Uncertainty in Historical GIS

L. Mamani Sanchez[1] and M. Bertolotto[1]

[1]School of Computer Science, University College Dublin, Ireland
Email: {liliana.mamanisanchez, michela.bertolotto}@ucd.ie

Abstract

In developing a system for storing and manipulating historical data related to the visitors of a library in the city of Dublin (Ireland) in the time period 1826-1926, we faced several challenges linked to the uncertainty associated with such data. These were due to several different aspects including lack of consistency in the recording of visits to the library over time, interpretation of different handwriting styles, varying levels of granularity in both the spatial (address of visitors) and temporal (date of visit) dimension of the records. The focus of our work is the development of a model for properly representing all aspects of uncertainty within a system to store and query this data.

Keywords: Uncertainty, Historical GIS.

1. Introduction

Our research focusses on uncertainty issues related to historical geographic information systems. A prototype is being developed in the context of a digital humanities research project. By aligning humanities research and GIS technology, the project seeks to explore and reconstruct the role and scope of Marsh's Library as a knowledge node in Dublin's book and reading culture between 1826 and 1926, and as such, to transform academic and popular understanding of Dublin's cultural and literary history in the nineteenth and early twentieth centuries.

Our aim is to build an open source historical GIS including a database of readers and readership, and a user-friendly visualisation interface which will aid the analysis of readership information by academic specialists. Both will help to provide a rich seam of evidence as to the broader evolution of Dublin's intellectual and literary history. Also, it will serve as means to make archival data and the results of an overarching analysis available to the general public.

Challenges related to the digitisation of handwritten data are relevant to our research, particularly because of uncertainty, non-uniformity and incompleteness present in our dataset. An objective of our research is to make the system easy to adapt to other archival data (e.g. data available in other libraries). Such a system will be of relevance to experts and professionals who curate and research data of a similar nature, where properly representing and dealing with different types of uncertainty is a major challenge.

2. Data and methods

Marsh's Library houses around 25,000 rare books and 300 volumes of manuscripts, many dealing with the sixteenth and seventeenth centuries. It was the only public library in Dublin for the first century and a half of its existence.

The Library's archival records consist of a series of "Visitors' Books" into which readers signed their names and addresses. For instance, the first people listed in the first Visitors' Book are Thomas Shaw of 19 Moore Street, John McCready of 43 Bride Street and Edward Burroughs of 13 Peter Street. Today, these addresses are in areas of socio-economic disadvantage. Therefore, these records provide a unique opportunity to chart the shifting socio-economic and cultural geography of Dublin.

Entries in the visitors' books are not uniformly structured and this poses an important challenge. For example, in some cases only the visitors' names are listed, while in other cases their address, profession as well as their referees' names is entered. Being handwritten, digitising entries presents the challenge of interpreting multiple handwriting styles. For difficult cases, interpretations are discussed with an additional human reviewer. Despite this procedure, digitisation cannot always be done accurately (if at all) and therefore, the system needs to record the level of confidence with which data was entered. Furthermore, some historical addresses have now changed, so a mapping between new and old addresses is necessary.

In addressing these challenges, we have developed a prototype system comprising a spatial database to facilitate the development of spatial queries such as display of the number of readers by address, metropolitan district, and, where known, their reading material. We have deployed the following technologies: PostGIS for the database, OpenStreetMaps (OSM) as the source of the most up-to-date spatial maps for the city of Dublin and Python and Javascript-based technologies for the prototype's backend and GUI. The GUI visualization comprises spatial, temporal and thematic descriptions which are shown as an output of two kinds of querying tasks: one driven by the user's specification of querying arguments, and another one driven by the concept of timeline.

3. Uncertainty in Spatio-Temporal data

There is a body of literature on uncertainty in spatio-temporal data. In particular, our approach is inspired by the work of Malizia (2013). He defines three types of inaccuracies: 1) locational, 2) incompleteness, and 3) temporal. Uncertainty in our data comes from these three different sources, and presents additional characteristics.

In relation to locational inaccuracy, Malizia discusses the issue of "success" in geocoding. A geocoding process of a set of addresses will be more successful if the percentage of addresses that can be associated with geographic coordinates is high. To achieve this high percentage, a loosening of conditions may be required to avail of trade-offs between match rate and locational accuracy. Therefore, the positional accuracy of those coordinates will likely decrease. He mentions three reasons geocodes are not found for a specific address: a) misspelled or abbreviated addresses; b) records which include alternative identifiers to address (eg. PO box numbers); or (3) missing street information in the reference file.

In our dataset, addresses mainly comprise house number, street names, county, and country. Therefore, if any of these items was left out, this may lead to an incorrect or less accurate geocoding. For instance, historically a college building such as Trinity College Dublin comprised apartments with a distinctive numeration. Today, the college layout has changed and it is not possible to find the corresponding coordinates for "25, Trinity College Dublin, Dublin, Ireland" using a crowdsourced geocoder.

