AddressingHistory:

a Web2.0 community engagement tool and API

by Stuart Macdonald[1]

EDINA

Abstract

This paper will chart the development and delivery of a Web 2.0 community engagement tool and Application Programming Interface (API) developed at the EDINA in partnership with the National Library of Scotland, as part of the JISC Digitisation and e-Content Programme. Such a tool enables members of the community, both within and without academia (particularly local history groups and genealogists), to enhance and combine data from digitised historical Scottish Post Office Directories (PODs) with contemporaneous large-scale historical maps.
The paper discusses the background to Post Office Directories and the corresponding geo-referenced historic maps for Scotland, the technical platforms deployed including sustainable software components, and web applications and services. It also examines issues surrounding user generated content (UGC) created by the community such as mediation, validation and cross-checking, and the use of social media amplification for community engagement and future directions.
To conclude, the paper argues that the success of online crowdsourcing tools such as the one developed for this project will ultimately be measured by continual and extended use within the wider community.

Introduction

Post Office Directories, precursors to modern day Yellow Pages, offer a fine-grained spatial and temporal view on important social, economic and demographic circumstances. They emerged during the late seventeenth century to meet the demand for accurate information about trade and industry due to the expansion of commerce during this period. They were published more frequently than the census and generally had information about local facilities, institutions and associations, listings for private residents, traders, trades and professions, sometimes details of important people, and advertisements.

The ways in which publishers collected data varied considerably. Some obtained information by personal canvassing and combined the results with existing trade listings. Other publishers simply asked people to send in their names together with a small payment if they wanted to be included in the directory.

By the early nineteenth century methods of compilation were more organised. In part, this reflected the growing links between directories and the Post Office. Many postal officials turned their hand to directory publishing as a means of both aiding their work and augmenting their income. Information was collected by letter carriers, who circulated forms during their postal rounds, and also delivered the finished directory on commission.

For Scotland there are at least 750 Post Office Directories spanning the period 1770 – 1912. The NLS are in the process of scanning using Optical Character Recognition (OCR) techniques and publishing this historic collection in conjunction with the non-profit Internet Archive.

During the 6 month project period the AddressingHistory 'crowdsourcing' tool focussed on three volumes (1784-5; 1865; 1905-6) of the Edinburgh digitised PODs and maps

from the same periods. However the specifications were such as to accommodate the full Scottish collection as and when they become available. The Web 2.0 interface and back-end storage solutions were built to be both scalable and as far as was practicable, self-standing so that multiple independent instances can be supported and customised for different audiences.

The Edinburgh Directories themselves are a unique and reliable collection of street, commercial, trades, law, court, parliamentary and postal information relating to the city of Edinburgh. They also provide a wealth of detailed information regarding residential names, occupations and addresses and include maps of both Edinburgh and Leith indicating trade and residential origins and development.

One significant deficiency of this collection, which the AddressingHistory online tool aims to redress by 'crowd sourcing', is that the addresses are not geo-referenced. Geo-referencing make possible explicit spatial search and discovery, whilst permitting a map based metaphor to be used in the exploration and visualisation of the resource. E.g. the historic distribution of shipwrights in Edinburgh can be plotted on a base map or the map itself can be used to explore the spatial distribution of selected phenomena (and their variation over time). Similarly, personalised maps illustrating family histories, maps tracking changes in local communities, and maps linking to other digitised materials such as census records and geo-referenced images, and historical addresses could all be explored through use of the Application Programming Interface (API).

The National Library of Scotland's Map Library is one of the ten largest in the world; as the Library of the Faculty of Advocates from 1689, maps of Edinburgh were actively collected; as a Copyright Library, the collections are particularly strong in the printed mapping of Scotland. Since 1998, NLS Map Library has scanned over 20,000 historical maps of Scotland, including over 500 of Edinburgh and its environs.

It is the pre-existence of large scale geo-referenced and contemporaneous maps against which the historic post office directories were contextualised that allows manual (geo)referencing down to individual house address level to be accomplished. This is achieved by simply moving a pin on the map; i.e. the map is the mechanism through which the geo-reference is allocated by the user to a particular POD entry.

