Manuscripts Online Advisory Group

9 March 2012
Humanities Research Institute

Present
Project Team: Michael Pidd, Kathy Rogers, Sharon Howard, Orietta Da Rold, Wendy Scase, Estelle Stubbs, Linne Mooney, Jeremy Smith, Darren Bailey (Mickey & Mallory)
Advisory Group: Bob Shoemaker, Ian Johnson, Aleks Drozdov, Stephen Brooks, Adam Farquhar, Peter Ainsworth
Apologies: Andrew Prescott, Tim Hitchcock, Jane Winters, Paola Marchionni, Keven Schurer, John Thompson

Overview

MP outlined the project to the Advisory Group (summary document).

Funded by JISC, a sister project to Connected Histories (www.connectedhistories.org) as a federated search service to provide a single point of access to search online resources. The resources included are written and early print sources (1000 to 1500), located in institutions in the British Isles or with a British provenance.

The departures from Connected Histories lie in the challenges of the materials to be included, such as variant spellings and languages; improving Web 2.0 features; and mapping dialectical information.

Data is processed in two stages: (if needed) natural language processing for selected information (languages, people, places and dates) and indexing.

RS: query about the greater variety of material compared to CH - genres, mix of primary and secondary sources - and implications for presentation
-See harmonization as a significant challenge, eg: diversity of datasets, languages, chronological range, variation in handling of special characters.

Indexing process, search specification, non-Latin characters

Search flowchart illustrates process. Particular issues identified so far:

Three types of date:
- cited date
- document (production) date
- content coverage

Non-Latin characters
Two types:
1. a core set that will be fully searchable using special popup keyboard - thorn, yogh, ash, eth, wynn(?);
2. a larger set that will be transliterated to Latin characters in the indexing process

AF: noted very English focus; will the project be able to deal with wider range of scripts - how easily extensible to eg Cyrillic?
- need to limit the project at this stage; resources will be individually audited for any additions to the existing list, but as we're dealing with transcriptions it's not so much about the possible range in the original MSS as the decisions made by editors in transliteration.

Current entity lists/gazetteers for NLP are primarily English or anglicised.

Natural Language Processing

KR overview of process:

The NLP is a two-stage approach: 1. languages; 2. people/places/dates. Stressed that MSSO doesn't do record linkage: not trying to identify (eg) "London" as a specific georeferenced place, but rather simply that it *is* a place (it's for users to work out where).

Different approaches to different types of dataset:
- structured datasets: most entities are already identified and we only have to do indexing.
- partially structured datasets: may need additional NLP to identify more entities or certain types of entity
- unstructured datasets: none of the information we're interested in has been identified - need complete NLP

The NLP process uses named entity recognition (using gazetteers) and additional grammar-based rules.

KR has started work on building language gazetteers
- Perseus for Latin
- identified some possible resources for French.

Place names gazetteers
- using gazetteers from Connected Histories, and a range of our own existing resources (including Taxatio).

PA - suggested Cassini Map for placenames (c16/17 map of France - not clear if there's a usable dataset...)
JS - possibly Early English Place Names Society, Gough map

Dates
- will need to write some new rules.

Action point: Editorial group: need small representative samples of data for their resources(including difficult material!) for testing, within the next fortnight.

RS: will the project attempt to distinguish between modern and medieval English?
KR: will distinguish between French and Latin, but not varieties of English - problem that there aren't sharp distinctions. With Latin, we find that the failures tend to be at boundaries - the NLP struggles to work out the start and finish of a language instance.

RS: need to alert people to success rates of NLP; give the caveats

KR: challenges of this material analogous to issues that Connected Histories experienced with OCR - small deviations from expected instances of gazetteers/rules easily throw off NLP

AF: question about what techniques/packages are being used for entity recognition
KR: GATE is used as the development environment (tokenizing and loading gazetteers), though may need to review that, but with entirely custom-built rules adapted from Connected Histories.

MP question for PA: is it possible to make distinctions between Anglo-Norman and French? PA suggested consonant groupings as the best way of distinguishing between the two.

Visual design and Web 2.0 features

Overview of first designs from Mickey & Mallory (image files)

Functionality largely based on CH

General look and feel
- mostly happy with colour scheme
SB: top logo seems a bit small and nondescript, could make it a bit larger and the decoration less prominent

Homepage
- quick search
- Discover resources
- news
- 'search paths' - sequence of searches in a session, anonymised and aggregated

Concerns about any recording of unique data - a full session would be a) unique and b) not very useful; segments of a session probably more useful
RS: you need enough traffic to make it useful!

Acceptable approach to protect privacy agreed:
- available to registered users within workspace
- default private, with option to make your own search paths public
- resource for recording search/methodology on publication

RS: why is that information on the homepage anyway? need something attractive to new visitors, but not sure this is best option
AF: suggestion of something like a Wordle of popular recent search terms, give idea of breadth/range of resources without privacy issues.

Quick search
- drop ‘quick’ (or drop label altogether?)

About project links
- suggested that it should be top level navigation
Include link in introductory text
‘Participate’ rather than ‘Contribute’ (suggests ‘donate’!)

Main Search form

PA: useful to have an ‘institutions’ category for searching (eg ‘the Curia’) - but this would require extra NLP

WS unsure about term ‘keyword’

AF: person and place fields - how will users build multiple (Boolean) searches?

AD: users have difficulty with understanding Boolean concepts and it is better to convey this through intuitive interface features rather than requiring users to learn Boolean operators.

AF: make the default Boolean relationship clear

AD: TNA search automatically recognises Boolean if in caps (indicating user knows what they’re doing), otherwise it uses system defaults

Cited date represents specific dates
KR: Search needs to be date range and dropdown, not free text
slider .

Source types and Resources
faceted search

RS: need to scroll down for this part of search - try to get into single screen?

Use of screen estate was discussed
- NB search tips need to be prominent for novice users
- remove text above search form / move to sidebar

Search results page

Layout of results from resources is modelled on CH in order to deal with varying size of resources

'View dialects'
- option to view variant spellings of a search, re-run if required

Faceted options in sidebar
doc types, dates - needs work

Map these results
- crowdsourcing dialect variants
(JISC view was that this should be open, not restricted access)
PA: scribal dialect is not always the same as location of provenance (scribes can move!), so would need caution. But crowdsourcing can be very productive and useful to researchers.
AF: is the search results page the best place to do this?

AF: concern that presentation format reinforces archival preferences over content - possibility of presenting alternatives during user testing?

SB: only showing 3 results on screen - possible to make better use of space?

Next Advisory Group meeting

Need to book well ahead and avoid clash with Quadrivium VIII (booked for 1-2 November).
Action point: SH to set up a Doodle Poll.

A.O.B.

Project collaboration agreement has been completed and signed by all partner institutions.

Associated documents

Most documents/files mentioned above are available for reference at
http://manuscriptsonline.wordpress.com/documentation/