Types, Granularities and Combinations of Geographic Objects in the Haiti Crisis Map

Iain Dillingham,[1] Jo Wood[2] and Jason Dykes[3]

[1,2,3]giCentre, School of Informatics, City University London, EC1V 0HB
[1]Tel. +44 (0)20 7040 0295 Fax +44 (0)20 7040 8584
Email: [1]iain.dillingham.1@city.ac.uk, [2]j.d.wood@city.ac.uk, [3]j.dykes@city.ac.uk

KEYWORDS: Uncertainty, Lineage, Ushahidi, Crisis information, Crowdsourcing

1. Introduction

The humanitarian community is reluctant to use social media when responding to crisis events, as the costs associated with inaccurate information are prohibitive (Coyle & Meier 2009). One solution is to crowdsource the verification process, although this clearly introduces further uncertainty into the information. As part of an ongoing programme of research into the geographic uncertainty associated with crowdsourced crisis information, we consider location descriptions from the Haiti Crisis Map, a dataset gathered in the wake of the 2010 earthquake in Haiti (Ushahidi 2009). Building on a previous study (Dillingham et al. 2012*a*) and using a named-entity recognition (NER) approach, we address three research questions that relate the types, granularities and combinations of geographic objects to their lineage and, as such, their geographic uncertainty.

2. Literature review

Several studies have explored location descriptions in datasets like the Haiti Crisis Map. For example, Wieczorek et al. (2004) and Guo et al. (2008) explore a collection of natural history records, and Doherty et al. (2011) a collection of historical search and rescue incident reports, with the aim of building geo-referencing methods that better account for geographic uncertainty. To better understand these methods, Guo et al. (2008) argue that a meaningful location description will contain a named place. This named place either will be the location, or will act (with a spatial relationship) as a reference to the location.

In a previous study (Dillingham et al. 2012*a*), we categorised location descriptions from the Haiti Crisis Map and found that locations were described in terms of named places more frequently than in terms of spatial relationships from named places; for example, as features, rather than as headings (e.g. 'south of'), offsets (e.g. '10km from') or 'between' or 'near to' spatial relationships from features. However, we were unable to explore the relationship between location descriptions and geographic uncertainty because we knew very little about the granularity of the named places, where granularity encompasses scale and definition (vagueness) (Jones et al. 2008).

To determine the granularity of the named places, it is necessary to annotate, either manually or automatically, location descriptions. Gelernter & Mushegian (2011) argue that automatic annotation has the advantage of scaleability. Consequently, we used an NER approach, which we discuss in Section 4.

3. Research questions (RQs)

1. What types of geographic objects are contained in location descriptions?
2. What granularities of geographic objects are contained in location descriptions?
3. How are types and granularities of geographic objects combined in location descriptions?

4. Methods

We wrote a small Java application to transform the Haiti subset of the GeoNames gazetteer (GeoNames 2012*b*) into the format used by the GATE/ANNIE gazetteer (Cunningham et al. 2002). This application retains unique, lower-case values of the `name`, `asciiname` and `alternatenames` attributes; performs a small number of character substitutions; and groups these values by their feature codes (GeoNames 2012*a*). We also manually created a second GATE/ANNIE gazetteer that contains common spatial relationships (e.g. 'south', 'between' and 'near to'). The entries in this gazetteer are based on our familiarity with location descriptions from the Haiti Crisis Map. In total, the two gazetteers contain 16,332 entries grouped by 92 feature codes. However, these entries are unevenly distributed amongst feature codes: for example, 'populated place' accounts for 62.20% of entries (10,159), whilst the next most frequent feature code, 'stream', accounts for only 7.84% (1,281).

We built a GATE application to annotate location descriptions using the two gazetteers. Although GATE can export to XML, for simplicity we preserved the original format of one location description on each line, with annotations encoded using an XML-like syntax (Figure 1). This syntax allows annotations to be nested; in the example, 'baie de henne' is annotated as 'bay', whilst 'henne' is also annotated as 'populated place'. Although this example is meaningful, others are not: 'port margot' and 'margot' are annotated as 'populated place' because both entries appear in the gazetteer. Consequently, we ignored nested annotations when calculating frequencies.

```
<BAY>baie de <PPL>henne</PPL></BAY>
<PPL>port <PPL>margot</PPL></PPL>
```

Figure 1: Two annotated location descriptions.

5. Preliminary results

The mean number of annotations for each location description is 1.639, with a range of 0 to 7. Figure 2 shows the frequency of feature codes. Figure 3 shows the frequency of adjacent feature codes. Table 1 is an index to feature codes.

