Published paper
Clough, P., Pasley, R.C., Siersdorfer, S., San Pedro, J. and Sanderson, M. (2007) *Visualising the South Yorkshire floods of '07.* In: Proceedings of the 4th ACM Workshop on Geographical Information Retrieval : Workshop on Geographic Information Retrieval run at the 16th Conference on Information and Knowledge Management, November 09, 2007, Lisbon, Portugal. ACM .
http://dx.doi.org/10.1145/1316948.1316972

Visualising the South Yorkshire Floods of '07

Paul Clough, Rob Pasley, Stefan Siersdorfer, Jose San Pedro and Mark Sanderson

Department of Information Studies
University of Sheffield
Western Bank
Sheffield, UK

ABSTRACT
This paper describes initial work on developing an information system to gather, process and visualise various multimedia data sources related to the South Yorkshire (UK) floods of 2007. The work is part of the Memoir project which aims to investigate how technology can help people create and manage long-term personal memories. We are using maps to aggregate multimedia data and to stimulate remembering past events. The paper describes an initial prototype; challenges faced so far and planned future work.

Categories and Subject Descriptors
H.3.3 **[Information Search and Retrieval]**

General Terms
Algorithms, experimentation.

Keywords
Visualisation, data aggregation, web mining, Google Maps.

1. INTRODUCTION
The web is a vast dynamic repository of information capturing all aspects of human life [1, 2]. This information consists of various sources including news, blogs, wikis, all of which are generated increasingly in an interactive and collaborative way. Content exists in varying formats, media types, languages and varying quality making the use of such data often challenging. However, information found on the web often relates to place and such information is fairly accurate and up-to-date [3]. Chen et al. [4] also add that "the World Wide Web is the largest collection of geospatial data; a resource that goes almost unexploited."

This paper describes an application to exploit geospatial data: mapping the content of different online media to preserve and help people remember past events. Information relating to world events can be found distributed online, ranging from formal news reports to more personal (collaboratively-generated) content. Through map visualisation we collate this information to provide a single aggregated view.

To develop and test the application we have focused on the flooding in June 2007 that devastated large areas of South Yorkshire (UK). We plan to create an interactive website for members of the local community to generate a shared memory of the event and help individuals recall and share their personal experiences. The work is part of Memoir[1], an EU-funded project investigating the technology, ethics and psychology of storing and accessing a life-time of personal information.

GIR'07, November 9, 2007, Lisbon, Portugal.
ACM 978-1-59593-828-2/07/0011.

2. PROTOTYPE SYSTEM
The current prototype is being used to explore some of the technical issues surrounding automatically collecting, processing and visualising content. The system gathers data from multiple sources, extracts geo-references (e.g. location names), assigns spatial coordinates and visualises the results using Google Maps.

2.1 Gathering Content
For the current prototype system we have used three main sources of information: Flickr, YouTube and BBC News Online. These provide varying information about events such as multiple media, and different reporting angles. Table 1 summarises the content used in the current prototype comprising a total of 5,505 files. The *total* column indicates the number of files gathered for each source and the *used* column shows the number of files from which we managed to extract geo-information.

Table 1. Content gathered for the current prototype

Source	Media	Total	Used
BBC News	Text	909	797
Flickr	Image	4,482	1,273
YouTube	Video	114	50

2.1.1 BBC News
This content was easiest to gather: searches for the query "flood" were submitted to the BBC News[2] website. This site provided a range of multimedia content and 909 news articles were collected which included the following metadata: title, date, story text, photo URLs, photo captions.

2.1.2 Flickr
Flickr is a large-scale photo-sharing website providing a wealth of personal content. The general query ("flood") was used together with a date range restriction (from 22/06/07). Data was gathered using the Flickr API and the following metadata was extracted: photo title, description, tags, owner, temporal information, spatial coordinates (if available) and user comments. Photos that already

[1] http://homepage.mac.com/jsanpedro/Memoir/index.html
[2] http://news.bbc.co.uk/

had spatial coordinates (793 out of the total 4,482) were plotted directly on a map.

2.1.3 YouTube

YouTube is similar to Flickr, but serves videos rather than still images. However, crawling YouTube is more challenging than Flickr because: (1) the YouTube API is more limited, and (2) standardised metadata does not exist for video content (unlike images). The crawling process was divided into two phases: (1) querying YouTube with "yorkshire flood" (without the API) and for each video result extracting the identifier of each video and gathering the following metadata (using the API): uploading author/user, title and name of video, rating (given by YouTube users), assigned tags, video description, creation/upload date, duration and collaboratively-generated comments.

