Manuscripts Online

Friday 20 January 2012
Humanities Research Institute

Present: Michael Pidd, Kathy Rogers, Orietta Da Rold, Wendy Scase, Linne Mooney, Sharon Howard
Apologies: Jeremy Smith, John Thompson

1. Collaboration Agreements (partner agreements) and funds

We're waiting on feedback from York and Glasgow. LM to chase with Sue Final.

2. Project Plan, blog and key JISC documentation

The project plan has been completed and posted on the project blog.

The blog is up and running.

Future content:
- list of the resources in MSSO
- profiles of the partners' datasets (and possibly others, later)
- news about Quadrivium
- technical posts (NLP)
- discussion and questions

Suggestions for dissemination of news
- user testing mailing list
- societies

The blog is already integrated into Twitter so that posts are automatically shared with followers.

3. Material Transfer Agreements (content licences)

Draft MTAs have been sent out to most content providers, including the project partners, for review, and awaiting responses. SH will monitor and chase these as necessary.

Exceptions:
METSO - MP has not yet heard from Patricia Hollahan at Western Michigan re: METSO texts. To chase..

BL Illuminated MSS - MP to have meeting with Kathleen Doyle before sending MTA

JISC Historic Books -

MP has met the JHB people and the situation is complex at present. Needs to discuss with both ProQuest and JHB and confirm situation for our MTA.

However, there is no problem with the EEBO-TCP texts (which will be accessed via ProQuest/JHB).

4. Design Agencies - invitations to tender - focus groups

Invitation has been drawn up and sent to 5 agencies, deadline 26 Jan. Brief is for visual look and feel of the site. The agencies are:

1. Mickey and Mallory
2. Headscape
3. Kambara
4. Chameleon Studios
5. Can Studios

Will circulate bids to EB for discussion.

Focus groups need change of date to allow the successful agency time to produce material. WS to try for end of Feb.

5. Stakeholder panel

MP to send out invitations next week.

6. NLP processing of Latin and other languages

KR report on Latin samples

Process:
1. make clean text
2. process, using rules to weight probability that text is in Latin, eg
- words in dictionary
- word endings
3. mark up

Two samples processed (from Vernon Manuscript and CUL Ii, 1.33), both good results, but a number of particular issues

boundaries: the NLP processor needs to work on ‘chunks’ of words that aren’t too long or too short - best way to break up texts?
punctuation - problematic in medieval texts; different uses from modern, and not consistent
lines - more promising, especially for verse

Very short fragments of Latin (1-3 words) are likely to be missed unless made up of words that are heavily weighted as Latin. Single words would have to be unambiguously Latin.

false positives: dictionary method is most successful where English/French vocabulary used is distinct from Latin

refinements
compile list of common English words to 'de-prioritise' or exclude
exclusions:
-tion
-ing
w, y, possibly k
thorn, yogh

the process also generates words that 'might be' Latin; this can be reviewed to be added to the Latin list

Once general principles are in place, individual datasets can be tweaked as necessary

Next task - French
word lists needed
We already have some resources (Froissart)
Anglo-Norman dictionary - MP to contact the people on the project to invite them to collaborate
http://www.anglo-norman.net/

NB for English: won't attempt to NLP for Old/Middle/Modern English variations, but will include variants in search engine.

7. NLP processing of other features

Places
place name resources
Taxatio, Cause Papers - tagged

should be able to find resources for English/ British place names; others might be more challenging - use blog to ask for help
+ spelling variants

People
need to think about rules that are likely to be useful - our existing name resources are for modern names; what's likely to be different in medieval sources

do we want to mark up names like 'wife of Bath'? 'name/title of place'

8. Special characters

For English words, the MED includes main variants.

Yogh = g, gh, y
Thorn = th
eth = th
ash = ae, e

Latin abbreviations
Crossed P = par/per
Looped P = pro
crossed b = ber
Macron (over u or o) = m or n

Superscript
Superscript hook = er/re
Superscript T = t
Superscript R = r, ur
Superscript M = m
Superscript N = n
Superscript E = e
Superscript O = o
Superscript U = u
Superscript I = i, ri
Superscript A = a, ra
Superscript 9 = is, es, us

MP to compile listing.

9. Quadrivium Workshops

<u>February, Glasgow</u>
- ask JS to flag up project in introduction
- OR and WS will also be present
- put together flyers
- publicise the November workshop

<u>November, Sheffield</u>
- full presentation and demo

10. Dissemination Plan

Flyers
- design using the agency's templates
- get a quote from a printer
- small print runs so we can update later as needed

Flag this up at March meeting

11. International data linking opportunities

MP to contact Europeana

The European Library - Alastair Dunning

12. AOB

Need to raise issue of access to Geographies of Orthodoxy resources with JT; also ask about what's happening to his site after 2015.