Visualizing Crisis News Briefs

Andrea Krukowski, Meredith McCarron, and Ady Sevy

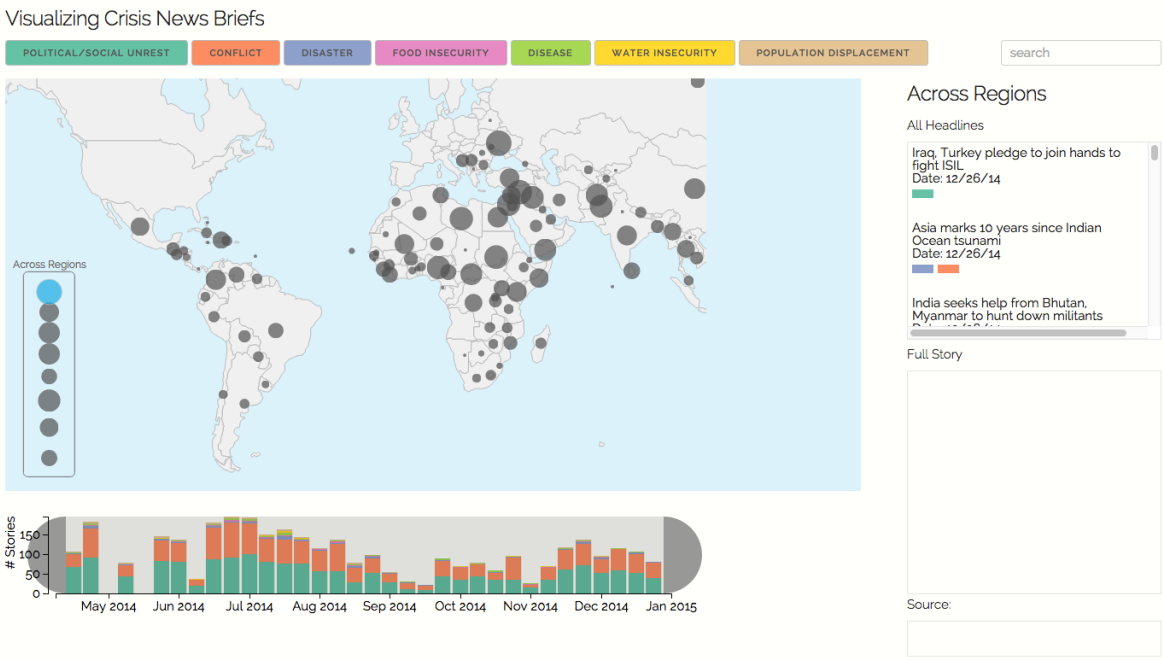

Fig. 1. Visualization tool upon loading.

Abstract—This goal of this project was to take texts containing summaries of important news events around the world ("Crisis News Briefs") and extract meaningful data to reveal spatial, temporal and categorical trends. Working with our client, UNICEF, we found that it was important to show broad trends while still making the most granular parts of the data accessible for users. Natural language processing tools, including Latent Dirichlet Allocation (LDA), Name Entity Recognition (NER), text parsing tools and k-means clustering, were used to process the raw textual data provided by the client. Briefs are sent out on a daily basis, which provided reliable date information for temporal trends. Geospatial data was extracted from the texts and mapped on the country level, unless documents specified the news stories as regional or global in scope.

Index Terms—Information Visualization, News, UNICEF.

1 Introduction

The Operations Center (OPSCEN) at UNICEF sends out via email extensive daily summary reports of recent security and crisis-related news to relevant staff members in every country in the world as well as headquarters policy staff and any other parties who may be interested (such as workers at other humanitarian or crisis-response organizations). The reports contain one-paragraph summaries of news articles, grouped by world region and country. Reports span about 10 single-spaced pages. Such large amounts of text data is very difficult for users to absorb, and comparing stories and themes over time is particularly burdensome. Headquarters staff primarily uses the briefs to browse, as reference for policy papers, and to keep track of risks to the security of the 12,000 UNICEF staff members who work in the field. Therefore, one major interest is following long-term and short-term trends on a country, region and worldwide scale. However, field staff may represent a very different use-case – less interested in browsing or events affecting in other regions. They are more likely to focus on the events in the country where they are stationed or the events in countries nearby that may spillover into their area.

Trying to cater to these varied use-cases for the data would be one of the biggest challenges of our project – therefore, we have decided to focus on the first use case since the primary client is the headquarters staff and we have regular access to this group of end-users. Also, we believe that this use case is the more general use case. A product tailored to this group may also be helpful to other potential end-users, as opposed to the other way around.

At present, the only way to use these briefs to identify trends is for users to go back through their email and scan past reports for relevant stories. Headquarters staff hope to be able to identify trends not just in the past, but to anticipate potential emerging trends and better identify weak signals. We asked the OPSCEN staff what they wanted field staff to do with the reports they compile, and they were unequivocal – they want field staff to take action.

- *Andrea Krukowski. E-mail: ajk583@nyu.edu.*
- *Meredith McCarron. E-mail: mkm318@nyu.edu.*
- *Ady Sevy. E-mail: as8202@nyu.edu.*

Information Visualization graduate class.
Polytechnic School of Engineering, New York University.
Instructor: Prof. Enrico Beritini.

2 Related Work

This project presents the challenges of mining documents and text and working with spatio-temporal data. As summarized by Brehmer et. al., there have been diverse approaches to analyzing large document collections and presenting visual clusters, including interactive tag clouds, hierarchical trees, and scatterplots. For example, the Overview tool (Figure 1) for investigative journalists allows users to hierarchically cluster text documents based on content similarity, visualizing the the collection as a tree. Our project has several similar characteristics: we will mine text to identify trends in current events. However, the project data is inherently structured, coming from news briefs selected by a team with specific goals and audience in mind, and we are able to pre-identify attributes for the users to cluster and search articles by, discussed in the Data section section below. Also, we are dealing with spatio-temporal data.

We also examined current crisis-mappers and related tools:

2.1 LRA Crisis Tracker

This tool (Figure 2) is a joint project of NGOs Invisible Children and Resolve, updated daily with recent activity by suspected fighters of the Lord's Republican Army (LRA), with the ability to filter by type of encounter and by date. This data is similar in that it is manually curated and added to the visualization, though their location data is much more granular. We found the date slider at the bottom of the map especially appealing.

