`katrina` R Data Package Codebook

Carter T. Butts* Ryan M. Acton† Christopher Steven Marcum‡

Monday 8th November, 2010

1 Overview

1.1 Philosophy

The extraction of relational information from field documents is intrinsically difficult. Given the highly variable nature of the source documents, great care must be exercised to permit reliable coding within a reasonable time frame. To the extent possible, the following principles have been followed in pursing this end:

1. Reduce the use of human judgment wherever possible;
2. Limit the number of coding decisions which must be made during each stage of the coding process (i.e., code only one thing at a time, separate tagging and coding tasks);
3. Employ multiple coders for all tagging and coding tasks;
4. Assess inter-coder reliability for all tagging and coding tasks;
5. Independently verify all critical procedures;
6. Track all changes by both personnel ID and date of change;
7. Store data at each phase of the tagging/coding process;
8. Document all procedures.

The end result of this process, it is hoped, is a data set of much greater quality and scope than could be obtained using ad hoc methods.

2 Raw Material Collection

Raw materials were collected from live websites and internet archives that were identified as potential sources of dynamic relational information by querying "Hurricane Katrina" (and variants) in a popular internet search engine. Data collection included semi-automated downloading and saving of digital materials at roughly 12:00AM until the source websites were exhausted each day from 08/24/05 and 09/05/05. Because many `urls` changed on daily basis, (i.e., through human error, archiving, or website reorganization) a human coder would verify each download and correct any missed attempts.

*Department of Sociology and Institute for Mathematical Behavioral Sciences; University of California, Irvine.
†Department of Sociology; University of Massachusetts, Amherst.
‡Department of Sociology; University of California, Irvine.

3 Raw Material Preparation

3.1 Cataloging and Processing of the Information Sources

This section details the general strategies employed to organize those documents that served as the raw source materials. These documents are herein called *source* documents, which refers to the *source* organizations that produced them.

Each source organization utilized in this dataset was given a unique four digit identifier, in addition to both abbreviated and full names. The first two digits of the ID identify state or organization kind. State codes, which typically comprise the first two digits of the ID, are given in Tables 1 and 2. Note that the District of Columbia is treated as a state here.

Code	**State**	**Abbreviation**
10	Alabama	AL
11	Alaska	AK
12	Arizona	AZ
13	Arkansas	AR
14	California	CA
15	Colorado	CO
16	Connecticut	CT
17	Delaware	DE
18	District of Columbia	DC
19	Florida	FL
20	Georgia	GA
21	Hawaii	HI
22	Idaho	ID
23	Illinois	IL
24	Indiana	IN
25	Iowa	IA
26	Kansas	KS
27	Kentucky	KY
28	Louisiana	LA
29	Maine	ME
30	Maryland	MD
31	Massachusetts	MA
32	Michigan	MI
33	Minnesota	MN
34	Mississippi	MS
35	Missouri	MO

Table 1: State Codes and Abbreviations

Code	State	Abbreviation
36	Montana	MT
37	Nebraska	NE
38	Nevada	NV
39	New Hampshire	NH
40	New Jersey	NJ
41	New Mexico	NM
42	New York	NY
43	North Carolina	NC
44	North Dakota	ND
45	Ohio	OH
46	Oklahoma	OK
47	Oregon	OR
48	Pennsylvania	PA
49	Rhode Island	RI
50	South Carolina	SC
51	South Dakota	SD
52	Tennessee	TN
53	Texas	TX
54	Utah	UT
55	Vermont	VT
56	Virginia	VI
57	Washington	WA
58	West Virginia	WV
59	Wisconsin	WI
60	Wyoming	WY

Table 2: State Codes and Abbreviations

The second two digits of the ID variable generally categorize source organizations by their geographic scope of operations. The ranges as defined below allow for multiple organizations to reside at each level when there exists multiple organizations that have as identical the first two digits in their ID:

Code Range	Description
00–09	state level source
10–39	city level source
40–69	county level source
70–99	other organization

If the source organization is a U.S. military organization, the first two digits of ID variable are '77' and the second two digits uniquely identify each organization. If source is a non-profit organization, the first two digits of ID variable are '88' and the second two digits uniquely identify each organization. If source is some other U.S. governmental agency or department, the first two digits of ID variable are '99' and the second two digits uniquely identify each organization.

