for the events being studied among a wide variety of data sources.

The technical elements comprising of the data infrastructure include data management, data discovery, data interoperability and data services. These must work seamlessly together to manage the diverse, distributed, multidisciplinary disaster-related data and information and to serve the various stakeholders (decision makers, the public and researchers). Now, data infrastructure that hosts and serves data to disaster studies has been developed by various groups, projects, institutions and agencies. Efforts have been directed at establishing coordination among different data repositories within countries and across country boundaries to share data. However, corresponding to the technical elements above, the main technology gaps exist in sustained infrastructure for managing large amounts of diverse, heterogeneous datasets. These retain difficulties in discovering and accessing data across multiple repositories and disciplines, thus limiting its use across applications and disciplines, with poor interoperability of “cross-disciplinary” datasets for disaster risk research, vulnerability and potential loss analyses of large amounts of unstructured data, which is all a critical part of data services.

For the issue of data classification based on dependency relationship of disaster event-supporting data, 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 (Al-Hamdan et al. 2014), macro statistics, on-site investigation, to models and simulations. Data heterogeneity not only sits in the differences in measurement units, recording methods, description of data or metadata, and formats, but also in the diversity of systems for data storage, discovery and access. The second challenge is that different disciplines classify 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 risk research community and implemented in data management systems. Such a standard would not only consider the scientific classification of disaster data but also take into account the overlap among data from different disciplines, as well as effectiveness in practice in order to promote for wider adoption and usage. Developing an appropriate classification system holds the key to broader utilization of open disaster data and advance research to help people and governments in dealing with disasters.

As for the issue of the conflict between specialists, the public toward open access of disaster data, 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. More and more, citizens are collecting and contributing disaster information in addition to consuming data. The participants in disaster data-related activities now extend from scientists to the public (Fan, Li, and Zhang 2006). The broadening participation poses a major challenge in terms of both technological and policy aspects. Therefore, 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 coexistence of massive amounts of heterogeneous data present significant challenges in reconciling and utilizing the data. The full extent of issues, both scientific and technological, and how to involve the public as a key component in disaster data collection and usage should be studied.

With regard to the autonomy of disaster data resources, in the long term, it is necessary to grant it. At the moment, disaster data resources typically reside within the confines of agencies or institutions, often with technology implementations and policies.
specific to the hosting institutions. 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, on policy shifts toward open access. The autonomy of disaster data poses two main challenges: policy and technology. First, there needs to be an appropriate framework to define policies related to access, rights, credit, etc. Such policies will be integrated into 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 needs to be based on existing data standards and embraced by the community. To access data via the heterogeneous data sharing system, one can selectively get the metadata 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.

Since the Organization 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 adopt legal frameworks and implement policy initiatives to encourage greater openness in science (Organisation for Economic Co-operation and Development 2015a). Policies on open access to scientific data would take the form of mandatory rules, infrastructure or incentives (Organisation for Economic Co-operation and Development 2015b). 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 prepublication inputs and postpublication outcomes including data. Also, with the development of metadata, data cleaning and curation require a substantial amount of resources, and this is not acknowledged adequately in evaluation mechanisms or grant allocation procedures. Therefore, this issue could be addressed to a certain extent by extending citation mechanisms to datasets, while a significant challenge in policy making is to incorporate these incentives for scientists into practical measures and instruments for encouraging sharing and providing access to data.

Additionally, open data access needs protection and is guaranteed by laws and policies, including information security and privacy. In the developed countries, laws of data access and data protection have been well established, while these are lacking in most developing countries. Besides, it has also been recognized that new issues may arise as a result of such laws and policies. For example, a dataset, such as medical records and statistics, or a social media communication, that was acquired originally for a specific purpose or by one type of application, could not be used in a different domain and disseminated to a completely different audience. If legally approved, data may be used in applications beyond the original intended use. This expanded scope of use may have implications related to ownership, privacy and ethics. Laws and regulations to protect scope of use and intellectual property are needed to ensure appropriate use and control of data. 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 (Okulov et al. 2007), while in some cultures, 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. Furthermore, the level of awareness toward citizens’ right to information varies among countries of different cultural backgrounds.

3.2. 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 risk researchers, as described in Section 3.1. 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. In 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. The 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, tools and services (Atkins 2003a, 2003b). 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. Several key areas of cyberinfrastructure relevant to an interconnected disaster data infrastructure are highlighted here as follows:

(1) **Networking and data movement**: Network connectivity has become ubiquitous thanks to the increased connectivity among institutions, academic campuses, government agencies, and even homes and personal communication devices. In the context of a data-centric cyberinfrastructure, the networking layer provides the foundation for data accessibility. Increasing connectivity and bandwidth (e.g. Science DMZ), as well as the availability of reliable and easy-to-use software to support the transfer of large scientific data, is critical to sharing distributed data sources.

(2) **Advanced computing**: Significant computational power is critical to meet the needs of modeling and simulation, data analysis and visualization at varying scales and dealing with large amounts of data of different formats and sources. Different computing paradigms such as parallel and distributed computing can be used for problems in the disaster risk research domain where large-scale resources are needed to compute at high resolutions. Many established models of operating computing resources, ranging from academic institutions, national computing infrastructure, to commercial clouds, should be considered based on the requirements of the disaster data infrastructure.

