LESSONS LEARNED FROM DATA MANAGEMENT ACTIVITIES AFTER GREAT EAST JAPAN EARTHQUAKE IN MARCH 2011

A. Kitamoto[*1, 2]

*[*1]Digital Content and Media Sciences Research Division, National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan*
Email: kitamoto@nii.ac.jp
[2]PRESTO Project Researcher, Japan Science and Technology Agency

ABSTRACT

This paper summarizes our effort towards managing the multi-disciplinary disaster-related data from the Great East Japan Earthquake, which happened on March 11, 2011 off the coast of northeast Japan. This earthquake caused the largest tsunami in the recorded history of Japan, killed many people along the coast, and caused a nuclear disaster in Fukushima, which continues to affect a large area of Japan. Just after the earthquake, we started crisis response data management activities to provide useful information for supporting disaster response and recovery. This paper introduces the various types of datasets we made from the viewpoint of data management processing, and drew lessons from our post-disaster activities.

Keywords: Disaster response, Data management, Meteorological data, Radiation data, Geo-tagging, visualization

1 INTRODUCTION

The Great East Japan Earthquake on March 11, 2011 off the coast of northeast Japan caused disasters that extremely impacted on Japan. As a response to the disaster, we decided to start a crisis response to provide various types of data on a website for people in need of emergency response and/or assistance. This was challenging because the variety and amount of data were unprecedented in terms of the number of disciplines involved, and this is especially true for the data from the Fukushima Daiichi nuclear power plant accident. General descriptions of the voluntary projects are already described elsewhere, such as (Utani 2011), so we focus on our own post-disaster activities, and draw lessons from our experiences. As of January 2012, we have produced the following datasets and services on our website (http://goo.gl/9knll).

1) Mass media online news about the earthquake and related events.
2) Meteorological observations and numerical data around Fukushima Daiichi nuclear power plant.
3) Radiation monitoring network data around nuclear-related facilities.
4) Radiation measurements on all schools in Fukushima prefecture.
5) Timelines of major events based on radiation measurements and accidents at the Fukushima plant.
6) Geo-tagging of mass media online news for the spatial understanding of events.
7) Comprehensive power plant database with aesthetically appealing visualization.

These datasets can be classified by using several facets, such as 1) real-time or archiving, 2) textual or numeric, 3) geographic or temporal, 4) alerting or informational, but we focus on the aspect of the data management process and classify it into several steps: 1) data collection, 2) data grounding, 3) data integration and analysis, 4) data visualization, and 5) data dissemination. The following sections summarize our experiences using the data management process for disaster response after 2011 Great East Japan Earthquake.

2 DATA COLLECTION

Datasets were collected from various Internet sources. Some were available before the disaster, and others were released after the disaster. Here, the challenge was to collect and integrate disparate datasets into a single database so that the data could be easily manipulated and visualized for multiple purposes.

The collection of data was a difficult task, not only because of the scattered sources of data, but also because of the non-uniform data formats defined for each data source. Machine-readable formats, such as XML (extensible markup language) and CSV (comma-separated values) are ideal formats, but those formats were relatively rare. The usage of standard metadata formats, which allows automatic integration of data from many sources, was far from reality. Instead, data were released in PDF (portable document format), where scraping is difficult to automate, scanned PDF, where OCR (optical character recognition) produced poor results, or encrypted PDF, where automatic processing was almost impossible. It seems that some of the organizations wanted to control the usage of their data, and were reluctant to let people use the data in their own manner. People like us, however, are highly motivated for digitizing data against difficulties to obtain a better understanding of the data.

The most reliable way for digitizing data was to manually type the data by reading from PDFs or other types of documents. Thanks to the availability of cloud-based applications such as the Google spreadsheet, even a very time-consuming task could be finished by the collaboration of volunteers. For example, a group of volunteers started the radmonitor311 project (http://sites.google.com/site/radmonitor311/) to collect radiation monitoring data from various websites. They used the Google spreadsheet to collect data so that many people can work at the same time and share data easily through the Google spreadsheet API (application programming interface). The usage of cloud-based applications was an important step for aggregating the power of motivated people.

It is important to note that the release of data in a machine readable format is an important step toward realizing an “open government,” which is now regarded as an important agenda in such countries as the United States and United Kingdom. Here, the role of government is to provide data, and the usage of this data can be left as the work for people. The advantage of an open government policy is that better visualization and analysis tools could be produced at a more reasonable cost thanks to the power of volunteers with diverse skills. Disaster response is when this type of policy yields the greatest value, because the need for information and motivation for contribution are at maximum.

3 DATA GROUNDING

Data grounding refers to mapping data on real coordinates such as space and time, which are the most basic facets of disaster-related data. Mapping to space means projecting data on a map, while mapping to time means projecting data on a timeline. When the space and time are represented in a textual representation, we need to convert textual representation into numerical coordinates. An example is the geocoding or geo-tagging of data from place names to latitude and longitude.

