Article:
Byrne, K. (2014) Event Mining in Our Rural Past. Working Papers of the Communities & Culture Network+, 3. ISSN: 2052-7268

Event Mining in Our Rural Past

Kate Byrne
University of Edinburgh
March 2014

Communities and Culture Network+
Seed Project RE4500753263
Final Report

1. Executive Summary

We are living through an Information Revolution, with as much data being made available over the internet in the past few years as over the entire course of history beforehand. Paradoxically, it is arguable that our most valuable data, which resides in the world's great libraries and archives and distills the best products of human endeavour, is itself amongst the least easily accessible. New data that is added to the Web by the gigabyte daily is generally designed for internet access, whereas our historical archives were designed for an earlier age and usually for a specialist, professional audience. This short project was a practical exploration of automated software tools that can extract structure from unstructured archival text and hence make it more easily searchable.

The data that is the subject of this project is from the First Edition Survey Project (FESP) carried out some years ago by RCAHMS (The Royal Commission on the Ancient and Historical Monuments of Scotland) to cull evidence of habitations from early OS maps. The FESP data is a small subset of the RCAHMS collection, but particularly apt for a Communities and Culture Network project as it concerns local rural heritage, and is of particular interest to local groups such as those fostered by the Scotland's Rural Past project.

The project successfully upgraded around 18,500 site records and returned the data to RCAHMS in a form suitable for loading into their online access system, Canmore. A further 4,500 site records were partially processed by a different method, to allow comparison of techniques and suggest avenues for future data projects.

=°.°=

Contents

2. Aims and Objectives

I - Background

In 2001, RCAHMS (The Royal Commission on the Ancient and Historical Monuments of Scotland) completed an ambitious and long-running project to map unroofed settlements depicted on the Ordnance Survey (OS) First Edition maps and add them to the Canmore[1] online database. The project (named *FESP*, "First Edition Survey Project") made over 25,000 new sites available for "medieval or later rural settlement" research [1]. These sites are not high profile visitor attractions – they are mostly ordinary cottages and farm buildings, now ruined – but they paint a picture of rural life over the past few centuries.

More recently, the audience for this data has become much broader, through the highly successful *Scotland's Rural Past*[2] (SRP) project, which encouraged communities around Scotland to find out about their local heritage and protect it for the future, through research and hands-on field investigation and survey. The SRP project has now finished but the groups it fostered are continuing. Its popularity with community groups encouraged the creation of a complementary 5-year project, *Scotland's Urban Past*, which is currently in the final stages of consideration for award of a Heritage Lottery Fund grant and if successful will start in the autumn of 2014.

This project, *Event Mining in our Rural Past* aimed to make the FESP data more accessible and useful to users such as the local history enthusiasts of SRP, by using text mining to extract structured data from the free text documents associated with FESP. The project outputs were a collection of data files ready for loading into Canmore, which is the online window to the data. The translation of free text into structured fields makes it much easier to run sophisticated queries against the data collected by SRP contributors. This in turn will allow non-experts such as the SRP community teams to explore the similarities and differences between their local heritage landscape and others in different parts of Scotland.

1 http://www.rcahms.gov.uk/canmore.html
2 http://www.scotlandsruralpast.org.uk/

The second goal of the project was to explore how well automatic text mining works on the kind of data RCAHMS hold in great abundance. There are plenty of other candidates for language engineering work, if this approach can be shown to be successful and cost-effective.

II - Methods and Outputs

In recent years heritage environment recording has increasingly moved to an "event based" model. This means that a historic site is described through the events in its life-cycle, as far as these are known. For a recently found site the events may be "survey", "excavation" and so forth, with attributes such as the date and organisation or individuals doing the work. For a building whose history is better known, the events might include the architect's original design, the construction, and who used the building for what purposes.

The "events" model allows consistent handling of a wide range of site types and makes searching for related sites easier for the user. However, converting existing data records and text documents into this form is a significant challenge. From the technical point of view the primary aim of this project was to explore the feasibility of text mining as an approach.

