From text to information: legal and ethical issues in the geographic analysis of unstructured text

Ross Purves (Corresponding author) and Olga Koblet
Department of Geography, University of Zurich, Switzerland
ross.purves@geo.uzh.ch
olga.chesnokova@geo.uzh.ch

In the last decade numerous researchers have explored the use of unstructured text as a source of geographic information, for example exploring the use of vernacular placenames, spatial relationships and language, cross-linguistic variation, extracting opinions and sentiments with respect to urban areas and first-person landscape perception. Scientists and policy makers are interested in such sources since they potentially give access to very large samples of the use of geographic language in the wild, unencumbered by complex experimental protocols. As such, the creators of the data are typically not aware of such downstream analysis. Particularly interesting genres of writing found on the web include first-person narratives written in the form of blogs and travel reports, newspaper articles, and historic and modern literature. Using such data throws up a number of legal and ethical issues.

Firstly, while copyright is more or less clear for the latter two genres, in the case of travel blogs this is often unclear. In practice researchers often simply scrape content without paying regard to any attached legal conditions, even where these are explicitly stated (for example in the form of creative commons share-alike licences). Furthermore, copyright rules vary geographically, and become more unclear when results are derived from content accessed through search engines (for example, is copyright important if we extract toponyms used with near in a newspaper corpus?).

Secondly, even if copyright issues are clear, what ethical issues arise when analysing such data? Is it for example acceptable to extract negative sentiment, written about local events, which may in turn lead to a decrease in house prices? And if so, should we inform the writers of these texts that we are using them in such a way? How can we be reciprocal when we analyse texts to generate information – is it enough to simply inform the writers of our analysis, or should they also be acknowledged as part of the process of information production?

We do not claim to have answers to these questions, but through a set of concrete examples will illustrate the challenges that arise.