2.2 CrisisNet (Ushahidi team)

This is a in-development project that trawls Facebook and YouTubes APIs for geotagged posts from a pre-determined location (thus far, Syria only). Content is then automatically translated, categorized and geotagged to a specific region or city. The data is represented on a heatmap, split into regions of Syria. A tooltip displays the number of reports for the highlighted region, and clicking on a region will brings up an example post. This data is similar to ours in that it has to be auto-tagged and categorized with natural language processing, and that it is plotted on a heatmap split by region. However, our data is curated, not automatically gathered by API, and thus needs to be added by the user. Additionally, we want users to be able to view all news stories for a given area, and text-based data is not well suited to being viewed exclusively in a tooltip.

2.3 Geovisual Analysis and Crisis Management

The Pennsylvania State University team have created an interactive visualization tool that integrates mostly georeferenced data from multiple sources, including news bulletins, emails, photos, and video feeds, into one environment to improve situational awareness, decision-making, and crisis management throughout a crisis. The tool provides another example for how to visualize text data with a spatio-temporal component. However, the end goal and user is drastically different than our project's: the objective of the Geovisual Analytics application is to aid crisis managers throughout a crisis and response.

3 Data Analysis and Abstraction

OPSCEN provided 283 files of news briefs, spanning 2013 to 2014. For our demo we have used only the 2014 files, i.e. 109 files. Each news brief is dated and provides a short one-line statement about a news item, as well as a paragraph-long summary and url to the source article. The news items are grouped by country and region. The attributes inherent in each news brief are: date, region, country, event title/one-liner, event summary, and source article url. These are all categorical datatypes, except for the date, which is quantitative. We will use all of these attributes in our project. Additionally, we plan to derive several other fields:

- Story category (categorical): we used a keywords dictionary and Latent Dirichlet Allocation (LDA) [2] to characterize each news item by topic to allow users to filter news items by item type. For instance, an item like "New malaria tool to help track insecticide resistance" may be tagged as "disease." Our categorization is not hierarchical, though an item may be tagged with multiple categories.

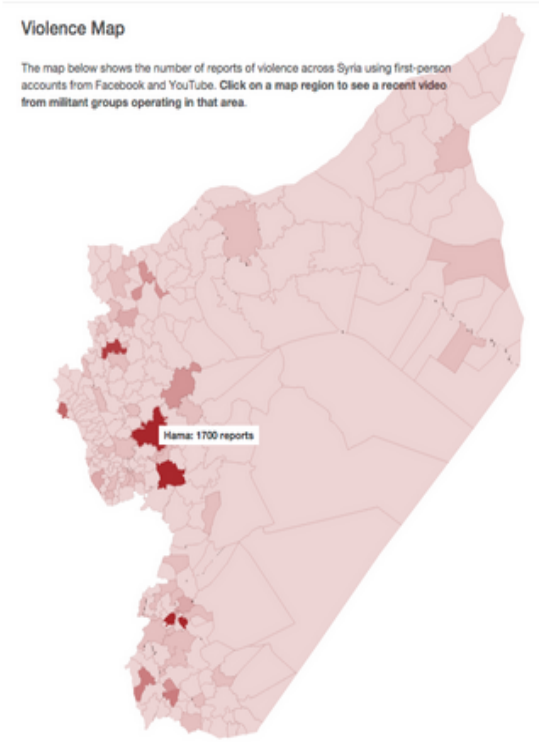

Fig. 4. CrisisNet.

- Entities (categorical): We used Name Entity Recognition (NER) software developed by The Stanford Natural Language Processing Group [4] to identify organizations, locations, and people and link news items to entities. This enabled the search functionality detailed in Section 5. Additionally, a basic normalization was applied to the output of the NER to handle results initially being case sensitive and having differences in punctuation.

- Geolocation - Latitude, Longitude (categorical): For each country we extracted its geographical location using Google Maps API [3]. The longitude and latitude represent the countryś centroid.

4 Task Analysis and Questions

Our visualization aims to answer two main questions detailed below. After outlining each question, we decomposed them into detailed tasks. The visual encoding and interaction design which support each task will be referred to by task number in the next section.

4.1 Question 1

What crises or security-related events happened today or within a specified timeframe, i.e. a snapshot of the data? We aim for users to be able to select a timeframe and be able to get a sense of the number and type of events that occurred in a country or region.

T1 What is the overall spatial distribution of events?

- **T1.1** How do those spatial patterns vary when filtering by specific tag/s?
- **T1.2** How do those spatial patterns vary when filtering by specific entities (name/location/organization)?

T2 What are the distribution of news bulletins categories (i.e. tags) within a specific country?