(3) **Data-intensive computing**: Computation in the domain of disaster risk research is typically data-centric. Whether researchers are modeling 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 (at terabytes (TB) level 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. Innovations in system architecture and hardware are extending the capabilities of supercomputers to better support data-intensive sciences, including massive high-performance data storage, large-scale Flash memory capable of reading and writing files at TB level 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. Leveraging the new technology will help address the challenges of delivering time-sensitive results of modeling and analysis to disaster management decision makers.

(4) **Cloud computing**: 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, Google Cloud Engine, Microsoft Office 365 and Apple iCloud apps). Five characteristics are considered essential in cloud computing by National Institute of Standards and Technology (NIST): On-demand self-service, network accessibility, resource elasticity, resource pooling and metered service. Many research groups and even institutions have moved their computing to the cloud. Users 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. The computational needs of the disaster risk research and mitigation domain encompass a wide range of applications, from modeling, data processing, data synthesis, data analysis, to visualization and decision support. Most of these can be well served by cloud computing service models.

(5) **Service-oriented architecture (SOA) and data services**: 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. Web services support SOA implementations over the Internet. They make functions, or services, accessible over standard Internet protocols across different systems, frameworks and programming languages. Many data repositories provide web services for data access and metadata information, although common protocols of data services are needed to address the specific challenges associated with scientific data from multiple disciplines for disaster risk research.

(6) **Data science:** 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 modeling, data visualization and many others. The focus on handling “big data” is rapidly increasing our ability to understand and gain insights into data at a scale that was impossible before. The growth of data science will benefit many fields, especially in risk and disaster management.

### 3.3. Case studies and lessons

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

(1) 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.

(2) In October of 2005, the ICSU 28th General Assembly agreed on the launch of IRDR. IRDR aims to address the challenge of natural and human-induced environmental hazards. The DATA group of IRDR is committed to establishing disaster loss data infrastructure for stakeholders.

(3) On 16 December 2006, the UN 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.

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

(5) 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.

(6) 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 (Lengert, Popp, and Gleyzes 2012; Rowan et al. 2013; Amelung and Lengert 2009).

(7) 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).

(8) In midle of 2014, Committee on Earth Observation Satellite (CEOS) proposed the idea of Recovery Observatory Infrastructure to further promote disaster data sharing and initiated three pilots (Flood, Seismic hazards and Volcano oriented) to demonstrate how satellite EO data can be used to improve monitoring and response to hazardous events, when accessing data at no-cost.

(9) The On-Site Operation Coordination Centre (OSOCC) originally developed by United Nations’ Office for the Coordination of Humanitarian Affairs (OCHA) aimed at improving the international assistance in support of the Governments of affected areas. In the case of emergency, OSOCC makes available a great number of data useful for making decisions about response and humanitarian assistance.

(10) Non-profit and citizen-driven projects have also been created to aid relief efforts using open data and technology, such as the Humanitarian OpenStreetMap Team (HOT), making free, up-to-date maps available, providing a critical resource when relief organizations are responding to disasters or political crises.

In the following section, we describe three examples of these initiatives in more detail to highlight the
efforts and showcase the specific actions in making them successful.

(1) Spatial Hazard Events and Losses Database for the United States (SHELDUS*, http://www.sheldus.org/) is a U.S. county-level hazard dataset covering 18 different hazard types and a time span from 1960 to present. The data originates from several existing national data sources such as the National Climatic Data Centre’s monthly storm data publications and the National Centers for Environmental Information’s (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. At its finest resolution (i.e., without any data aggregation), SHELDUS* outputs the beginning 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. The SHELDUS* database is widely used, particularly among researchers and planners. Research and planning represent the main usage areas of SHELDUS*. For example, as of May 2015, there are more than 85 peer-reviewed publications and 50 thesis/dissertations that heavily utilize SHELDUS* data. The topics of research range from resilience models for the tourism and hospitality industry to statistical analysis 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. Besides, 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 in the US.

(2) The primary goal of 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 HDDS developed by USGS provides quick and easy access to the remotely sensed imagery and geospatial datasets that are essential for emergency response and recovery operations (Bewley 2011; Lamb and Jones 2012; Jones and Lamb 2010, 2013). The HDDS (http://hddsexplorer.usgs.gov/), as a disaster response system, incorporates satellite tasking and data acquisition, product development, Web applications and data storage. Currently, the 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 248 TB (7,761,504 files), which provides the 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 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.