The FESP programme focussed on desk-based analysis of OS First Edition maps supported in some cases by field work. The first stage of the present project was a detailed analysis of the FESP data records to categorise them appropriately for different text mining techniques. The results of this analysis are presented in Section 3, *Key Findings*. The objective was to extract separate "events" in a standard format: the event type, a description of it, the agents involved in it, the "patient" that experienced the event (typically the site itself), and the date of the event. RCAHMS already have a data structure to support such records, and are gradually populating it by hand – from the overall collection of around a quarter of a million sites – as time permits.

The preliminary stages of the project included working with the RCAHMS partners (Peter McKeague, the lead partner, and Graham Ritchie, IT specialist) to agree the scope and format of the FESP data to be examined. A local Oracle database was set up and the data loaded into it, along with template tables for the output data, in the event format RCAHMS have devised.

=°.°=

3. Key Findings

The time scheduled for this project amounted to 12 days FTE work. Slightly more than that budget was used but it was insufficient to achieve all the goals that could have been attempted with more time available. Around 80% of the data was processed fully and delivered in the required format, with the remaining – and more complex – 20% being partially processed as an indication of what the tools might achieve in a future project.

I - Data Analysis

Almost a quarter of the project time was spent in analysing the data, to understand its character and complexities, and in working out the relationship model for the event tables that would receive the output data. The model is attached as Appendix 1.

A total of 25,843 sites in Scotland were identified as being part of the FESP project[3] and of these 2,859 were found to have already been manually processed by RCAHMS staff and therefore out of scope for this project – leaving 22,974 site records as the focus to work on. Various minor coding errors were found as a by-product of the analysis work (incorrectly assigned event codes on a very small number of the manually created event records), and identifying them allowed RCAHMS staff to correct them.

The analysis was done partly by running SQL commands against the database tables copied from RCAHMS and partly by processing the text report data associated with each site record. For each record a text document was created by exporting the free text database notes, so that the text could be processed in large batches, outside the Oracle database, using tools not available within it. By a combination of these two approaches, the 22,974 documents were divided into "simple" or "complex", depending on whether they seemed likely to contain only a single event – typically a "desk based assessment" event for FESP records – or multiple events, perhaps including field visits, antiquarian observations or suchlike. The technical report at Appendix 2 contains more detail on the analysis results. Around 80% of the texts, 18,552 documents, were classified as "simple", leaving 4,422 "complex" ones. Clearly the sensible strategy was to concentrate limited resources on the 18,552 as these were easier to process as well as being by far the larger category.

3 The original number was 25,845 but the analysis work found two that had been incorrectly entered.

II - Processing simple events

The majority of the simple FESP events follow a regular pattern: the first line of the document contains an OS grid reference for the site, then there is a paragraph of descriptive text and finally a line of formulaic text which should conform to a pattern such as:

```
"Information from RCAHMS, (<initials>), <date>."
```

This signifies a desk based assessment event. The corresponding pattern for a field visit by RCAHMS investigators would be:

```
"Visited by RCAHMS, (<initials>), <date>."
```

The natural way to deal with patterns such as these is using regular expressions, and this was the approach taken. The initials are those of RCAHMS investigators, past or present; these were matched against a list supplied by RCAHMS, with ambiguous entries being resolved manually. The opening words of the pattern indicate the classification of the event, such as "desk based assessment" or "field visit".

A simple event matching the layout described can be processed to populate database fields for the event description (the text paragraph), the event location (the OS grid reference line, parsed to a valid reference), the event agent (found by parsing and resolving the initials) and the event date (found by parsing the date).