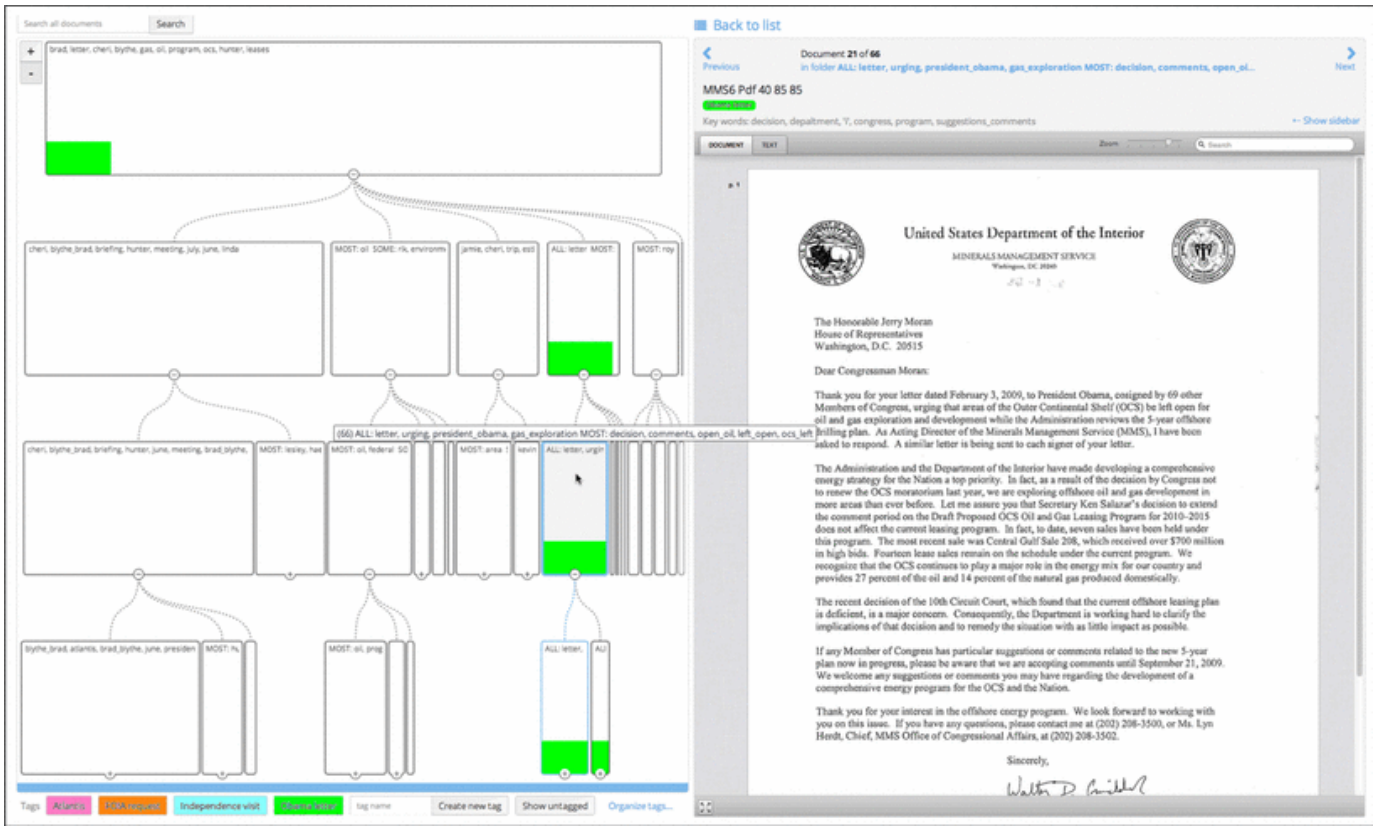

Fig. 2. The Overview tool.

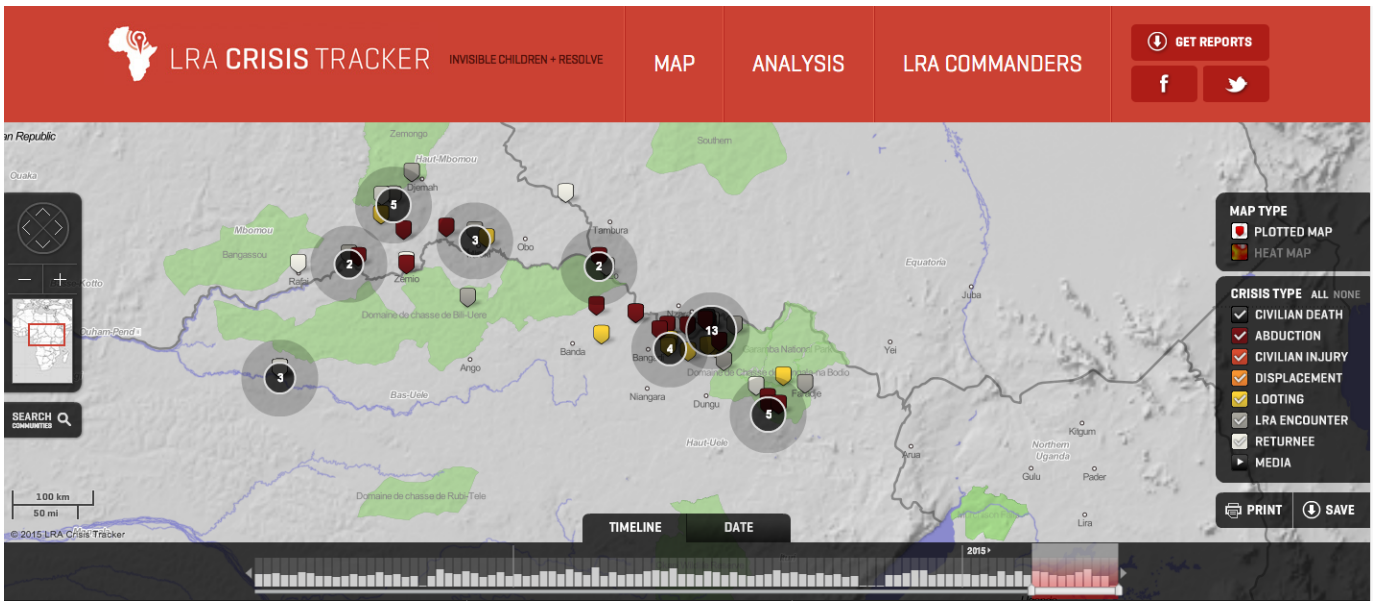

Fig. 3. LRA Crisis Tracker.

T3 What are the details of an individual event?

T4 What is the source of the original story?

4.2 Question 2

What are long-term or nascent trends? How do regions or countries compare? Our goal is to allow the user to identify spatial-temporal patterns. The user will be able to filter news items by location, category tags, and entities and see trends on the map; the user will also be able to see news items in aggregate or split by the selected filters (country or category) in the accompanying line graph. We hope to allow the user to aggregate by either number of events/news items or people impacted. The filtering tools will allow users to compare two or three countries or category types.

T5 What are the temporal trends of number of events?