The list of source organizations, along with their ID numbers and abbreviated names, is given in the

`katrina.bysrc` manual page within the R `katrina` data package.

3.2 Cataloging of the Raw Materials

This section describes the creation of ID numbers for each of the source documents (raw materials) from which the data arise. ID numbers for these raw materials are eight digits in length. The first four digits correspond to the source organization ID number, outlined previously. The next four digits were assigned sequentially by publication time, where possible in the format MMDD (there was no need to indicate the year, as all of these documents were published in 2005). This method of numbering was only feasible for sources that did not issue multiple documents in the same day. Otherwise the next four digits consist of '55' as the first two characters, and the last two are used to uniquely identify each document.

3.3 Rendering Raw Materials to Text

Source documents were typically in either `.doc`, `.pdf`, or `.html` format at the time of data collection. The following techniques were used to convert source documents into a common format (`.txt`):

- **For `.pdf` files:** Two programs were used to implement the `.pdf` to `.txt` conversions: Adobe Acrobat 7.0 Professional and ABBYY FineReader 8.0 Professional Edition. Most conversions were implemented using Adobe Acrobat, but in cases where the conversion output was too difficult to read by a human, ABBYY FineReader was used instead to carry out the conversion. Many times Adobe Acrobat performed the nicer conversion between the two programs. Output was always specified as 'Text (Accessible)' format in Adobe Acrobat. Both programs were used on a Microsoft Windows XP Professional machine.
- **For `.html` files:** All conversions from `.html` to `.txt` were performed using the Detagger 2.4.0.12 program for Microsoft Windows (`http://www.jafsoft.com/detagger/`). Detagger was configured to remove all html tags from the `.html` source file.
- **For `.doc` files:** The conversion of `.doc` files into `.txt` format was comparatively simple, requiring only that the user save the file as a `.txt` file from within Microsoft Word.

After conversion, all files were renamed with their corresponding 'ID' variable value. For example, if a document's value on the 'ID' variable is 99045503, its new name becomes `99045503.txt`.

3.4 HTML Conversion of the Raw Materials

Once all documents were converted into `.txt` format, copies of each file were made. Into these copies were inserted basic html tags to render the files readable within a web browser, such as Mozilla Firefox (e.g., the
</br> and tags). These copies were then given an `.html` extension. These resulting HTML files served as the final versions of the documents which were processed and analyzed in order to produce this data package.

4 Coding Methodology

4.1 Phase 1: Tagging of Organizational Actors

A detailed explanation of the coding procedures used with the source documents is given in Butts et al. (2010).

The preliminary identification of organizations included any formal organization of entities or a group of people who appeared to be acting as or serving as representatives of an organization. Less obvious examples of organizations included (but were not limited to):

- Details, operations, missions, tasks, resources (e.g. a fencepost & wire resource)
- Teams, sections, task forces, branches, units, departments
- Voluntary organizations/agencies
- Personnel/representatives/liaisons of organizations
- Schools, churches, farms, camps
- Airports, air force bases, air reserve bases
- Canteens
- Stations
- Emergency Operations Centers (EOCs) / Divisions/Departments of Emergency Management (DEMs) / Emergency Management Agencies (EMAs)
- Emergency Support Functions (ESFs)
- Courthouses
- Facilities
- General groups of people (inspectors, workers, crews, squads, patrols, troopers, troops)
- Vehicles carrying out duties (rescue boats, trucks of ice)
- Representatives of organizations (e.g., planning section chief, municipal contacts (if they're named specifically, not just a phone number))
- Shops, retail locations, gas stations

This preliminary pass at identifying organizations yielded a high number of false positives (i.e., entities that initially appeared to be an organization, but upon further research, were not actually organizations at all). Through the next step, named entity coding, the final set of actual organizations was realized.