6. Discussion

Although the most frequent feature codes refer to settlements (e.g. PPL, PPLC) (RQ1), settlements are likely to be over-represented and streets under-represented because many streets are numbered by their settlements but generally only settlements are entries in the gazetteer (e.g. Delmas, which is a settlement within Port-au-Prince, contains Delmas 50, which is a street). Similarly, airports are likely to be over-represented and capitals under-represented because many location descriptions use the abbreviation 'pap' to refer to 'Port-au-Prince' rather than to 'PAP', the IATA code for Port-au-Prince International Airport. Turning to frequencies of adjacent feature codes, capitals frequently follow populated places (PPL–PPLC, 334) (RQ3). Given that all 32 entities for capitals in the gazetteer refer to Port-au-Prince,

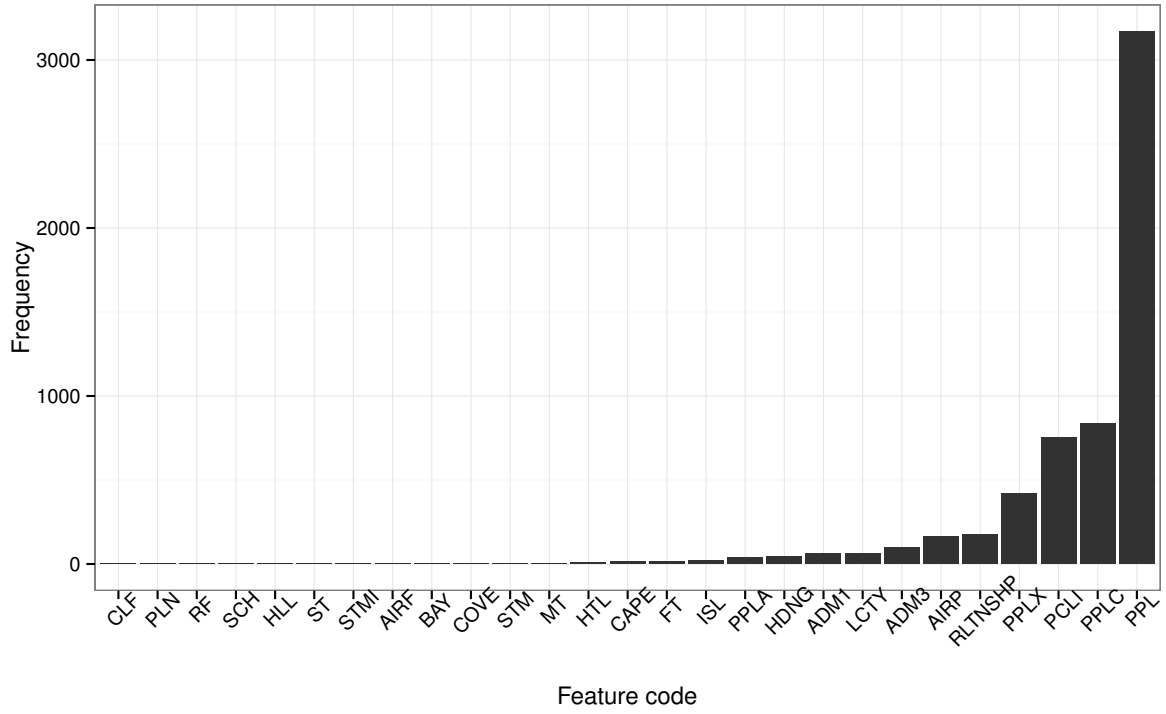

Figure 2: Frequency of feature codes.

these populated places are likely to be districts of, or streets in (see above), the Haitian capital. Further work is required to use JAPE, a GATE-based pattern language, to identify common street-like, airport-like and district-like patterns and improve the NER approach.

Coarser-granularity administrative areas (e.g. PPLC, PCLI) are more frequent than finer-granularity administrative areas (e.g. PPLA, ADM1) (RQ2). Indeed, political entities (PCLI, 753), of which all refer to Haiti, and capitals (PPLC, 835), of which all refer to Port-au-Prince, are the most frequent administrative areas (RQ1). Political entities tend to follow populated places (PPL–PCLI, 319) and capitals (PPLC–PCLI, 197) (RQ3). Given that the spatial and thematic foci of the Haiti Crisis Map are clear, these pairings cannot reduce place-name ambiguity. Instead, their existence suggests that redundancy characterises many location descriptions.

