Developing a Social Media Archive at ICPSR

Libby Hemphill
Director, Resource Center for Minority Data
SOMAR Goals

- Evaluation
- Replication
- Novel analysis
- **FAIR data principles**
SOMAR Audiences

- Sociotechnical researchers studying social media
- SBE researchers using social media data
- Social science methodologists
- Digital archiving and curation researchers
- Researchers in the future
SOMAR Challenges

• Technical infrastructure
• Ethical and legal infrastructure
• Metadata enhancements
• Adoption
# Ethical Considerations

**Table 4.** “How Would You Feel If a Tweet of Yours Was Used in a Research Study and . . .” (n = 268).

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Very uncomfortable</th>
<th>Somewhat uncomfortable</th>
<th>Neither uncomfortable nor comfortable</th>
<th>Somewhat comfortable</th>
<th>Very comfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td>. . . you were not informed at all?</td>
<td>35.1%</td>
<td>31.7%</td>
<td>16.4%</td>
<td>13.4%</td>
<td>3.4%</td>
</tr>
<tr>
<td>. . . you were informed about the use after the fact?</td>
<td>21.3%</td>
<td>29.1%</td>
<td>20.5%</td>
<td>22.0%</td>
<td>7.1%</td>
</tr>
<tr>
<td>. . . it was analyzed along with millions of other tweets?</td>
<td>2.6%</td>
<td>18.7%</td>
<td>25.5%</td>
<td>30.0%</td>
<td>23.2%</td>
</tr>
<tr>
<td>. . . it was analyzed along with only a few dozen tweets?</td>
<td>16.5%</td>
<td>30.3%</td>
<td>24.0%</td>
<td>20.2%</td>
<td>9.0%</td>
</tr>
<tr>
<td>. . . it was from your “protected” account?</td>
<td>54.9%</td>
<td>20.5%</td>
<td>13.8%</td>
<td>6.0%</td>
<td>4.9%</td>
</tr>
<tr>
<td>. . . it was a public tweet you had later deleted?</td>
<td>31.3%</td>
<td>32.5%</td>
<td>20.5%</td>
<td>10.4%</td>
<td>5.2%</td>
</tr>
<tr>
<td>. . . no human researchers read it, but it was analyzed by a computer program?</td>
<td>2.6%</td>
<td>14.3%</td>
<td>30.5%</td>
<td>32.3%</td>
<td>20.3%</td>
</tr>
<tr>
<td>. . . the human researchers read your tweet to analyze it?</td>
<td>9.7%</td>
<td>27.6%</td>
<td>25.0%</td>
<td>25.4%</td>
<td>12.3%</td>
</tr>
<tr>
<td>. . . the researchers also analyzed your public profile information, such as location and username?</td>
<td>32.2%</td>
<td>23.2%</td>
<td>21.0%</td>
<td>13.9%</td>
<td>9.7%</td>
</tr>
<tr>
<td>. . . the researchers did not have any of your additional profile information?</td>
<td>4.9%</td>
<td>15.4%</td>
<td>25.1%</td>
<td>34.1%</td>
<td>20.6%</td>
</tr>
<tr>
<td>. . . your tweet was quoted in a published research paper, attributed to your Twitter handle?</td>
<td>34.3%</td>
<td>21.6%</td>
<td>21.6%</td>
<td>13.1%</td>
<td>9.3%</td>
</tr>
<tr>
<td>. . . your tweet was quoted in a published research paper, attributed anonymously?</td>
<td>9.0%</td>
<td>16.8%</td>
<td>26.5%</td>
<td>28.4%</td>
<td>19.4%</td>
</tr>
</tbody>
</table>

Note. The shading was used to provide a visual cue about higher percentages.

---

**Social Media Ethics Framework:**

1. Does your research involve social media data?
   - Yes
     - Have you consulted the terms and conditions of the specific platform?
       - Yes
         - Have you consulted the relevant disciplinary, funding, legal or institutional guidelines?
           - Yes
             - Can the social media user reasonably expect to be observed by strangers?
               - Yes
                 - Are the research participants vulnerable? (i.e. children or vulnerable adults)
                   - No
                     - Is the subject matter sensitive?
                       - No
                         - Will the social media user by anonymised in published outputs?
                           - Yes
                             - Can you publish or share the dataset?
                               - Yes

---

Fiesler and Proferes 2018

Townsend and Wallace 2016
In [ ]:

```python
import matplotlib
import calendar
import codecs
import datetime
import json
import sys
import gzip
import string
import glob
import requests
import os
import numpy as np
```

**Twitter Crisis Analytics**

The following notebook walks us through a number of capabilities or common pieces of functionality one may want when analyzing Twitter following a crisis. We will start by defining information for a set of events for which we have data.

In [ ]:

```python
crisisInfo = {
    "boston": {
        "name": "Boston Marathon Bombing",
        "time": 1366051740,
        # Timestamp in seconds since 1/1/1970, UTC
        # 15 April 2013, 14:49 EDT -> 18:49 UTC
    }
}
```
SOMAR Open Questions

• What about content that’s integral but not native to the social media post (e.g., links, images, videos)?
• What are the right metadata enhancements?
• How should SOMAR fit/model data management practices?
• How should we connect to existing collections and tools?
• How should we sustain the enterprise?
SOMAR Next Steps

- SFM test bed on ICPSR infrastructure
- Tweet Sets test bed on ICPSR infrastructure
- Reddit harvester for SFM
- Automate rehydration loop
- Kibana exploration
- Jupyter connector to ElasticSearch
- URL expansion and web archiving
More Resources

- ICPSR Data Management and Curation
- FAIR principles for scientific data management
- Documenting the Now
- Social Feed Manager
- Tweet Sets
- GESIS Twitter Dataset example
- GETAR