2.2 Processing Content

Content was converted into a standardized XML format for further processing. The General Architecture for Text Engineering (GATE) system [5] was used to extract geographical information (as used in previous work [6]). The Ordnance Survey (OS) 50k Gazetteer and Locator resources were used to assist extraction of locations and assignment of spatial coordinates. The 50k Gazetteer lists places that appear on the 1:50000 OS maps (e.g. populated places and certain landmarks), together with their co-ordinate points. OS Locator was converted into a gazetteer containing UK street names (with co-ordinates). Locations with multiple coordinates were resolved to the referent closest to the centroid for the first place name mentioned, e.g. if Sheffield, Chesterfield and Doncaster appear in a document then it is likely they refer to places in South Yorkshire (the 50k Gazetteer contains a very small place in Cornwall named Sheffield but this is not close to places called Chesterfield or Doncaster). A stopword list was used to remove false hits such as 'flood', 'the' and 'hey' which appeared in the gazetteers but used frequently in a non-geographical sense (e.g. in the caption "Hey, have a look at my flood pics"). Certain parsing errors also occurred; tags like 'sheffieldflood' were not recognised.

2.3 Mapping Content

Content is combined through visualisation using Google Maps. Pins are labelled to indicate the content, and coloured to show content parsed by GATE and content containing spatial data suitable for mapping directly (some Flickr data). Users click on the pins to show the media and associated metadata.

3. DISCUSSION

The prototype gathers a range of media; content such as video has features that make it especially interesting for sharing memories and experiences. The inherent temporal dimension tends to convey messages of a more complex nature than those achieved with still images. The audio stream enclosed in the video also has a role and in combination with the sequence of images, creating a context that effectively narrates events. Technologies such as mobile phones and digital cameras are accessible to more people: it was noticeable that soon after flooding began in South Yorkshire, both Flickr and YouTube were being populated with relevant media. Although many applications exist which gather content from various sources and visualize this as layers on a map (e.g. ononemap.com), the main focus of this work is to gather, process and present a variety of *personal* content from all types of media.

4. SUMMARY & FUTURE WORK

Our long-term objective is to develop techniques to gather and aggregate information on different events, providing enhanced user interfaces for visualizing geographic and temporal contexts. The potential of exploiting geospatial data from online content for helping to record local (and national) events is huge. However, it is obvious that automatically gathering and processing this content is non-trivial and requires processing for each resource. For example, when considering text types, we have seen that geo-coding performance is better on BBC news reports indicating the need for further improvements in the geo-coding methods for metadata; whereas the structure of Flickr and YouTube data makes extracting information simpler than from unstructured text.

Ongoing and future work includes: developing techniques to extract relevant information from various data sources and standardising their representation, applying statistical methods on context information (such as captions or tags) to improve the automatic assignment of geo-tags, using relevance and diversity measures to select the most representative samples of content (if the number of photos, videos or other items becomes too large for a map), designing enhanced interfaces and animations to depict temporal developments of an event, exploiting existing spatial information for geo-coding other sources (e.g. exploiting previously geo-coded Flickr images to other media through matching keywords), and providing an integrated and unified view from different information sources, such as Web portals, news repositories, multimedia sharing systems, blogs and newsgroups.

ACKNOWLEDGMENTS

Work partially supported by the EPSRC and Ordnance Survey (CASE/CAN/06/67) and EU-funded Memoir project.

REFERENCES

[1] Chakrabarti, S. (2002) Mining the Web: Analysis of Hypertext and Semi Structured Data, Morgan Kaufmann.

[2] Lin, J. and Halavais, A. (2006) Geographical Distribution of Blogs in the United States, *Webology*, **3** (4).

[3] Himmelstein, M. 2005. Local Search: The Internet Is the Yellow Pages. *Computer* **38**, 2 (Feb. 2005), 26-34.

[4] Chen, Y., Suel, T., and Markowetz, A. 2006. Efficient query processing in geographic web search engines. In *Proceedings of the 2006 ACM SIGMOD international Conference on Management of Data* (Chicago, IL, USA, June 27 - 29, 2006). ACM Press, New York, NY, 277-28.

[5] Cunningham, H. (2002), 'GATE, a General Architecture for Text Engineering', *Computers and the Humanities* **36**, 223-254.

[6] Clough, P. (2005), Extracting metadata for spatially-aware information retrieval on the internet, In *Proceedings of the 2005 workshop on Geographic information retrieval*, ACM Press, New York, NY, USA, pp. 25-30.