As is to be expected with data that has been entered manually, over many years and by many different people, there were wide variations in the actual patterns found in the "simple" documents. The punctuation, the format of the date, the layout of the single or sometimes multiple grid references all varied, plus there were spelling mistakes and typographical errors to allow for. For example, there were nine different variants on the word "Information", at the beginning of the desk based assessment pattern. This is a standard problem in data cleaning and is time-consuming but not difficult to deal with, requiring patient analysis and checking of results to ensure that every variant has been captured. (In the example mentioned, the regular expression snippet `"/[Ii]nf(?:or|pr|ro)mati?o[nm]/"` captures the single word "Information" in all variants present in the data.)

An advantage of using regular expressions in this way is that the process is deterministic and, with sufficient error trapping, one can be sure that every single record has either passed or failed the parsing step. The law of diminishing returns means that is simpler to process wild variants by hand than to write software to handle very rare exceptions, so beyond a certain threshold, records that could not be processed were listed as errors to be corrected manually. Out of the 18,552 text documents, only 193 "NGR error" reports were listed and 31 errors in parsing the organisation, initials and date information. The NGR errors were where the grid reference line could not be parsed with confidence, for reasons such as invalid grid square letters for the UK or unmatched easting and northing strings.

This part of the project, to transform 80% of the project documents into structured event data, was successful. The software engineering was quite time-consuming, but in no way comparable to the man years of effort that would be required to do this job manually. Furthermore, manual editing of this kind is repetitive and requires enormous attention to detail, and it is therefore likely that new errors would be introduced by the cleaning process itself.

III - Experiments with complex events

By their nature, the complex events are much harder to deal with systematically. In this case the text is to be parsed into multiple events, each with different categorisation, date, agents and so forth. RCAHMS recording practice has been to follow the basic pattern explained above, with paragraphs of descriptive text followed by single lines that give clues to the category of event; but there is room for wide variation. Figure 1 contains a sample text document marked with the events it should be split into. It might be possible to parse text of this kind using the regular expression approach, but a quite different technique is suggested here.

Named Entity Recognition (NER) and Relation Extraction (RE) are well-established information extraction techniques within the broad umbrella of natural language processing. For RCAHMS data the entity categories include the standard ones such as personal names, organisations and dates plus categories specific to the domain such as site type and event category.

