

Information Integration and Disaster Data Management (DisDM) ^{*†}

Felix Naumann Louiqa Raschid
HPI, University of Potsdam University of Maryland
naumann@hpi.uni-potsdam.de louiqa@umiacs.umd.edu

Recent disasters such as the 2003 SARS outbreak, the 2004 Asian tsunami, the 2005 Kashmir/Pakistan earthquake and 2005 hurricanes Katrina and Rita clearly identified the shortcomings of IT solutions for disaster rescue and recovery. Of particular note was the *accompanying disaster* of the lack of viable or deployed data management solutions, specifically information integration and information sharing solutions. In this position paper, we identify some specific challenges to information integration and sharing that is driven by the requirements of disaster data management (DisDM). We then identify a brief summary of needed technologies and methods for DisDM.

DisDM Challenges

All major disasters affect the lives and the welfare of many individuals; in many other aspects they differ significantly. These differences include the demographics of the affected people and the affected area (local, such as a hurricane; national, such as the 2005 earthquake; regional, such as the 2004 tsunami; global, such as the 2003 SARS outbreak and the predicted H5N1 avian flu pandemic). The time dimensions are also quite different; an earthquake strikes within a few seconds, while an epidemic such as the bird flu spreads much more slowly. The cause of the disaster can be man-made, such as terrorist attacks, or they can have natural causes, or both, such as the 2005 flooding of New Orleans. Disasters are dealt with in several phases including immediate response and rescue, and short-term and long-term relief and rehabilitation. Additional phases may include long-term tracking of people and places (post disaster) and planning or prevention or assessment phases (pre and post disaster). Next, we briefly enumerate the data management and data integration requirements of DisDM.

Diversity: Major disasters involve multiple autonomous organizations (governmental, NGOs, individuals, communities, and industry). This leads to a diversity of client interests, compounded by the fact that many of the data sources and data consumers are neither IT specialists nor have experience in data management. Most important is the need to both exchange and disseminate information.

Heterogeneity: The data to be integrated and disseminated is extremely heterogeneous, both structurally and semantically. This is because data from a large number of different domains is needed to successfully perform the DisDM mission. There is data about persons—both about victims and about relief personnel; data about damages to buildings, infrastructure and belongings; weather data; geographical data about roads and other landmarks; logistics data about vehicles, delivery times, etc.; communication and message data; financial data needed to manage the collection and distribution of donations; data in blogs; etc.

Data Quality: To further complicate the technical challenges raised by the heterogeneity of data, data quality is critical in DisDM scenarios. Poor data quality can squander resources and can even be life threatening.

On-the-fly Integration: Finally, there is a need for on-the-fly integration. It is difficult to plan for all possible disasters and potential participants cannot be predicted. The participants also change over the course of the disaster and the phases of disaster response. While there is the potential to model past disasters and to learn information integration solutions, there is a clear need for a flexible platform.

*Partially supported by the National Science Foundation Grant IIS0533986.

†Thanks to Monica Scannapieco and Snehal Thakkar for their feedback.

To summarize, DisDM relies critically on data management and data integration and at the same time DisDM presents new challenges.

Data Integration Approaches to DisDM Challenges

We have compiled a list of needed technologies and methods for DisDM solutions. We note that this is not meant to be exhaustive. We briefly identify them within four task areas, showing what exists and what needs to be developed.

Data integration: The particular DisDM challenge for data integration is twofold. First, the ad-hoc nature of DisDM makes automated data integration a necessity. Second, the diversity of data sources and user interfaces and their volatility suggest a virtual integration solution as opposed to a central materialized warehouse. In both areas of automation and virtual integration, research has made much progress. For instance, schema matching techniques semi-automatically align heterogeneous schemata and graphical interfaces allow easy schema mapping. The area of virtual integration has made much progress in query answering (Local-as-View, Global-as-View, etc.) and query optimization. DisDM will be a challenging application domain in which we can apply existing data integration solutions. For example, bioinformatics applications do consider a diversity of data, but they typically do not have to discover data resources on-the-fly, or to deal with the volatility of clients in the DisDM environment.

Flexible architecture: An information architecture for DisDM must be able to deal with highly volatile data sources, disrupted network connections and frequent changes to individual sources as experts adapt the information systems to the demands of the particular disaster. Schema mappings have proven to be a simple yet effective way of connecting to heterogeneous sources; they are easily (and visually) manageable by an expert and can automatically be interpreted as GAV/LAV-views for data import and transformation. Peer-data-management-systems (PDMS) apply many ideas of P2P file-sharing networks to the domain of distributed databases. PDMS are especially robust in volatile environments, they are simple to set up, simple to join (using schema mappings) and need no centralized maintenance. Finally, all data integration architectures and solutions will need to consider data security and privacy.

Dissemination: The dissemination of relevant information to interested parties is particularly difficult due to two factors. First, there is usually not enough time in the aftermath of a disaster to develop sophisticated applications on top of data sources. Thus, data can often only be acquired through a query language (not suitable for lay users) or through keyword search (not suitable for complex data demands). Research in profiles, wide area monitoring, Publish/Subscribe and HCI must all come together to address the data delivery needs for DisDM.

Quality: The quality of data is especially relevant and especially vulnerable in DisDM. Approaches to cope with poor data quality can be broadly classified into two parts: approaches to detect and assess poor quality and approaches to improve data quality. The former is relevant for users or applications accessing autonomous sources where the data itself may be of poor quality but cannot easily be updated or improved. In such cases IQ assessment methods and methods of quality-driven query planning can be applied. Technologies for the latter approach to improve data quality include error and format correction, outlier detection, duplicate elimination and data fusion. It is expected that in a DisDM scenario there may be multiple sources and sensors that may provide information relevant to a particular event, e.g., the condition of some road may be reported from aerial surveys and people on the ground. Intelligent merging that makes use of geographic or other models could lead to improved quality.

To summarize, data management and in particular information integration and information sharing, has a critical role to play in DisDM. While there is much existing research ideas that can be applied, it is clear that DisDM will challenge many of the assumptions made, and in many cases introduce new complexities. An overall challenge is to develop workable solutions that are adaptive, flexible, and customizable, that can be quickly deployed with some minimum basic features, and that can evolve with the phases of the disaster. As a direct response to the Asian tsunami, Sahana, an open source disaster management system is under development.

(http://en.wikipedia.org/wiki/Sahana_FOSS_Disaster_Management_System). We encourage database researchers to contribute their expertise to the many information management challenges of the Sahana project.