To assist the geo-referencing exercise, addresses from each of the directories were parsed using Google's geocoding software[2] in order to assign a geo-reference. There were issues with the legibility of the OCR'd text (especially for the POD for the earlier period) in addition to period addresses no longer being in existence or having suffered name changes. Thus within the interface a ranking mechanism makes explicit the relative 'accuracy' of the geo-coded content.

The user interface to the tool and associated API is intuitive and easy-to-use to encourage researchers, local historians, genealogists and members of the wider community from across the age spectrum to discover, explore and contribute to rich records of social history and to create their own related maps and data sets for both academic and personal research. These were also developed to be sympathetic to tools developed by related projects including Visualising Urban Geographies (VUG), an online resource developing new insights into the spatial character and historical development of Edinburgh (http://geo.nls.uk/urbhist/).

Technologies

Overview

The AddressingHistory tool and API comprises several software components, each built with resilience and sustainability in mind. Open Source software was chosen in several instances, allowing for great flexibility and a feature-rich application, whilst containing costs. AddressingHistory is built as a typical 3-tier web application. For the user-facing client presentation component, JISC recommended standards such as XHTML, CSS and other relevant W3C web standards (i.e. images etc) were employed. The web interface is supported by mainstream browsers and OGC standards including the Web Map Service (WMS) Interface Standard and OpenLayers were used for web mapping components.

An API is available, allowing access to the raw data via multiple output formats. It is accessible via a RESTful web service

Development

The project followed best practice for technical development, making extensive use of a number of common and well-established libraries including the Java SDK, Spring MCV framework and the jQuery Javascript libraries. Unit testing was performed via the jUnit libraries.

Development initially began by scoping the application's requirements, designing a database structure to store the information contained in the Post Office Directories in conjunction with pre-processing and data-loading software. The structural interpretation and translation of the varied content from three eras of directories proved to be a time consuming exercise. The directory data was processed, and additional metadata such as the locations of addresses were added to the database.

The API, following JISC recommendations for API Good Practice[3] was designed to allow access to the raw data using a number of HTTP GET procedure queries including a parameter which allows web developers to specify the format (JSON, KML or TXT) they want the result returned in.

The client application was built upon the API, featuring web based mapping. To the OpenLayers mapping, we added a collection of historical maps from NLS, contemporary to the three Post Office Directories of interest. A user registration system, facilities to edit the stored data and suggest specific changes were added towards the end of the development, together with various enhancements – including a view to the original scanned directory pages.

All components of the web accessible service and API are hosted via a Solaris 10 virtual container, together with an established PostgreSQL database, hosted at EDINA. Throughout the project, the source code, tests and configuration files were stored in a Subversion version controlled repository. Software builds and releases were automated via the Apache Maven software project management tool. Documentation is stored in a shared repository.

User Generated Content

The AddressingHistory project raised a number of issues regarding user generated content (UGC) created by the community such as mediation, validation and cross-checking of UGC. At present the AddressingHistory team retain the option to check UGC and will do so on a periodic basis. As part of a sustainability plan it is envisaged that once community participation reaches a certain level an 'engaged

user group' comprising active members of the user community may volunteer to conduct validation and cross-checking of UGC through a devolved mediation process.

A logging facility has been installed in order to identify inappropriate behaviour (e.g. spam) or inaccurate UGC. A registered user can be contacted in order to justify behaviour. Potentially the user or more accurately the username can be prevented from editing further content.

The AddressingHistory backend database has been designed so that the original database and that containing the UGC are maintained as separate instances thus allowing inaccurate or inappropriate user generated content to be removed from the database.

Social Media

A key element in determining the success of the project was the establishment of a mechanism whereby the 'crowd' could contribute to the creation of a fully geo-coded version of the digitised directories. In part an avenue through which such community engagement could be realised was through working with Edinburgh Beltane – a national co-ordinating centre for public engagement and with the University of Edinburgh College of Humanities and Social Sciences Knowledge Transfer Office. Social media channels were also deployed to engage the public, to develop links within the local and family history communities, and to act as a vehicle to expose the tool and API to a wider audience. The following section describes both method and mechanism used to engender public collaboration and community engagement

Building and Developing Community Connections

At the outset of the project an information page was created on the EDINA website[4] and later updated to connect to additional AddressingHistory presences. A WordPress blog[5] , was deployed as the key space for communicating and engaging with interested members of our target audiences.