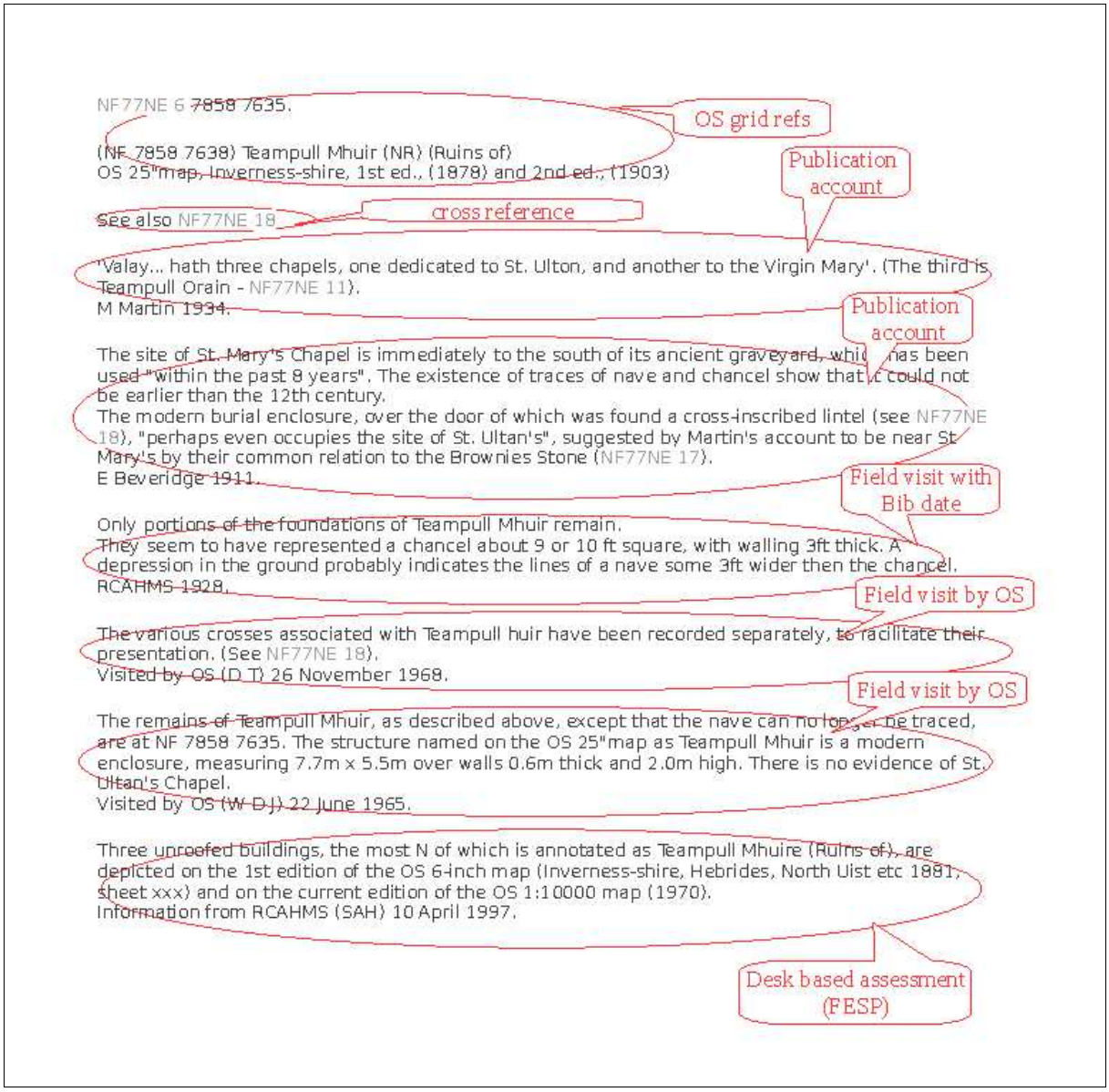


Figure 1: A "complex" event text. Taken from http://canmore.rcahms.gov.uk/en/site/10055/ (© copyright RCAHMS).

Once entities have been found and categorised in the NER step, one can look for binary relationships between pairs of entities in the RE step. By grouping such binary entities around an "event" it is possible to characterise the date, agent, location and so forth for the event in the way desired for this project. See [2] for a full treatment of the procedure, using data from the RCAHMS collection that is similar to the FESP data. The software developed in this research was re-used for the present project.

The first step in this approach is the NER process over the document collection, to identify significant strings such as dates, initials, grid references and so forth and allocate them to the

correct categories. There wasn't time to tailor the pre-existing software tools properly for the present task but some simple experiments were done over the FESP text, to give an idea of the kind of extraction that might be possible. Taking the text from Figure 1 as an example, the NER experiment produced a list of 71 entity strings, classified as place, date, period, personal name, organisation, sitetype, grid reference and so on. The process also identifies events, where it postulates a relationship between entities. For example “M Martin 1934” is correctly identified as a bibliographic reference, linking a person to a publication date, and this string is classified as an event, specifically a bibliographic event. The table below shows a sample of the first few of the 71 entities detected in the text from Figure 1, and how they were categorised by the classifier software.

Entity string	Category
NF77NE_6_7858_7635	Grid reference
NF_7858_7638	Grid reference
Teampull_Mhuir	Site name
OS	Organisation
Inverness-shire	Place
1878	Date
1903	Date
Valay	Site name
Chapels	Site type
M_Martin	Personal name
M_Martin_1934	Event
1934	Date
St_Mary's_Chapel	Site name

A full explanation of the process is beyond the scope of this report but it should be clear that if the significant strings can be found automatically in this way, there is a good chance of being able to construct events of the kind produced for the “simple” FESP documents.