4.2 Phase 2: Coding of Named Entities as Organizational Actors

This step finalized the set of organizations to be included in the data set. The rule of thumb utilized here was that the organization in question must be clearly identified as an existing organization from other information sources (such as Internet searches or through identification in other documents), or must be a sub-unit of a larger organization. Often times locating an organization's main website and/or address and contact information confirmed the organization's actual existence. Some organizations, however, were difficult to verify. Vague references to groups that initially appeared to be organizations but that could not be traced to any actual organizations in other information sources, were not included. For example, 'a team of electricians' would have been ignored if there was no clear reference to some parent organization or such from which the team arose.

There was some debate as to whether an Emergency Support Function (ESF) counted as an organization. An ESF at face value is not an organization per se, rather it is a collection of organizations that are designated to manage a given task area during an emergency situation. In the case of many of the SITREPs examined, an ESF was often referred to as engaging in activity with other organizations. For example, "ESF 13 deployed the Florida Army National Guard to Palm Beach this morning at 0900." Because of this convention, an ESF was treated as an organization herein only when it was shown to be engaging in some specific relational activity with other organizations.

Another interesting aspect of these data is that many of the identified organizations are sub- or sub-sub-units of larger organizations. Every effort was made to provide the most detailed information about an organization and its parental lineage. For example, the state of Florida's Department of Health might have a coalition named ABC, which in turn sends a task force of specialists called XYZ to help Georgia with its recovery efforts. Where the information is available, all parental lineages like this are recorded in these data.

There were times when it seemed fairly obvious to the coder that a given named entity truly was an organization, but the referent text left an element of ambiguity. The named entity may even have appeared to be engaging in some relational activity with some other organization. A problem with these SITREPs, largely an artifact of being converted from `.html` or `.pdf` to `.txt` and then back to `.html`, was that the formatting of tables and lists was sometimes lost. This often led to scrambled or otherwise unreadable lists, which made deciphering a given named entity quite difficult. In such cases when a named entity was difficult or impossible to properly code, the entity was not considered to be an organization.

It is also important to remember that these SITREPs were generated by humans, which means they are prone to error. Because of this, a coder of documents like these must often make use of contextual clues in order to deal with quirky phrasing, abbreviations, misspellings, or even the improper naming of organizations. While this introduces room for coding error, it is an essential part of the coding process. Using contextual clues reduces error in the coding process to the extent that not employing them would result in more error (i.e., some prior information is better than none).

A clear example of this was evident in the Florida SITREPs. Within several of these SITREPs, the abbreviation 'DoF' was used to stand for 'Division of Forestry'. Then, in one instance, after several SITREPs prior had been using the 'DoF' abbreviation, 'DoF' was used to stand for 'Division of Fire'. This was initially quite troublesome, as it led the coder to question all previous coding of 'DoF'. Upon investigation, no such Division of Fire existed in Florida. The Division of Fire fluke appeared in the same sections that the Division of Forestry references had appeared in SITREPs prior, and made no appearances anywhere else. It was determined that this was a mistake made by the preparer of the SITREP, and the Division of Fire references were coded as references to the Division of Forestry. This is only one example of many where some detective work and some context-based guessing revealed what was behind the confusion. Of course, there are instances when the coder had no idea what the referent text was intended to mean. This scenario points to the vital role that humans play, both in preparing the documents in the first place (which, as suggested here, are not immune to errors), as well as in interpreting ambiguities during the coding phase.

4.3 Phases 3 and 4: Relationship Tagging and Coding

In these data, two organizations are said to be relationally linked if the following actions were documented to have occurred between them within the time frame of the Hurricane response being reported in a given SITREP:

- There was an exchange of communication/ideas between organizations. The most common example is the sharing/providing/requesting of information from one organization to another.