Twitter was an unexpectedly useful space for the project (via the @ addresshistory account) and a Facebook page[6] was also created for AddressingHistory for sharing short updates, useful links and to encourage viral sharing and recommendation.

Community Engagement

From the outset the project team encouraged blogging and discussion of the project and the Project Officer proactively sought out potential contacts, followers, and bloggers and responded to any comments on the project to ensure that mentions were complemented with links to the website and that questions were responded to.

AddressingHistory benefited from pre-existing genealogy and local history blogging and online communities, receiving regular mentions and links from a wide variety of sites and discussion boards[7] . The social media and web presences helped reach out to many interested parties[8]. However outreach activities such as events, presentations, and print communication were also instrumental in exposing the project to a wider audience.

Ongoing Activity

As a longer term strategy we intend to maintain where practicable blog activity, Facebook and Twitter presences. A mailing list has been set up to ensure we can remain in contact with those interested in AddressingHistory developments and a Google group has been established aimed at users interested in using the AddressingHistory API for their own websites, projects, or mashups.

Future Directions

Features and Functionality

AddressingHistory was an ambitious project from the outset, covering a range of technologies and featuring several disparate problems. The initial processing of data extracted from the historical directories through OCR, presents a unique challenge in terms of data errors and lack of structure.

In spring 2011[9], as part of a second phase of development made possible through further funding, the AddressingHistory team will investigate and further streamline data pre-processing and loading processes with a view to providing 'cleaner' output. In conjunction with this, it will add further content (for other areas of Scotland) to broaden the user community and subsequent utility of the tool and API, as well as incorporating computer-generated metadata to POD entries such as categorising places and professions, or extracting multiple addresses.

Interesting further work may also involve investigating how best to capitalise on the social mechanics of AddressingHistory. This unique application offers opportunities for game-mechanics, awarding users for their crowd sourced information and challenging users to contribute.

Another avenue of development under consideration is the inclusion of facilities to upload and attach geo-referenced content such as images, census records, videos and sound files to AddressingHistory entries in the database. In this way, the directories would be extended to include photos of people, buildings, landmarks thus enriching the resource and broadening both utility and appeal.

Learning and Teaching

The AddressingHistory project team have met with representatives from Glow (a national intranet for education hosted by the Scottish Government[10]) with a view to using AddressingHistory as a means to create learning and teaching materials by Glow subject specialists for school pupils both within Edinburgh and beyond. Materials developed may have resonance with the recently launched *Digimap for Schools*[11], an online mapping service for use by teachers and pupils in schools hosted by EDINA

Linked Open Data

> "*Great work! but (cries) this *really* should be done with Linked Data and RDF, endless scope and endless data!*"
> - Comment received on the AddressingHistory Blog

The POD data has been processed and structured in such a manner that every person entry in the directories has a unique identifier in the AddressingHistory XML database. Each entry is accessible via the API through a URI - in much the same way that a unique place in Geonames (an open geographical database) is referenced by a unique URI. As the RDF output from GeoNames gives you the data in an XML document by using a schema of tags defined by the GeoNames ontology/vocabulary, AddressingHistory could theoretically provide output that included some place information using said GeoNames tags (i.e. you get a result for a person who lives in Leith and the result also links to information describing the spatial footprint, population, economy, demography of Leith). Or, indeed, 'celebrities' present within the PODs could be linked to DBpedia entries using the 'sameAs' tag,

which declares a link between two resources that describe the same real-world thing.

Context versus Content

"A major feature of this project is the offer of maps, and maps which enable the user to explore and present historical information spatially. The outcome is visually attractive and exciting. There is a danger that the fun of producing the map acts as a barrier to thinking about what is happening."

*- Quote from Professor Robert Mor*ris at the Launch Event

Professor Robert Morris, Emeritus Professor of Social and Economic History at the University of Edinburgh who provided the introductory presentation at the AddressingHistory launch[12] had reservations regarding context versus content. He indicated that, where applicable, explanatory notes providing information about background, construct and content of the original directory listings should be made explicit. In addition underlying assumptions and rules about both the structure of the processed data and the translation of the structured data into a consumable and interactive format should be made clear.