T5.1 What are the temporal trends of number of events when

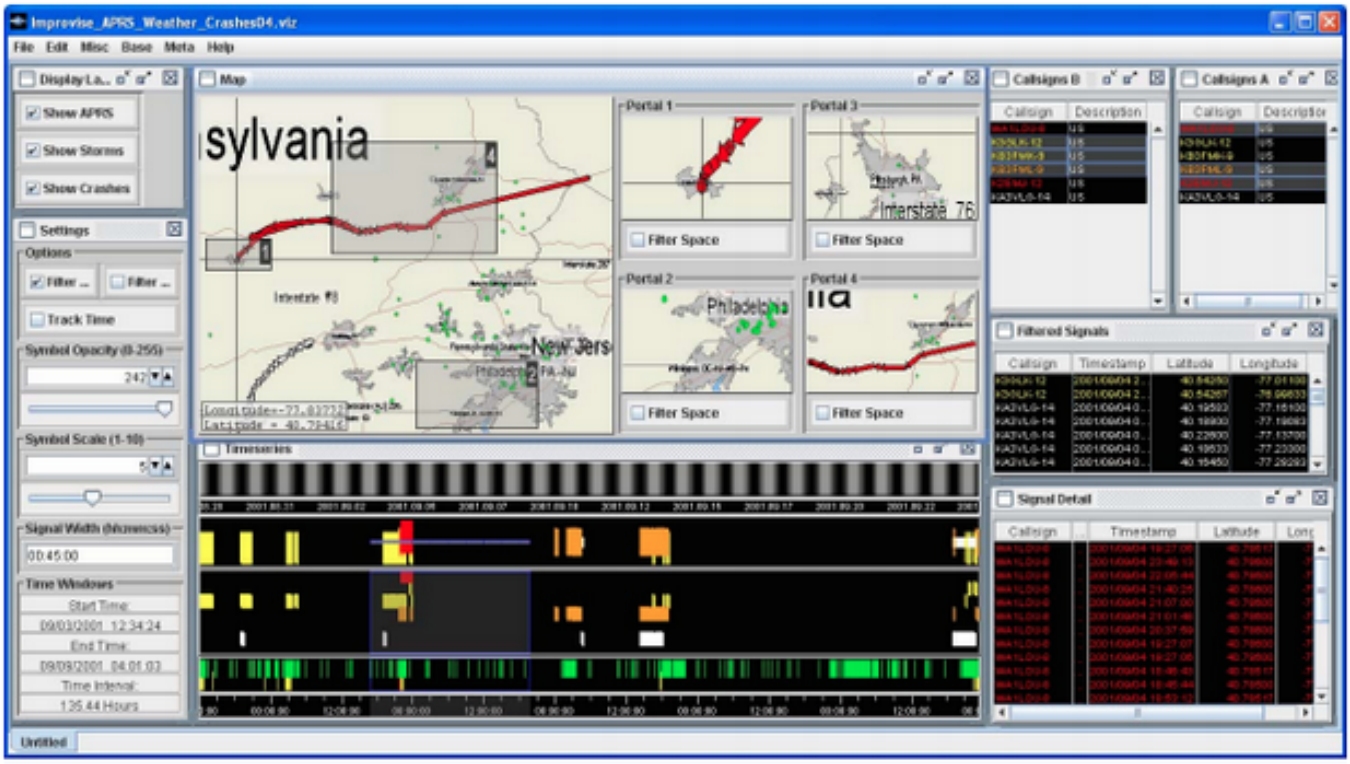

Fig. 5. An Improvise Visualization snapshot.

filtering by category (i.e. tag)?

T5.2 What are the temporal trends of number of events when filtering for specific entities (name location)?

T6 Compare trends of events by countries/ categories/ regions.

5 Visualization and Interaction Design

Figure 5 displays our visualization upon loading. The tool is comprised out of five pieces: filter buttons, search field, map, info panel (headlines, stories and source) and trend chart. Table 1 and Table 2 lay our visual encoding and maps it back to the tasks defined in section 4.

Interaction		Supported Task
Filtering	by categories	T1.1
	by search	T1.2
Clicking on a circle		T2
Selecting trend chart subset		T1, T5

Table 2. Interactions and tasks mapping.

How would a user go about using the visualization to gain insights on the data?
The following section will simulate a user trying to explore more about Israel and Hamas conflict in the summer of 2014.

1. User starts to explore the functionality, upon hovering on a circle - the name of country is displayed, along with the total number of stories for that country under the selected filters. Not surprisingly, it seems like many of the circles are concentrated around the Middle East (Figure 5).

2. Filtering by category - the user starts to explore the different filtering options. Since her current focus is on conflicts, she decides to display only the stories related to conflict. Notice that the trend chart has changed to reflect this selection (Figure 6).

3. Filtering by time frame - the user wants to focus her attention on the last violent conflict between Israel and Hamas last summer, hence she adjusts the trend chart brush to include the relevant weeks (Figure 7).

4. Searching for specific entity of interest: person, location or organization. Since the current focus is the Gaza conflict, the user narrows down her selection using the search bar (Figure 8).

5. The user added political and social unrest category to the view. While the circles around the Middle East make sense, the user noticed that there is one story in India. Clicking on the India circle and selecting the story reveals an interesting insight. This story was related to an anti-Israel protest occurred in Kashmir, where tragically a teenager was killed (Figure 9).

6. This exploration allowed our user to go beyond the expected news article and to discover evidence on how of instability of a certain region affects other regions (Figure 10).

6 Findings and Insights

Finding #1. Spatial Patterns:
The visualization tool can be used to reveal spatial patterns. For example, at the starting view, one can see that there are no crisis events tracked by UNICEF in the U.S. or Western Europe. Also, after filtering by categories, the placement of the markers shows where certain types of events are most common. For example, food insecurity is an issue in Central and South America, as well as throughout Central and Northern Africa.

Finding #2. Temporal Trends:
The tool can also be used to see temporal trends; for example, after searching for ?Islamic State? in the search bar, one can see that there are no stories before June 2014. By panning the timeline with a fairly size timeline window, one can see that the influence of the Islamic State across the region; initially, there are only stories in Syria and

Visualization Piece	Object	Mark	Channel	Encoding	Supported Task
Map	News stories	Point	Map Position	Country	T1
			Area (2D size)	Number of stories	T1
Trend graph	News stories	Bar	X-Position	Date (week)	T5
			Y-Position	Number of stories	T5
			Y-Position	Number of stories	T5
			Color Hue	Primary category	T6
Text Sidebar	News stories	Text			T3, T4
		Tags	Color Hue	Category	T2

Table 1. Visual encoding and tasks mapping.