GeoNames defines a locality as "a minor area or place of unspecified or mixed character and indefinite boundaries" (GeoNames 2012*a*). The low frequency of localities (LCTY, 64) suggests that location descriptions tend not to contain *vague places*, or places that have vernacular names and vague spatial extents (Jones et al. 2008) (RC1). Finer-granularity named places (e.g. HTL) are infrequent (RQ2), as are spatial relationships (RLTNSHP, 178; HDNG, 46) (RC1), a result which supports our previous study (Dillingham et al. 2012*a*).

In summary, although there is scope to improve the NER approach, location descriptions tend to contain coarser-granularity administrative areas, especially settlements followed by references to Haiti. This suggests that redundancy characterises many location descriptions. Furthermore, location descriptions tend not to contain spatial relationships, which is surprising, given that so many buildings were destroyed by the earthquake (Inter-American Development Bank 2010). Consequently, we suggest that many location descriptions were edited to conform to a model of space that was itself highly uncertain. Although the original, unedited location descriptions are not available in the Haiti Crisis Map—and may never be available in future crisis maps, given the ephemeral nature of these 'mashups'—exploring patterns of types, granularities and combinations of geographic objects in location descriptions may allow us to infer their lineage, which is one aspect of their geographic uncertainty (MacEachren et al. 2005).

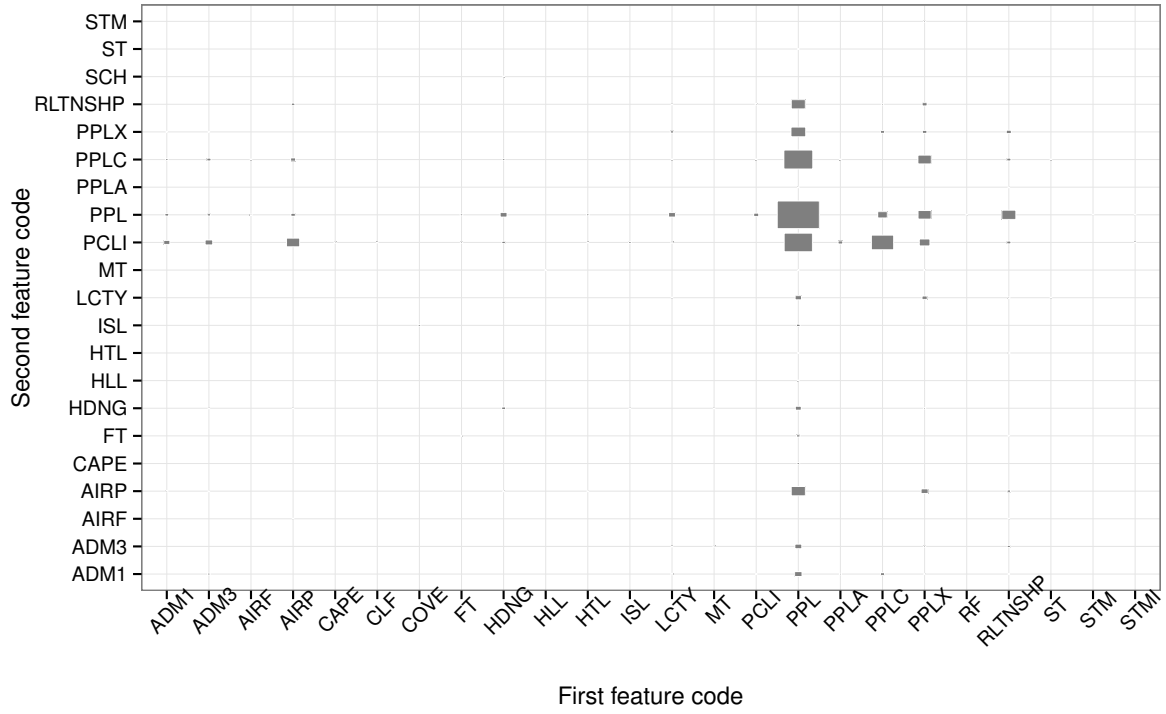

Figure 3: Frequency of adjacent feature codes.

7. Conclusion

We used an NER approach to explore the types, granularities and combinations of geographic objects contained in location descriptions from the Haiti Crisis Map. We argue that doing so may allow us to infer the lineage of location descriptions and highlight several characteristics that suggest that many location descriptions from the Haiti Crisis Map were edited. Although further work is required to improve the NER approach, we see the most interesting prospects for future research to be in the development of both analytical methods and tools that allow for the analysis of the characteristics of location descriptions over space and time (that is, to explore the effects of the uncertain model of space), and their application to location descriptions from alternative crisis maps. Consequently, we are developing visual analytics software (Dillingham et al. 2012*b*) that we hope to present in the near future.