Our case study potentially presents two additional reasons for missed geocodes: undecipherable sources and names changes. Undecipherable sources relate to handwriting, so the best a transcriber can do is providing a wild or informed guess for the address. In order to capture this aspect, in our model we record a level of confidence with which the data was digitised.

Change in street names is more difficult to track as it may be costly and unreliable. In a set of records, it may even cause ambiguity and therefore inaccuracies in geocoding. For instance, the current 'Tara Street' in Dublin city was called 'George's Street' until 1855. Currently, geocoding by 'George's Street' returns completely different coordinates. There is a lack of studies about the effect of inaccuracies in the temporal dimension. In our case, inaccuracy in temporal data is closely intertwined with the issue of changing names. Not many structured sources that help us to refine our dataset with temporal information exist. For example the information above "George's Street" being called so until 1855 was extracted from prose describing history of Dublin Street names (Clerkin, 2001). However, there is no compact and complete resource or easy-to-access literature where this kind of information is recorded. In order to overcome this problem, we built a table of equivalences of street names that are associated to the same geographic coordinates together with the range of dates when a street was officially called a specific name.

This attempt to tidy historical data does not address the issue of changing geography such as the change of city maps (blueprints), disappeared streets, or modification of streets. An example of this is Sean Macdermott Street Lower and Killarney Street which together used to be called Gloucester Street at some stage and earlier Great Martin's Lane. This information also has to be recorded accordingly.

As for temporal inaccuracies, some come from the heterogeneous granularity with which dates were recorded. Some events are associated to specific dates, while other ones are linked to ranges of dates, where it is difficult to know for sure the exact date. However, in our data, as visits to the library were recorded in temporal order, we can establish relations between individual records. Therefore we link a visit which has an unknown date to the range determined by the last recorded date of a previous visit and the next recorded date of a subsequent visit. Temporal queries such as "Who visited the library on March 23rd, 1829?" for which there is no exact result will return a set of visits which might have occurred on such a date (with the associated range).

4. Discussion and Future Work

So far our work has focused on defining meaningful descriptions of uncertainty in our dataset. In particular, we deal with different levels of confidence in the interpretation of handwritten records as well as different levels of granularity/completeness in the recording of spatial and temporal aspects of the data.

We are researching reliable methods that deal with this uncertainty and methods to keep track of geographic information changes as part of a historical knowledge repository. While we have already addressed the change of street names over time in the city of Dublin, even if a street name was officially changed to a new one, people may query by an old name. This suggests another dimension of fuzzy queries for spatial data which needs to be further investigated.

Our future work will include processing of complex queries over uncertain data. Another important development will be the creation of advanced visualisations based on space-time cubes and timelines (Bach et al. 2014; Ding et al. 2016; Gautier et al. 2016) that take into account the aspect of uncertainty. Visual inspection will also potentially help identify inaccuracies in the data. Not many systems for spatio-temporal data visualisation provide a high level of interactivity that allows domain users to analyse their data in detail and discover unknown patterns. Therefore, we aim at developing a system with an interactive map-based interface that allows users (particularly domain experts) to investigate non-obvious details and relationships in the data.

4. Acknowledgements

This research was supported by an Irish Research Council grant to the project "Mapping readers and readership in Dublin, 1826-1926: a new cultural geography".

5. References

Bach, B., Pietriga, E., Fekete, J. 2014. Visualizing dynamic networks with matrix cubes. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14).* New York, USA: ACM, pp.877-886.

Clerkin, P. 2001. *Dublin Street Names*. Dublin: Gill & Macmillan.

Ding, L., Krisp, J.M., Augsburg, U., Meng, L., Xiao, G., Keler, A. 2016. Visual exploration of multivariate movement events in space-time cube. In: *Proceedings of the 19th AGILE International Conference on Geographic Information Science-Geospatial Data in a Changing World*, *14-16 June 2016, Helsinki, Finland*. [Online]. Dresden: AGILE [Accessed 07 June 2017]. Available from: https://agile-online.org/conference_paper/cds/agile_2016/shortpapers/127_Paper_in_PDF.pdf

Gautier, J., Davoine, P.-A., Cunty, C., Lyon II, I.R.G. 2016. Helical time representation to visualize return-periods of spatio-temporal events. In: *Proceedings of the 19th AGILE International Conference on Geographic Information Science-Geospatial Data in a Changing World*, *14-16 June 2016, Helsinki, Finland*. [Online]. Dresden: AGILE [Accessed 07 June 2017]. Available from: https://agile-online.org/conference_paper/cds/agile_2016/shortpapers/122_Paper_in_PDF.pdf

Malizia, N. 2013. The Effect of Data Inaccuracy on Tests of Space-Time Interaction. *Transactions in GIS.* **17**, pp.426–451.