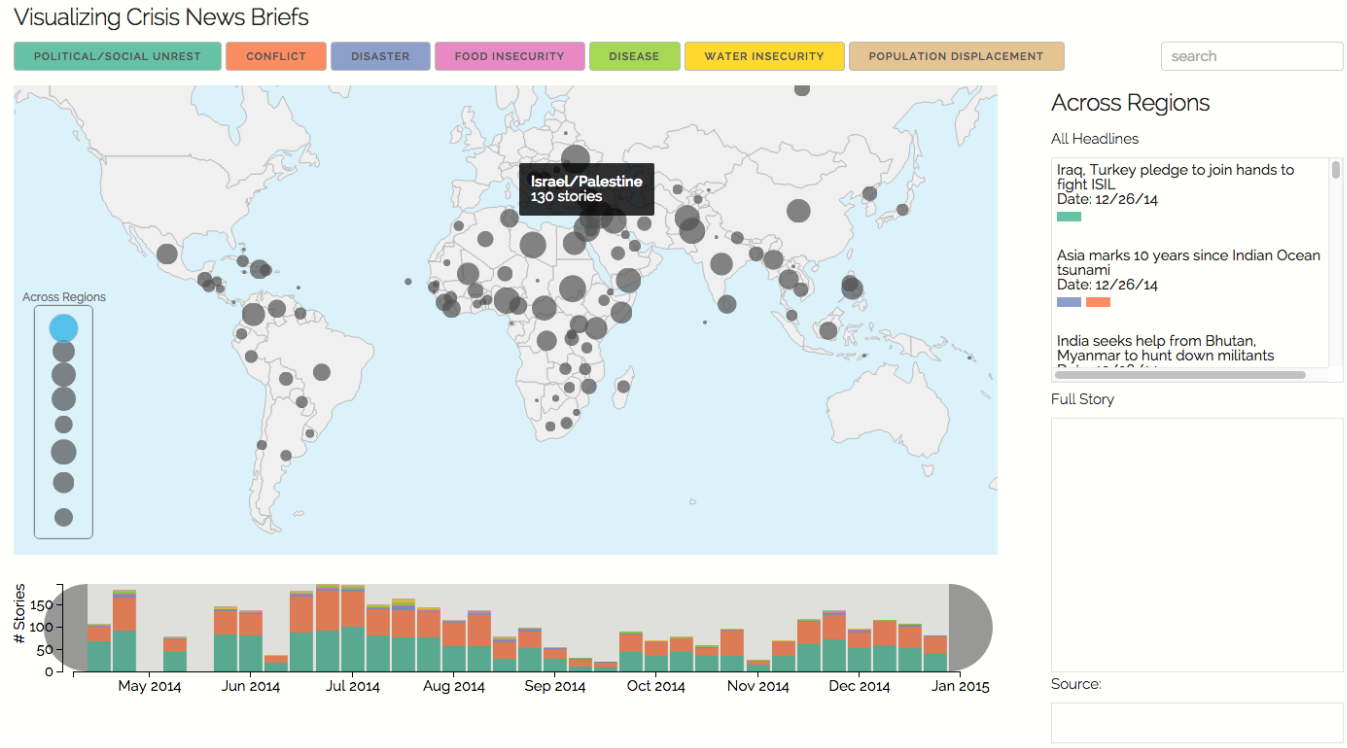

Fig. 6. Tooltip upon hovering on a circle.

Iraq but, over time, stories pop up in other countries, like Algeria and Libya.

Finding #3. Unexpected Connections:
Disaster, food insecurity, disease, and water insecurity often are often linked with conflict or political/social unrest, as shown by the fact that many stories with categories in the first group are also tagged with "conflict" or "political/social unrest". Additionally, A user can also reveal more interesting, unexpected insights. For example, a user may be interested in the Gaza conflict and search for all stories connected with Gaza using the search bar. The tool filters all circles, revealing many in the Middle East and, unexpectedly, one story in India. Clicking on the India circle and selecting the story reveals an interesting insight: this story was related to an anti-Israel protest that occurred in Kashmir, where tragically a teenager was killed.

7 Limitations and Future Work

The main limitation of our project is the source of the data. As mentioned in the Data Analysis and Abstraction section, our data source was manually structured Word documents. After overcoming the parsing phase - finding good tools, and setting up the parsing logic - because the data was generated manually, we faced multiple inconsistencies that, in some instances, forced us to discard some of the data. We envision later iterations of the tool including normalized tags for country names, categories, and entities that users would choose from when writing up the briefs, so that data inconsistencies could be largely avoided.

Another key limitation we had was related to assigning categories to the news stories. Since the earliest stages of the project we knew that filtering by category was key functionality, as it was crucial for our defined tasks. We ended up using both an unsupervised method (LDA), and a supervised method to assign categories by keywords. However, both solutions involved some amount of manual work, which we decided was an acceptable trade-off for ensuring the quality of our categories. As stated above, we suggest implementing user inputs for tags and categories so that: (1) the groups could be more flexible and refined; (2) this user feedback can be used later as input for an automatic classifier. The more tagged stories we have, the better the search and analysis functionality will be.

Additionally, the NER normalization can be improved such that many fewer unique entities would be displayed on the drop down menu. For instance, items such as "ISIS", "ISIL" and "Islamic State" could be consolidated. We could also add keywords that are not necessarily names, places, or organizations. A user may want to search for all hurricanes or earthquakes–nouns like this are not currently in the

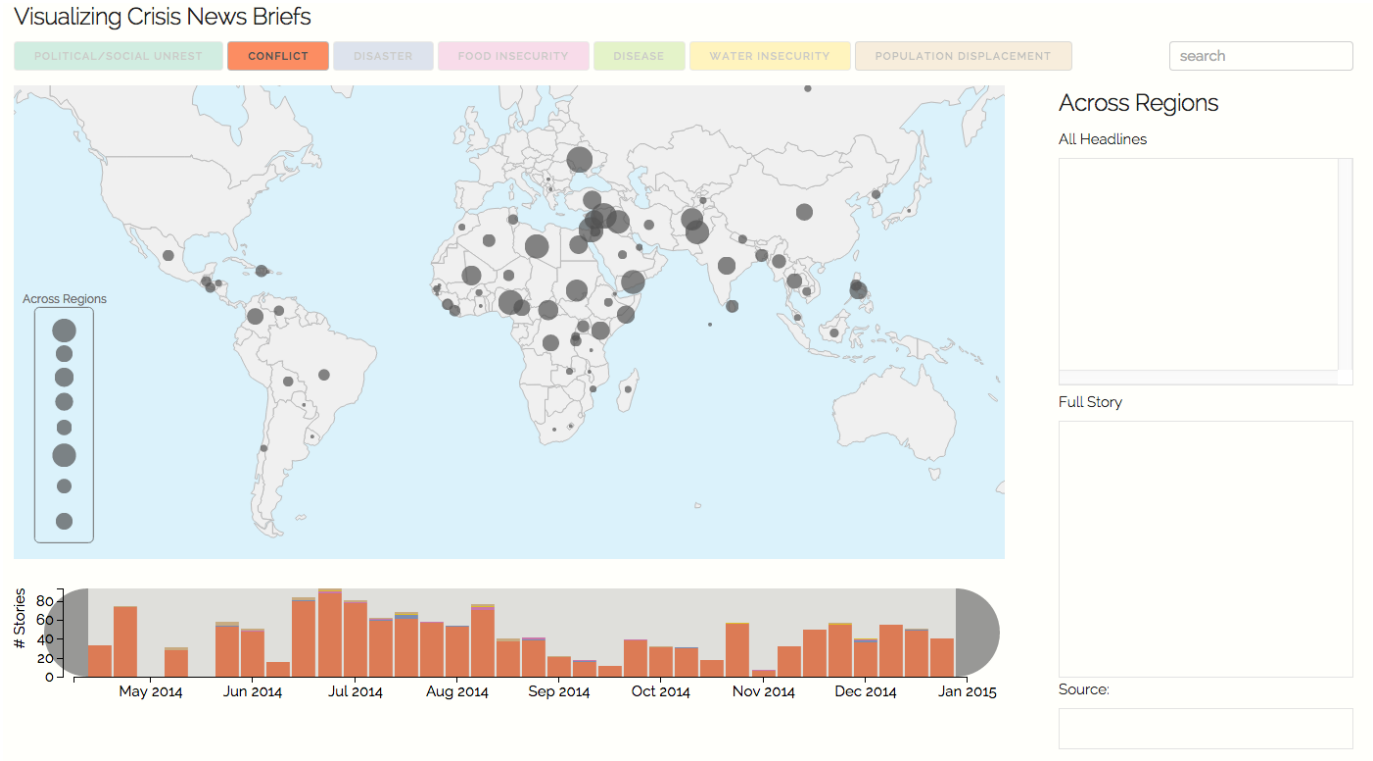

Fig. 7. Filtering by category.