- There was an exchange of manpower between organizations. Common examples include the sending/receiving of assistance, liaisons, representatives of organizations to/from other organizations.
- There was an exchange of material or financial support between organizations.
- There was an exchange of power between organizations. A common example is when one organization deploys/disengages or takes control of another organization.

It should be noted that the above detailed actions need not have been reciprocated between organizations.

For this pass through the raw materials in the coding procedure, no distinction was made about the kind or value of the relational ties between organizations. All relations were coded as undirected, and self-reflexive loops (a case where an organization was documented to be engaging in relational activity with itself) were not allowed.

Once all instances of documented relational activity between the mentioned organizations were recorded, various adjacency matrices were generated to represent the relational data.

Matrices were compiled in three different ways: (1) as indicating the relational activity between organizations mentioned within the same source document; (2) as indicating the relational activity between organizations mentioned by the same source organization; and (3) as indicating the relational activity between organizations which were documented as relating to one another within the same day (across all source organizations and source documents). These three methods of computing matrices directly yielded the data contained in the `katrina.bydoc`, `katrina.bysrc`, and `katrina.bydate` data objects, respectively. A "union rule" version of network, representing all organizations and relations aggregated over the whole period, can be found in `katrina.combined`.

4.4 Coding Organizational Lineage

Where possible, for each organization identified in the response networks, we identified their parent organization lineages. Ties in these lineage networks were defined by ownership and/or administrative authority of one organization over another. The sources of the lineage data were varied and included: websites, promotional material, organization hierarchy charts, newspapers, and organizational informants cold-called by our coders.

In many cases, parent organizations were easy to identify. Often it was the case that a given identified organization in the network was clearly a child of some broader organization or agency. Within the federal government, for instance, it is not uncommon to find such a nested structure of departments, agencies, offices, divisions, and so forth. We smooth over the variability in how some municipalities are structured by assuming that, unless strictly defined, towns and cities are child organizations of counties and parishes, which are child organizations of states. In these data, we found that this set organizations could be traced to a maximum of order four parent organizations with a maximum lineage length of five organizations. The network object `katrina.lineage` contains the data for the parent organization lineage network.

Because not every parent organization was observed in the SITREPS from which we built the response networks, the `katrina.lineage` contains 260 more organizations than the master `katrina.combined` network. An vertex attribute called "org.inset" is included in the `katrina.lineage` that indicates whether or not an organization was part of the observed response networks.

Every effort was made to ensure that the lineage reflects the state of the world at the time of Hurricane Katrina.

5 Organization Attributes

Several key covariates were collected on each organization appearing in the SITREPS. Coders obtained this information through a combination of sources, including the SITREPS, organization websites, and published research articles. These variables are described in Table 3, below.

We followed the coding schema of Tierney (2002) for the scale and type variables. The "group" variable was created to mimic, as closely as feasible, the coding of Lind et al. (2008) for their Katrina communication network.

Attribute name	Vector type	Description
first.appearance	numeric	The date of first appearance of each vertex. Entry corresponds to location in **katrina.bydate**
fema.region	numeric	The FEMA region codes for the headquarters location of each vertex
group	numeric	The organization group code
hq.city	character	The headquarters location (city only) for each vertex
hq.state	character	The headquarters location (two letter U.S. state abbreviation only) for each vertex
scale	numeric	The scale of operations code for each vertex
source.org.status	numeric	Indicator for each vertex whether it is also a source organization. If so, its unique source ID code.
type	numeric	A vector containing the organization type code for each vertex.

Table 3: Table of Organization Covariates

6 References

Butts, C. T., Acton, R. M., and Marcum, C. S. (2010). Interorganizational collaboration in the Hurricane Katrina response. *Journal of Social Structure*.

Lind, B. E., Tirado, M., Butts, C. T., and Petruscu-Prahova, M. (2008). Brokerage roles in disaster response: Organizational mediation in the wake of Hurricane Katrina. *International Journal of Emergency Management*, 5(1):75–99.

Tierney, K. (2002). Lessons learned from research on group and organizational response to disasters. Technical report.