When datasets only contain place names without postal addresses or geographic coordinates, geocoding can be performed using the following two steps: 1) place names to addresses and 2) addresses to geographic coordinates. For the first step, we used Internet search engines to look up the postal addresses of places. For the second step, we used an Internet geocoding service to convert a postal address to its given latitude and longitude coordinates. Sometimes the location is only available as a point on a scanned map, or the location is given in a descriptive form (such as 100 m from the park along the road), but in these cases we have to rely on the manual process of geocoding to improve the reliability of points through the interpretation of information from multiple sources.

A more difficult case is the grounding of natural language text, in contrast to the geocoding of a postal address, which is well-structured. This process also consists of two steps: 1) extracting place names from the text, and 2) disambiguating them from the candidates. However, we do not have a simple way to segment place names from other words especially in Japanese where there are no delimiters for a word boundary. Words are also ambiguous, especially when we have two place names with the same spelling. We need additional information to resolve multiple candidates, but text has insufficient information to do this.

To solve this problem, we are developing a piece of software called “GeoNLP” (http://goo.gl/5Jq1T), whose task is roughly described as two steps: 1) toponym recognition and 2) toponym resolution, following the classification made by Leidner (2007). The first step deals with the recognition of a text span that constitutes a toponym, which is a special case of generally named entity recognition. This process depends on a dictionary of place names, or gazetteer. The second step deals with the selection of the correct referent among all the candidate referents (possible locations). We applied several types of heuristics to pick the best referent from among the given candidates, such as the proximity of multiple place names or the similarity of the place name category appearing in the same sentence. GeoNLP was applied to the grounding of mass media online news to categorize the online news articles by their place.

4 DATA INTEGRATION AND ANALYSIS

The data from multiple sources can be compared on a uniform coordinate after data grounding. The task here is to compare different types of data to get a better understanding of the whole picture. The first task was to integrate multiple timelines to interpret the dispersion of radioactive materials. The following four datasets were integrated: 1) timeline of radiation measurement events from radiation monitoring of Fukushima prefecture since March 15, 2) radiation monitoring for all schools in Fukushima at the beginning of April, 3) timeline of accidental events reported from Nuclear and Industrial Safety Agency since March 11, 4) meteorological simulation data from Japan Meteorological Agency (JMA) for wind and rain every hour since March 11. These four types of data differ in type, scale, and time, but a careful comparison of those datasets helped us understand the mechanism of dispersion for Eastern Japan. We focused on the major release events on March 15 and 21. We found out that the event on March 15 was especially difficult to analyze because of the complex constantly changing wind patterns on that day. However, we identified three phases of dispersion, namely a plume moving to the south, a plume moving to the west, and a plume moving to the northwest. Our interpretation was difficult to validate, however, because there was insufficient observation data in terms of space and time. We suggest that the analysis of the data should be backed up by using a carefully designed simulation study.

The second task was to analyze the relationship between the radiation level and total rainfall on March 15 using the Fukushima school monitoring and weather radar data. This analysis was based on the hypothesis that more rainfall brought more radioactive materials on the ground because rain was an important factor for the fallout of radioactive materials. The Fukushima school monitoring data was obtained at the beginning of April, which is before the decontamination campaign, so we expected the radiation levels to reflect the amount of fallout just after the accident. We divided Fukushima prefecture into nine regions, because the absolute level of radiation differed for each region. Figure 1 shows the analysis results. This indicates that the radiation level and total rainfall does not show a significant correlation. This is probably because the total rainfall is not accurate. We know from JMA observations that the weather in Fukushima on March 15 was weak snow or rain in the afternoon, but weak precipitation is poorly captured especially when the radar site is distant. The calibration of the rainfall by ground observations was not successful because weak rainfall is often measured as 0 mm of rainfall. Since the fallout of radioactive materials is highly affected by small-scale weather conditions and the location of a plume, we concluded that the relationship between the radiation level and total rainfall cannot be analyzed at a reliable level of accuracy using the available data.

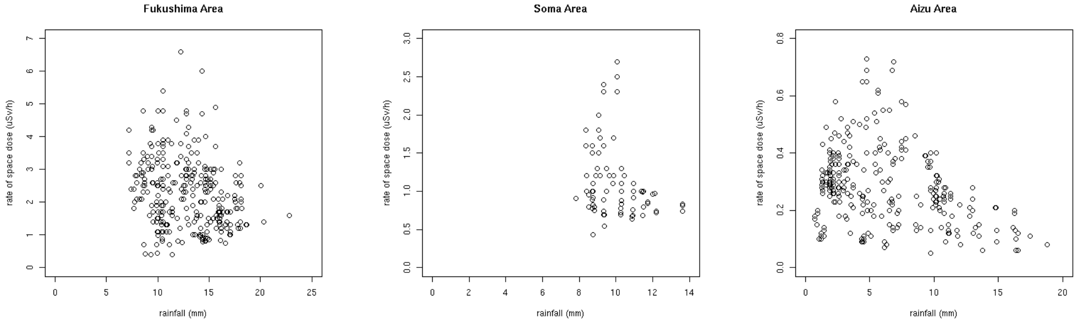

Figure 1. Relationship between radiation level (obtained from the Fukushima school monitoring data) and total rainfall (on March 15 and 16 from the weather radar data). Panels represent three regions in Fukushima prefecture, Naka-dori, Hama-dori and Aizu, respectively.