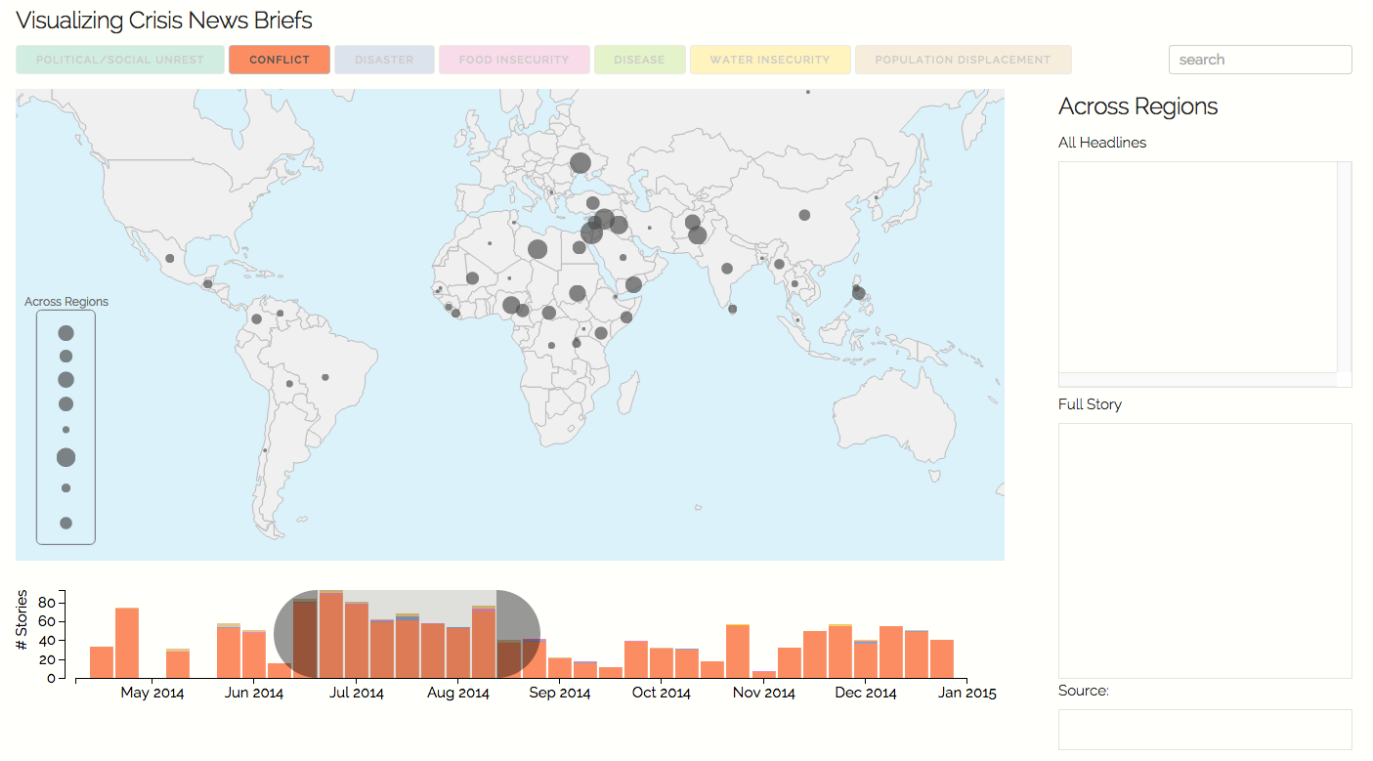

Fig. 8. Filtering by time period.

keyword search but may provide a more granular type of filtering than the categories but less specific than specific organizations or names.

Lastly, our customer is interested in an additional application of the NER, such that users could, for example, create a query for all organizations that are related to Iraq to get an output of a detailed list of the items, including magnitude of stories and trend over time.

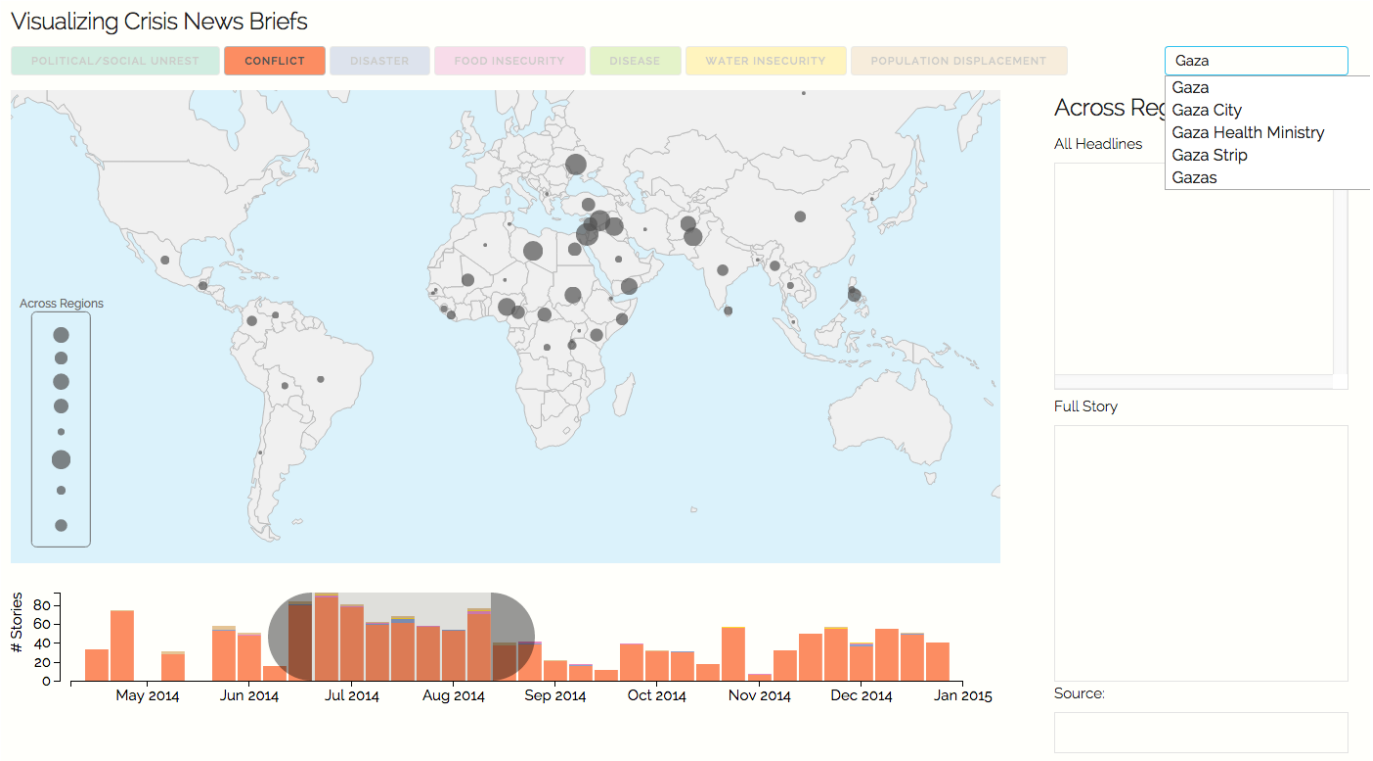

Fig. 9. Searching specific entity: person, organization or location.