It is important to recognise that this procedure, of NER and RE steps, is not deterministic in the way that the regular expression technique is. The software uses machine learning techniques to build a model of what the various NE types are like and uses this model to recognise candidates in new text, with a certain probability. On average the process is around

80% successful in terms of precision (whether the NEs it detects are correct) and recall (how many of the NEs in the text it finds). This is of course an error rate of 1 in 5 – and to some data owners such inaccuracy is unacceptable. The counter-argument is that getting 80% of the data processed automatically is worthwhile and one should work out a strategy to deal with the errors an automatic process is bound to make.

=°.°=

4. Key Issues

The project showed that simple data cleaning work can be undertaken very successfully and with guaranteed accuracy by straightforward software processes. It seems probable that significant quantities of records, not just from RCAHMS but from similar archive curating bodies, could be identified and worked on with these methods, saving enormous amounts of manual correction work. It would be important to find coherent subsets of the data collection in order to be reasonably sure of success; the FESP set – that is, the "simple" 80% of it – was ideal in this respect.

Sadly there was insufficient time for a thorough exploration of how best to deal with the messier, complex data. The experiments that were done showed promise. It would be good to take them further.

=°.°=

5. Next Steps

This was a seed project and as such it fulfilled its brief entirely. The next steps should be to undertake further experiments on the complex sites, now that a good test bundle has been assembled for the purpose. The foundations for finding the key data items – names, dates and so on – are in place, and a careful, step by step procedure seems requisite.

The FESP data used in this project is a very small subset of the entire RCAHMS collection. For experimental work a small coherent set of data is a boon, but clearly the goal should be to develop adaptable tools and techniques that can be applied to larger data sets.

Almost every archive body world-wide has data that needs cleaning, because it will have been assembled over decades or centuries, with evolving recording standards and priorities. The present drive to make such collections available to as wide as possible a cultural community – seeking understanding of shared heritage and sense of belonging in time and place – is very welcome but means that large scale reorganisation of data records is a pressing necessity.

A promising line for future exploration is the notion of "assisted curation", where human experts work in collaboration with automated tools. This is currently an active research field (see, for example [3] and [4]) though the take-up in the cultural heritage world has been slow so far. The goals include getting software to deal with the routine correction tasks, while flagging up the uncertain or low probability items for expert intervention. It will also be important to develop better interfaces to allow humans to examine and alter results easily.

=°.°=

6. Impact

The impact of this seed project will be largely felt by RCAHMS itself and by the users of its data – a community that includes professional archaeologists and architects, planners, and of course the ever growing band of those interested in their local culture and heritage. If the methods proposed are successful and adopted by the wider archive community, the ultimate impact of projects like this one will be a complete change in the way we view the archive curation task. There is good evidence from scientific and medical fields that assisted curation is a promising way forward.

=°.°=

7. Dissemination

We hope to find support for publishing the results through the Computer Applications in Archaeology Conference, which is one of the main routes to dissemination in the heritage world. The results will be shared with RCAHMS' sister bodies in Wales and England, through joint meetings and committees. We will also seek opportunities to present findings at the various workshops regularly occurring in the historic environment community.

=°.°=

8. Funding

The project partners will seek further funding both for dissemination of results and also to carry out the Next Steps identified above. So far no specific funds have been targeted but the completion of a seed project like this one should make it easier to justify investment. The kind of work proposed will ultimately pay for itself because it replaces extensive manual effort.