(3) ChinaGEOSS Disaster Data Response Mechanism (CDDRM), which is established by ChinaGEOSS Data Sharing Net in 2016, is responsible for coordinating Chinese high-resolution satellite images and disaster analysis products to affected countries (Zhang et al. 2018). The response process includes data acquisition, data sharing, data application and data publishing. At the data acquisition stage, diverse instant messaging technology is used by operation office to publish data requirements and update feedback information. At the data sharing stage, multiple pre- and post-disaster high-resolution satellite images are distributed to service nodes in America, Europe and Asia. Based on the cloud storage technology, scientists from different countries and institutions can query metadata through a unified portal and select the best network connectivity node to access data. It helps them rapidly map injuries, damages and identify possible safe areas. At the data publishing stage, all of the sharing data are published in the form of a data journal to protect the data provider’s intellectual property right. As of March 2018, the CDDRM has responded four times for major natural disaster events around the world (e.g. the New Zealand earthquake in November 2016, the Mexico earthquake in
September 2017, the Iraq–Iran earthquake in November 2017 and the Cyclone Gita in February 2018 (as shown in Figure 1). More than 7 agencies and 12 satellites in China have been involved in the past activities. As a new nongovernmental disaster cooperation mechanism, ChinaGEOSS Data Sharing Net has also established the cooperative relationship with GEO, UNOSAT, UNESCAP, IRDR, CODATA, DBAR and so on. It has been regarded as a complement of the intergovernmental disaster reduction under the Sendai framework.

4. Conclusions

What can the scientific 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 the impact of disasters on lives and properties? For these purposes, the following are recommended:

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 risk research.

2. Organizations like the 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 efforts to ensure that these standards meet the needs of disaster risk 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 EOs to include other types of data, especially the economic, population, public health, infrastructure, 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 an easy access to highly usable, distributed, multidisciplinary datasets for disaster mitigation stakeholders and applications on a 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 the 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 risk research and reduction.

(8) Review and where appropriate consider establishing a series of exemplary pilot projects on applications of cross-disciplinary approaches and use of data for studying disaster risk reduction. This could involve academic experiments, datasets, infrastructure, test beds 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.

One way to heed the call of the Sendai Framework for greater science use for understanding risk is to build strong links between such UN member states National focal points or 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 Science Council (ICS) and the United Nations Office for Disaster Risk Reduction (UNISDR), aims to serve as that link to bring more science and data-driven approaches to disaster risk management in partnership with the LODGD Working Group of CODATA.

Acknowledgments

The authors wish to express their most sincere appreciation to colleagues in the broader academic community for their thoughtful comments and suggestions on the previous versions of this paper. Special thanks to Drs Philippe Bally and Theodora Papadopoulou from the European Space Agency, and Dr Fang Chen from the Institute of Remote Sensing and Digital Earth (RADI) at Chinese Academy of Sciences, for their thorough review of the paper and constructive suggestions; and to Dr Simon Hodson, Executive Director of CODATA, for his guidance and support for the activities of the LODGD Task Group and our outreach to the broader community; and the help from Integrated Research on Disaster Risk (IRDR).

This work was written on the basis of LODGD White Paper 1, so the authors are thankful to the contribution from the LODGD members: Edward T. H. Chu, Susan L. Cutter, Nuha Eltinay, Bapon Fakhruddin, Jinlong Fan, Mingrui Huang, Shifeng Huang, Brenda K. Jones, Xiaotao Li, Rishma Maini, Xiuling Qing, Michael Rast, Xinlu Xie, Masaru Yarime, Hongyue Zhang and Siyka Zlatanova.

Funding

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences [grant number XDA19020201].

Notes on contributors

Guoqing Li is a professor and the head of Satellite Data Technology Division, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences (CAS). His work focuses on high-performance geo-computation, big earth data management and spaceborne disaster mitigation.

Jing Zhao is a PhD candidate of Satellite Data Technology Division, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences (CAS). Her work focuses on global change specially related to the impact of human activity on climate and disasters, and on spatial distribution of socioeconomic data with a high precision.

Virginia Murray is a professor and the head of Global Disaster Risk Reduction, Public Health England. She is a medical doctor and her work focuses on the health emergencies and disasters risk management and links to the implementation of the 2015 UN Landmark agreement of the Sendai Framework for Disaster Risk Reduction 2015–2030.

Carol Song is a senior scientist and director of Scientific Solutions at the Rosen Center for Advanced Computing, Purdue University. Her work focuses on connecting multi-disciplinary science with advanced cyberinfrastructure to accelerate discovery and innovation. Her current research areas include high-performance and data-driven computing, advanced cyberinfrastructure and geospatial data infrastructure.

Lianchong Zhang is a PhD candidate of Satellite Data Technology Division, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences (CAS). His work focuses on Earth Observation data integrating, management and sharing.

ORCID

Guoqing Li http://orcid.org/0000-0003-0654-5426
Jing Zhao http://orcid.org/0000-0002-7838-9832
Virginia Murray http://orcid.org/0000-0002-1423-9990
Carol Song http://orcid.org/0000-0003-0123-9067
Lianchong Zhang http://orcid.org/0000-0001-5902-7126

References

Al-Hamdan, M. Z., W. L. Crosson, S. A. Economou, M. G. Estes, S. M. Estes, S. N. Hemmings, S. T. Kent, et al. 2014. “Environmental Public Health Applications Using Remotely Sensed Data.” Geocarto International 29 (1): 85–98. doi:10.1080/10106049.2012.715209.

Amelung, F., and W. Lengert. 2009. “Supersites: An Initiative Towards Open Access for InSAR Data.”