This has in part been addressed by inclusion within the interface of a range of Help documents (including a Post Office Directory Guide and People, Place & Profession Search Guides), an API Guide and Frequently Asked Questions. As part of any future development, use cases and contextual essays (such those available from the Statistical Accounts for Scotland[13]) should be considered.

Two videos which help to demonstrate context were created for the launch and shared via Vimeo . One video discussed the background to the PODS[14] explaining their usefulness to researchers (amateur and professional), the second video, featured on our Facebook page on the launch day, explains the POD digitisation process[15].

Sustainability

In accordance with the project plan the AddressingHistory project partners are committed to supporting the resource for a minimum of one year whilst it gathers community traction. During this time consideration will be made to the processes necessary for ongoing dissemination, community take-up of the deliverables and their adoption by the community. AddressingHistory aim to achieve this through those social media channels established as part of the project and an on-going relationship with Edinburgh Beltane[16] and, in turn, to appropriate organizations engaged in local and family history projects. Given the broad applicability of the resource it is envisaged that a range of communities may be interested in the longer term curation and continuance of the project tools e.g. the Open Street map community has an active user base interested in both contemporary and historical addresses. It is also anticipated that the active involvement of 'engaged users' throughout the project and beyond will provide direction on longer term sustainability issues.

Project partners will evaluate possible business models of sustainability based on levels of demand provided they remain consistent with the underlying open philosophy e.g. revenue generation through an online donations facility, subscription model (e.g. per annum, per month, per use), a 'freemium model' (e.g. free API download of a certain number of records with payment being required for further downloads), or academic advertising.

Conclusion

AddressingHistory was an ambitious project which combined a range of technologies from data processing and database design, to Web 2.0 and web mapping services. Much was achieved within the relatively short project in terms of public engagement and amplification through social media facilities and channels, and the delivery of a robust and scalable website and API capable of empowering the 'crowd' with the facility to search and edit geo-referenced content from the Scottish Post Office Directories and digitised historic maps from the same era.

However, gauging the success of the project goes beyond the delivery of engaging and innovative online tools. It will ultimately be measured by continual and extended use within the wider community[17].

Notes

1. This paper was shown as a poster presentation at the IASSIST conference 2010 at Cornell University. Stuart Macdonald is the AddressingHistory Project Manager at EDINA & Data Library, University of Edinburgh. Email: stuart.macdonald@ed.ac.uk
2. http://code.google.com/apis/maps/documentation/javascript/v2/services.html#Geocoding
3. http://ie-repository.jisc.ac.uk/344/
4. http://edina.ac.uk/projects/addressinghistory_summary.html
5. http://addressinghistory.blogs.edina.ac.uk/
6. http://www.facebook.com/AddressingHistory/
7. E.g. Clan MacLea/Livingstone (http://clanlivingstone.info/forum/viewtopic.php?f=5&t=1084&p=9800&hilit=addressinghistory#p9800),

Rootschat (http://www.rootschat.com/forum/index.php?topic=496764)

8. Indeed many of the social media monitoring techniques that were trialled on AddressingHistory are now successfully being used to better monitor social media mentions of other EDINA projects and services.
9. At time of publication (February 2012) phase 2 of the project is nearing completion. Work focused on streamlining the geo-parsing, and extending geographic and temporal coverage of Post Office Directories within the online tool. An AddressingHistory Augmented Reality Application will also be available shortly.
10. http://www.ltscotland.org.uk/usingglowandict/glow/
11. http://digimapforschools.edina.ac.uk
12. http://tinyurl.com/375czrb
13. http://edina.ac.uk/stat-acc-scot/reading/
14. http://vimeo.com/16902845
15. http://vimeo.com/16906333
16. The partnership cultivated between AddressingHistory and Edinburgh Research & Innovation and Edinburgh Beltane has initiated ongoing communications between EDINA and both organisations with a view to enhancing community engagement from broader service level perspectives.
17. Note: A free to access index for the Glasgow Post Office Directories from 1783-1911 is now available - http://bizdirs.from-mt.com/glasgow/

MAG