=°.°=

References

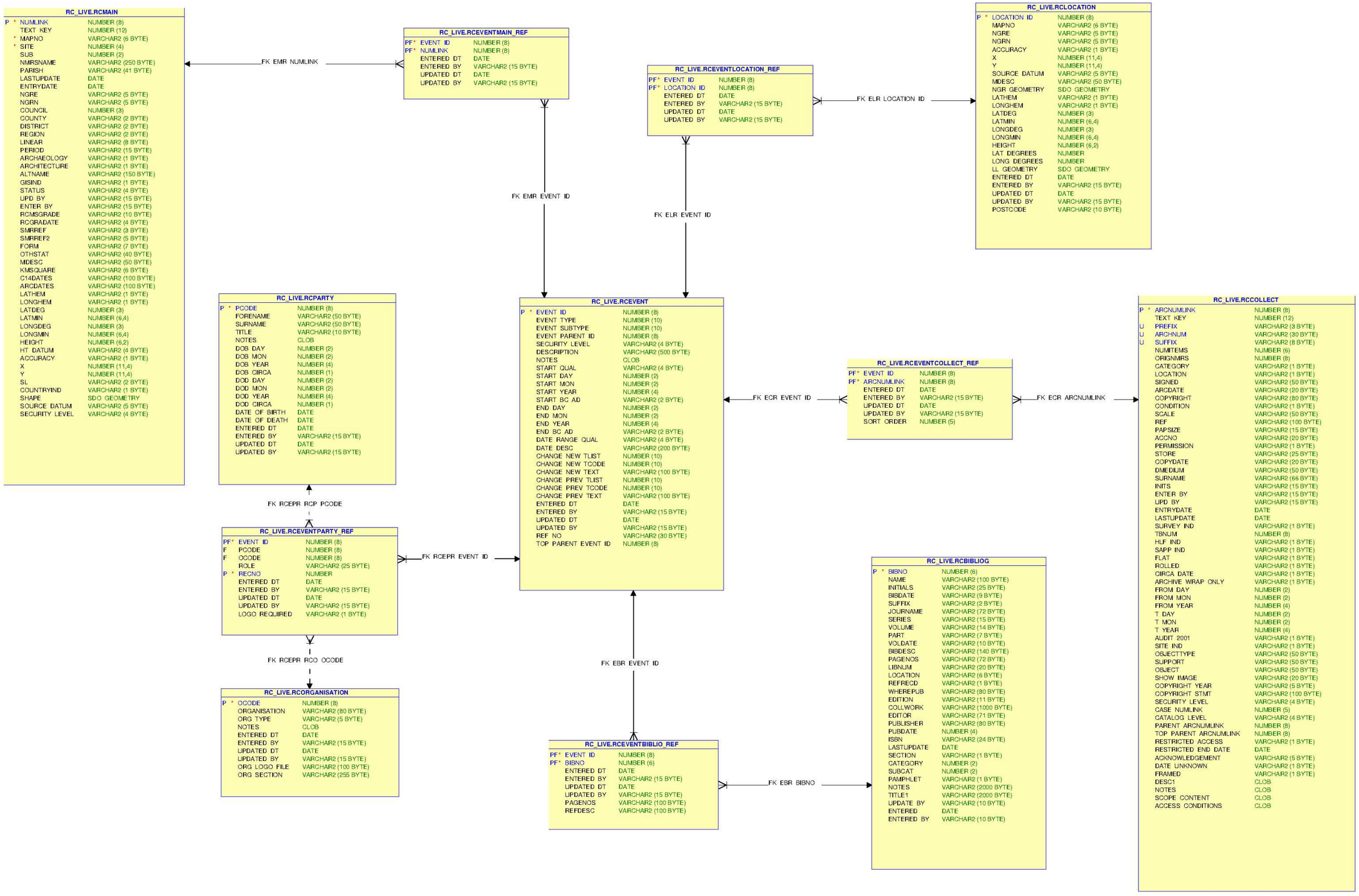
[1] **L McInnes**, Medieval or Later Rural Settlement: 10 Years On, : (2003)

[2] **Kate Byrne and Ewan Klein**, Automatic Extraction of Archaeological Events from Text, : (2009)

[3] **Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjor, Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin and Xinglong Wang**, Assisted Curation : Does Text Mining Really Help?, :556-567 (2008)

[4] **Chih-Hsuan Wei, Bethany R. Harris, Donghui Li, Tanya Z. Berardini, Eva Huala, Hung-Yu Kao and Zhiyong Lu**, Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts, **Volume 2012**: (2012)

Appendix 1 : Event Data Model



Appendix 2 : Technical Report

This report describes the data supplied to RCAHMS as output from the project. The data was zipped in a bundle containing the files listed in the table below. The six “kb_*.sql” files are SQL import files for the correspondingly named tables.

