

## **GeoParsing: the Digitization and Historical Georeferencing of Text Documents**

**Stuart DUNN**

King's College London, UK

Many documents and metadata records refer to geographic locations in their text; and some 80% of material on the World Wide Web contains a direct or indirect reference to location. Yet a very large proportion of these are not tagged with full geographical coordinates (e.g. an OS National Grid Reference), or related to controlled vocabularies or gazetteers. Significant value is added to any research resource if location references can be isolated and used as an information structuring and retrieval tool in its own right. If the resource is consistently geo-tagged, this is possible. When metadata contains quantitative geographic information, resources can also be searched for using geographical queries, not just the "what", "when" and "who" queries of standard resource discovery interfaces. This frequently is not the case however: for example, large-scale text digitization projects do not provide comprehensive indices of toponyms or geographic features. This paper will report on an exemplar project that employed Natural Language Processing (NLP) to identify references to location in previously digitized texts (the Stormont Papers, proceedings of the Parliament of Northern Ireland between 1921 and 1973), convert them to the KML open standard supporting by platforms such as Google Earth, and present them in a spatio-temporal interface. It will identify the advance computational infrastructure needed for this experiment and what lessons have been learned from this exercise, identify what elements of geographic metadata should be added at the start of large text digitization projects, and outline research questions for the future.