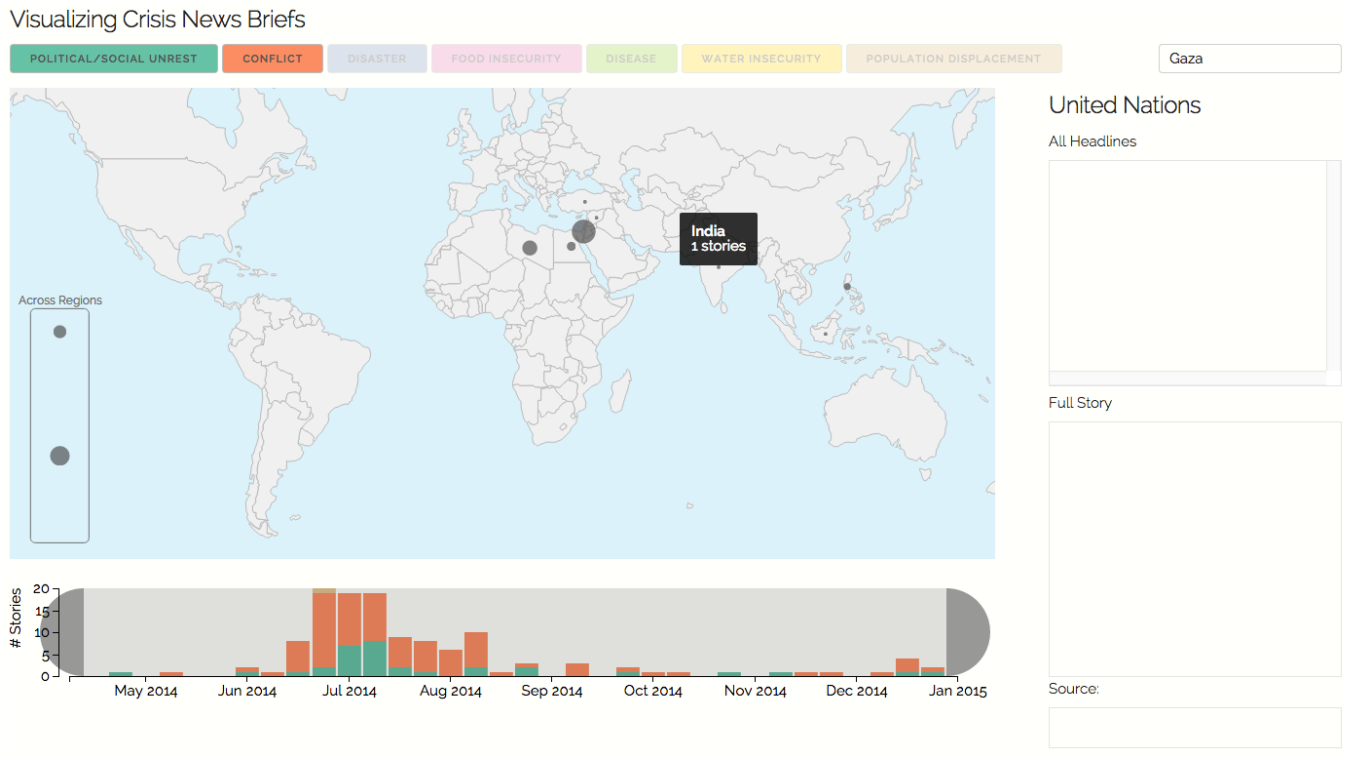

Fig. 10. Surprising pattern of Gaza related story in India.

8 Conclusion and Lessons Learned

This project was very instructive in forcing us to design a visualization that could serve several needs without overwhelming the user. The fact that we have spent most of our time on the planning rather on execution has proven to be very beneficial, as our final product did not deviate much from our original sketch. We strove to address the need to observe trends over time, space and broad themes, while still making the underlying data browsable so that users could make their