8. Acknowledgements

We thank the Haiti Crisis Map for making their dataset publicly available. This work is funded by a doctoral studentship from City University London.

9. References

Coyle, D. & Meier, P. (2009), New technologies in emergencies and conflicts: The role of information and social networks, Technical report, UN Foundation–Vodafone Foundation Partnership, Washington DC, USA and London, UK. [Accessed 2 June 2011].
URL: *http://www.unfoundation.org/press-center/publications/new-technologies-emergencies-conflicts.html*

Feature code	Description
ADM1	first-order administrative division
ADM3	third-order administrative division
AIRF	airfield
AIRP	airport
BAY	bay
CAPE	cape
CLF	cliff(s)
COVE	cove(s)
FT	fort
HDNG	heading
HLL	hill
HTL	hotel
ISL	island
LCTY	locality
MT	mountain
PCLI	independent political entity
PLN	plain(s)
PPL	populated place
PPLA	seat of a first-order administrative division
PPLC	capital of a political entity
PPLX	section of populated place
RF	reef(s)
RLTNSHP	spatial relationship
SCH	school
ST	street
STM	stream
STMI	intermittent stream

Table 1: Index to feature codes (GeoNames 2012*a*).

Cunningham, H., Maynard, D., Bontcheva, K. & Tablan, V. (2002), GATE: A framework and graphical development environment for robust NLP tools and applications, *in* 'Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)', pp. 168–175. The 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, PA, USA, 7–12 July 2002.

Dillingham, I., Dykes, J. & Wood, J. (2012*a*), Characterising locality descriptions in crowdsourced crisis information, *in* 'Proceedings of the GISRUK 20th Annual Conference'. Lancaster University, Lancaster, UK, 11–13 April 2012.

Dillingham, I., Dykes, J. & Wood, J. (2012*b*), 'Exploring patterns of uncertainty in crowdsourced crisis information', Poster presented at the EuroVis Workshop on Visual Analytics (EuroVA). Vienna, Austria, 4–5 June 2012.

Doherty, P., Guo, Q., Liu, Y., Wieczorek, J. & Doke, J. (2011), 'Georeferencing incidents from locality descriptions and its applications: a case study from Yosemite National Park search and rescue', *Transactions in GIS* **15**(6), 775–793.

Gelernter, J. & Mushegian, N. (2011), 'Geo-parsing messages from microtext', *Transactions in GIS* **15**(6), 753–773.

GeoNames (2012*a*), 'GeoNames feature codes'. [Accessed 3rd October 2012].
URL: *http://www.geonames.org/export/codes.html*

GeoNames (2012*b*), 'GeoNames Haiti'. [Accessed 3rd October 2012].
URL: *http://download.geonames.org/export/dump/HT.zip*

Guo, Q., Liu, Y. & Wieczorek, J. (2008), 'Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach', *International Journal of Geographical Information Science* **22**(10), 1067–1090.

Inter-American Development Bank (2010), 'Haiti reconstruction cost may near $14 billion, IDB study shows'. [Accessed 13 November 2012].
URL: *http://www.iadb.org/en/news/webstories/2010-02-16/haiti-earthquake-reconstruction-could-hit-14-billion-idb,6528.html*

Jones, C. B., Purves, R. S., Clough, P. D. & Joho, H. (2008), 'Modelling vague places with knowledge from the web', *International Journal of Geographical Information Science* **22**(10), 1045–1065.

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M. & Hetzler, E. (2005), 'Visualizing geospatial information uncertainty: What we know and what we need to know', *Cartography and Geographic Information Science* **32**(3), 139–160.

Ushahidi (2009), 'Haiti Crisis Map'. [Accessed 19 October 2011].
URL: *http://haiti.ushahidi.com/*

Wieczorek, J., Guo, Q. & Hijmans, R. (2004), 'The point-radius method for georeferencing locality descriptions and calculating associated uncertainty', *International Journal of Geographical Information Science* **18**(8), 745–767.

10. Biography

Iain Dillingham is a PhD candidate at the giCentre, City University London, researching the geographic uncertainty associated with incident reports gathered in the wake of crisis events.

Dr Jo Wood is a Professor of Visual Analytics at the giCentre, City University London, with research interests in geovisualization and visual analytics of data with spatial and temporal components.

Dr Jason Dykes is a Professor of Visualization at the giCentre, City University London, undertaking applied and theoretical research in, around and between information visualization, interactive analytical cartography and human-centred design.