5 DATA VISUALIZATION

Data visualization for disaster response uses a standard presentation tool such as a map or timeline, because the most basic facets of disaster-related information are space and time. We intensively used Google Maps to visualize the spatial data because of the convenient API they provide. Here, we introduce two examples of data visualization, namely “Wind Map around Fukushima Daiichi Nuclear Power Plant” (http://goo.gl/gzPmR) and “Electrical Japan” (http:// goo.gl/HwLVC).

The first map visualizes the dispersion of radioactive materials from Fukushima Daiichi nuclear power plant. Many people were desperately in need of meteorological data on the wind and rain just after the accident, because they provided information on the transport of radioactive materials; wind contributes to horizontal movement, while rain contributes to vertical movement. Meteorological data, however, were not easily accessible after the accident for three reasons: 1) ground observations became inactive due to the shock of the earthquake, 2) weather forecast simulation data were only available to experts, and 3) the Japanese government decided not to release dispersion simulation results from SPEEDI (System for Prediction of Environmental Emergency Dose Information) to avoid panicking people. Our activity is to solve the second barrier by improving the access to the visualization of existing data. Atmospheric simulation data for weather forecasting, namely numerical model GPV (Grid Point Value) data from JMA, are the best data for this purpose, because it offers meteorological elements such as pressure and wind on grid points from the surface to the upper atmosphere for selected pressure levels. We developed a system to visualize the GPV data on Google Maps, and released it on March 22. We tried to release the service as quick as possible, but it took 10 days after the accident, which was later than major release events on March 15 and 21. However, the service quickly attracted people's attention, and the page view of the service reached about one million as of January 2012.

The second map, Electrical Japan, focuses on a long-term issue, the energy policy of Japan, because the usage of nuclear power became a hot issue. As the basis of the discussion, we built a comprehensive database of the power plants in Japan with the location and power information to summarize the quantitative data on the current status of electricity generation. This was much more difficult than we first thought because complete information should be reconstructed from the complicated interpretation of fragmented information available from many sources. Moreover, we focused on DMSP (defense meteorological satellite program) nighttime lights of the earth to use nighttime lights as the proxy of electricity consumption. The power plant map for nighttime lights clearly visualizes the relationship between areas where electricity is made and used. The energy consumption of the Tokyo metropolitan area was supported by the electricity generated in rural areas, such as Fukushima. The map was designed for aesthetic beauty to reduce the unnecessary stimulus on this hot issue.

6 DATA DISSEMINATION

After the data are visualized on the website, the final step is to promote the services to the general public. Social media, especially Twitter, was used for this purpose, because social media was intensively used for sharing information after the earthquake. We built several Twitter bots to push information, such as @wind_f1 tweeting wind direction for the next 24 hours. The information in a tweet was limited to 140 characters, but it was enough for the minimal role of a tweet, namely the notification of an update and the "heartbeat" of the system, with a link to a detailed web page. This notification service is only useful for real-time data, not for archival data, but it is always valuable to think about the usage of social media to attract people's attention to valuable data.

7 LESSONS LEARNED

We tried our best in the limited time available after the disaster, but the result is far from satisfactory. Lessons learned from our activity can be summarized as follows. First, we need to foresee the evolution of a disaster, and prepare data in advance before they are actually needed. As long as we start a project after something happens, the data become available later than they are actually needed. Solutions are first to improve the speed of software development. Second is to increase the flexibility of software development to make it improvisational and, third is to expand the imagination for potential disasters to prepare data before the crisis actually happens.

Secondly, a mechanism may be necessary to coordinate the voluntary projects started just after the disaster. At the time of a crisis, we cannot afford good coordination among activities due to the limited amount of time and continuous changes to the situation, and we observed many overlapping projects with similar purposes and methods. Of course, independent and rapid development is necessary for extremely rapid response to a disaster, and the situation could be improved if we had a single platform powerful enough to aggregate related efforts. A good example is Google Person Finder, which was quickly recognized as the single central database of safety information and volunteers' power was focused on this platform to achieve a large amount of work.

Finally, we provided most data in Japanese due to limited time, but the data should be provided as multilingual resources, at least in English. This earthquake raised worldwide interests, but Japanese people were accused of providing little English information. A practical solution is to use machine translation, but the accuracy is unsatisfactory. A solution is to limit the type of information using a fixed structure and a small vocabulary.

8 REFERENCES

Utani, A., Mizumoto, T. & Okumura, T., (2011) How geeks responded to a catastrophic disaster of a high-tech country – Rapid development of counter-disaster systems for the Great East Japan Earthquake of March 2011. *ACM Special Workshop on the Internet and Disasters 2011*.

Leidner, J.L. (2007) *Toponym Resolution in Text*, PhD thesis, School of Informatics, University of Edinburgh.