README.pdf	This file.
dbaUpdates.err	Lists the errors or anomalies detected in the input data. These are either “NGR error” records (193 errors) or “PARTYREF error” records (31), making a total of 228 site records requiring manual intervention. The NGR errors are where it was impossible to parse the grid reference line, for various reasons. The PARTYREF errors relate to problems parsing the “Information from...” line, for organisation, person and date.
dbaUpdates.out	The output corresponding to the .err file above. This data was loaded into the KB_FESPDATA table and exported to kb_fespdata.sql but this file is included as it may be more convenient to deal with. It enables visual inspection of the error records to see what was parsed out into fields.
kb_eventmulti.sql	Export of table KB_EVENTMULTI, which is a list of (2859) numlinks of FESP sites that are outwith the scope of this project as they have multiple event records and are assumed to have been processed already.
kb_eventsingle.sql	The 22984 records that are in the scope of this project are listed in table KB_EVENTSINGLE. It contains columns *numlink, event_id, processed, eventtype, typecode*, which are as follows*:* *numlink:* for the FESP site *event_id*: corresponding event *processed*: Y/N depending on whether this site's event has been processed *eventtype*: “desk based assessment”, “field visit”, “architectural notes”, “unclassified”, or “complex text notes” – see below for further details *typecode*: the event_subtype where known.
kb_fespdata.sql	Table KB_FESPDATA contains 18370 rows and is a straight load of the dbaUpdates.out file; one row for each FESP site processed by the CCN project.
kb_fespevent.sql	Table KB_FESPEVENT is a copy of the 22984 FESP event records from table RCEVENT. Of these, 18370 have been updated (*updated_by*='CCNproj') and are destined to **replace** the matching rows in RCEVENT.
kb_fespngr.sql	Table KB_FESPNGR contains 22984 rows of which 18370 have been updated (*updated_by*='CCNproj') to contain NGR fields parsed out of the first line of the text notes (the line is removed from the event *notes* field). Where parsing failed the string is held in the *errors* column. The *mdesc* field contains “centred” where the NGR is so described in the text.
kb_fesppartyref.sql	Table KB_FESPPARTYREF corresponds to RCEVENTPARTY_REF and contains 18370 rows to be **added** to that table; one for each of the 18370 FESP events processed and given a single person/org/role link record.

NoEdn.txt	A list of the 50 event texts that were queried by me as not apparently containing a reference to the first edition of the OS map (to be expected in FESP notes). This file reports the reasons – most are typos (eg “editon” instead of “edition” and several are cancelled sites.

Summary of the Data Analysis

1. FESP records and associated events

There are 25,843 FESP sites, each with at least one event. The original number was 25,845, with two having no events; but these turned out to be data entry errors, corrected by Peter in the live database. Of these, 2,859 sites are associated with multiple events and are therefore assumed to be already processed and outwith the scope of this project. That leaves 22,984 FESP sites with single events, that are the target of this project. (22984+2859=25843)

2. *singleEvent502502* and *singleEvent502881* sets

The singleEvent502502 set contains event notes (one file per event record, named with the numlink) for FESP records having only a single event with code 502502, ie "archaeology notes". These are unprocessed.

None of these event records has a link to other tables, ie no extraction of people/orgs has been done.

There are 22974 events in the 502502 set, out of a total of 22984 single event FESP sites. (We assume that if there are multiple events for a site, it's been processed already, by hand.)

The remaining 10 sites (6 field visits, 2 architecture notes and 2 dbas) are these:

numlink	event_id	eventtype	eventsubtype	
128832	650740	502348	505881	architecture notes
132272	781095	502346	502551	desk based assessment
149735	801223	502348	505881	architecture notes
180137	642079	502346	502551	desk based assessment
315898	886076	502346	502399	field visit
315901	886078	502346	502399	field visit
315978	886218	502346	502399	field visit
315979	886219	502346	502399	field visit
315980	886220	502346	502399	field visit
316020	886360	502346	502399	field visit

Only the two architecture notes (numlinks 128832 and 149735) have not already been processed. These two are in the singleEvent502881 set.