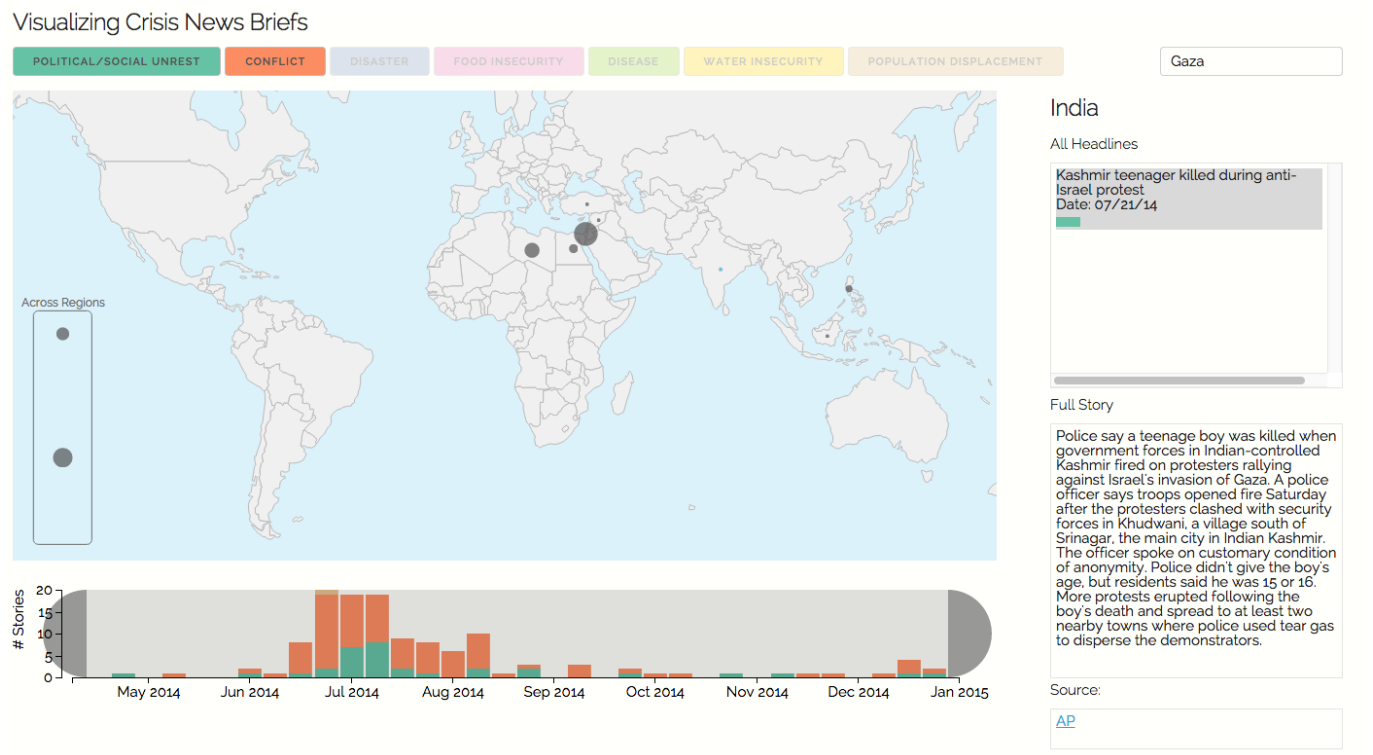

Fig. 11. Revealing the detailed story.

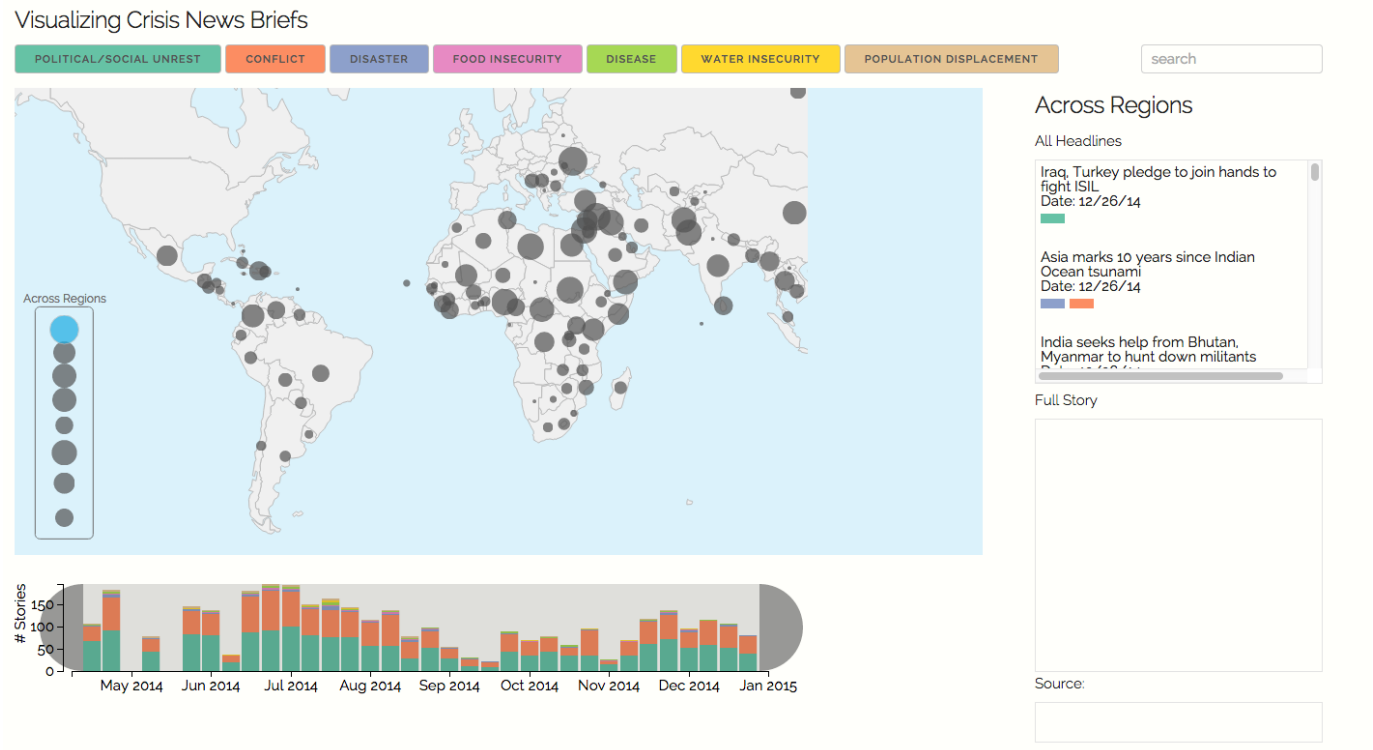

Fig. 12. Visualization Crisis News Brief upon loading.

own discoveries of trends that we did not quantify, find sources for position papers and publications, or just to keep up with current events in their country and across the world.

The unstructured nature of our data means that the avenues for exploration are almost endless, so keeping focused was challenging but crucial. We see great potential for all kinds of additional functionality for this tool, if the client wishes to go further with the data.

9 LINKS

Github https://github.com/adysevy/unicef

Demo http://adysevy.github.io/unicef/WebApp/

Video https://vimeo.com/127946758

ACKNOWLEDGMENTS

We would like to thank Eva Kaplan from UNICEF, for her support and cooperation in this project.

REFERENCES

[1] Brehmer, M., Ingram, S., Stray, J., & Munzner, T. (2014). *Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists*.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. emphLatent dirichlet allocation . J. Mach. Learn. Res. 3 (March 2003), 993-1022.

[3] Google. *Google Maps API*. May, 2015. https://developers.google.com/maps/

[4] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. *Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling*. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/ manning/papers/gibbscrf3.pdf

[5] LRA Crisis Tracker. (n.d.). Retrieved May 9, 2015, from http://lracrisistracker.com/

[6] Mapping the Syrian Conflict with Social Media. (n.d.). Retrieved May 9, 2015, from http://crisis.net/projects/syria-tracker/

[7] The Operations Center (OPSCEN) at UNICEF. (n.d.). Retrieved May 9, 2015, from https://www.dropbox.com/sh/a9e62mraviveldt/AACcjVTmMPZtRqzByDL8n82ia?dl=0

[8] Tomaszewski, B. M., Robinson, A. C., Weaver, C., Stryker, M., & MacEachren, A. M. (2007, May). *Geovisual analytics and crisis management*. In Proceedings of the 4th International ISCRAM Conference (pp. 173-179). Delft, the Netherlands.