3. *simpleEvents* set and subsets

The 22,974 singleEvent502502 files were subcategorised into “simple” and “complex”, where “simple” means that the notes appear to correspond to only one event (indicated by there being only one paragraph of text) and “complex means that there are multiple paragraphs and therefore we expect multiple events. In fact one of the simple files (numlink 136420) contains two events within a single paragraph so it transfers to the complex category.

As expected, about 80% of the FESP events (ie singleEvent502502 files) are simple: 18,552 files. These are held in the simpleEvents set, broken down as follows:

	18,370	desk based assessment – the standard FESP event (*dba* subset)
	156	field visits (*fieldVisits* subset)
	26	miscellaneous (*unclassified* subset)
	18,552	
Add	4,422	*complexEvents*
to get	22,974	total *singleEvents502502*
Add	10	noted above (6 field visits and 2 dbas already processed; 2 architecture)
to get	22,984	total FESP events.

Of the 156 field visits, 4 do not mention “first edition” (in any valid variant including typos). Many of the unclassifieds don't either (mostly cancelled sites). The file noEdn.txt lists all 50 that were picked up.

KB_EVENTSINGLE table

This table contains a summary of the processing done and left undone, on the 22,984 FESP sites. The columns are as described earlier and the first few rows of the table are as follows:

NUMLINK	EVENT_ID	PROCESSED	EVENTTYPE	TYPECODE
51	640968	N	complex text notes	
58	641057	N	complex text notes	
75	641618	Y	desk based assessment	502551
173	641518	N	complex text notes	
176	641521	N	complex text notes	
378	641125	Y	desk based assessment	502551
396	641750	N	complex text notes	

Simple SQL reports on this table will show how many FESP events have been processed (processed='Y') and how the events were categorised in the analysis. For example, the following SQL query returns a summary corresponding to that given in the analysis section above:

```
select distinct eventtype, count(eventtype), processed
from kb_eventsingle
group by eventtype, processed
order by count(eventtype) desc;
```

This query returns (with the sum added, for readability):

EVENTTYPE	COUNT(EVENTTYPE)	PROCESSED
desk based assessment	18372	Y
complex text notes	4422	N
field visit	156	N
unclassified	26	N
field visit	6	Y
architecture notes	2	N
	22984	

Software used

The text notes fields were exported to individual files for convenient processing, and divided into sets as noted above. Odd bits of pre-processing, checking and analysis were done with standard tools: awk, grep etc.

The programs *prepareDBAupdates.py* and *doDBAupdates.py* process the main set, of 18,370 desk based assessment events.

The output of *prepareDBAupdates.py* is the dbaUpdates.out file (which is loaded as table KB_FESPDATA, in case required). This is produced by parsing the text files in a loop, using simple regular expression patterns.

The second python program reads this file in and runs updates and inserts against the relevant database tables. The main event record is updated (through its surrogate copy, KB_FESPEVENT) and new rows are inserted into the eventparty_ref table (through its surrogate, KB_PARTYREF). It was decided that attempting to transfer the grid ref data into RCEVENTLOCATION_REF and RCLOCATION was not worth the effort, but since most of the grid ref parsing had already been done, the data is included in KB_FESPNGR in case useful in the future.

Manual processing

A total of 18,370 sites were processed – the simple, single event, “desk based assessment” ones. The small sets of unprocessed data (*viz* the 156 field visits and 26 unclassified events) are left for manual processing, as this would probably be as fast as adapting the software to handle them. The relevant site numbers and event_ids can be easily found from the KB_EVENTSINGLE table.

Event Mining in Our Rural Past

Keywords
(data cleansing, community access, language processing)