COST STSM Final Report

Miranda Lewis

Design Sprint: Keywords and Text, 11–15 April 2016

Last month I had the good fortune to attend a week-long Design Sprint which was organized as part of the COST Action Reassembling the Republic of Letters Working Group 3 programme and held at the HuygensING in The Hague. A Design Sprint brings together a pre-selected group of specialists — in this case IT developers, project managers, scholars, and editors — and enables them to work on a single test-case in a series of intensive back-on-back sessions. The objective of a Sprint is to compress into a limited number of days what might otherwise take months of interrupted discussion, experimentation, and development, and at the end of the set period a prototype is unveiled to serve either as a starting point or as a proof of concept.

This Design Sprint focussed on one of the core considerations of Working Group 3, namely the subjects discussed in early modern correspondence and their standardization, capture, and representation in a digital catalogue. One working week was set aside for nine individuals to grapple with the following tasks: to understand and articulate the problems and pitfalls encountered by scholars as they work to define the subjects and keywords under discussion in early modern correspondence; to establish what, in an ideal scenario, might meet the demands of this research as scholars work in a digital environment; to impose a structured format on hitherto unstructured data; to discuss and assess potential solutions; and to identify and work on a prototype that could be developed to demonstrate 'proof of concept'. All this in five days — no mean task.

I was particularly interested to participate in my capacity as Digital Editor of Early Modern Letters Online [EMLO]. In EMLO's current data model it is possible to capture both keywords and subject headings for individual letters, but it is not possible for scholars to reference and draw automatically on an established taxonomy, nor at present is it a simple matter to ensure standardization of keywords across multiple correspondence catalogues. When EMLO was in development, few scholarly contributors were volunteering to record in the union catalogue the topics discussed in the correspondences under their investigation. Such work can take a disproportionate amount of time, something contributors to EMLO have in fearfully short supply. Many contributors, who were offering for publication calendars collated in the course of research carried out at earlier stages of their careers, were not in a position to return to supplement their metadata with this information. Thus, in EMLO's early years, basic fields for subject words and keywords were included in the data model, but it was not thought necessary at that stage to develop a structured and referenced system. More recently, however, a number of contributing scholars have expressed interest in assigning both broader subject and more specific keywords to their corpora of letters. Although at EMLO we have investigated the potential use of a number of different ontologies, we have found neither one that fulfils the often quite different needs of these scholars, nor a means of allowing the flexibility that each individual scholar requires. EMLO receives contributions from large numbers of scholars and projects working on a range of correspondences from different centuries in a variety of languages, and the topics found within these correspondences are diverse.

The five days of the Sprint at HuygensING were divided clearly into separate sessions by the indefatigable and clear-headed Sprint coordinator, IT project manager Astrid Kulsom. With admirable skill and aptitude, she guided the assembled group through the initial stage of identifying and 'unpacking' the problem, which came in the form of the subject of doctoral student Karen Hollewand's subject of research —Dutch scholar Hadriaan Beverland (1650–1716).

Beverland proved an inspired — and entertaining — case study. Banished from the United Provinces in 1679 for reasons you may guess when you read the second half of this sentence, he was obsessed with exploring in his written work and correspondence (in alphabetical order, rather than priority of subject) scholarship, scripture, sex, and sin. Best known from his somewhat louche depiction in a portrait attributed to Ary de Vois and now in the Rijksmuseum (I can't resist mentioning it is worth seeking out Beverland as he leans back in his chair, pipe in hand, with a scantily clad woman, usually described as a prostitute, at his side), Beverland gave us all the *double entrendres* we could have asked for, as well as proof that there are certain topics only a human eye is able to tag: Venus; or sin; or *voluptas*. We found in Beveland problems that arise from the use of different languages, the 'you say *prodraga*, and I say gout; he says *kramp* and she says *gicht*' issues. Using a small selection of his letters, we were able to discuss how ideally scholars need access to multiple ontologies (for example, the Library of Congress classifications, or Iconclass, or InPhO, etc.) as well as to early modern classification lists; how they need access to a range of dictionaries (Oxford English Dictionary, Lewis and Short, Duden, Larousse, etc.). We determined quickly that established ontologies were not detailed enough for our needs, that scholars would have to be provided with a structured method of adding their own terms, sometimes terms very specific to the correspondence upon which they worked and sometimes terms shared across other early modern correspondences. We recognised that scholars would need the ability to add annotations to certain keywords to explain their usage, their meaning(s), and the editorial decisions behind the word's selection so that other scholars could see at a glance whether a particular 'user-added keyword' was the one they needed to select (rather than creating a duplicate or variant). And we agreed that to make these 'user-added' keywords meaningful and to connect them usefully to larger schemas, they should be linked to terms (and thus, by extension, to branches) that exist in established ontologies. Following demonstrations from Walter Ravenek (IT Developer of the Circulation of Knowledge project's *ePistolarium* database) and Charles van den Heuvel (Professor of Digital Method in Historical Disciplines) and with significant input from Ludovica Marinucci (Università di Cagliari, currently on an internship at the HuygensING during which she is working on the use of keywords for mapping between the letters and texts of Christiaan Huygens), we considered the potential for extracting automatically generated keywords, subsequently gathered under topic headings, from digitized text and then discussed how this would have to be combined with oversight from scholars who would need to be able to link back to the original digital transcription.

We spent a day pondering these problems, brainstorming, and compiling a wish-list. Day Two was a devoted to sketching ideas and coming to terms (no pun intended!) with what might be possible. 'Decision day' was midweek: what would be possible within the constraints of the Sprint. And it was at this point that IT Developer René van der Ark withdrew to his workstation to focus on the prototype. His work moved centre stage on Day Four as the prototype was run back and forth through various iterations as suggestions were sought and updates delivered, authority files downloaded and incorporated, and what was not deemed possible or practical to include was channelled into a 'roadmap' for future consideration. And then came the 'Day of the Proof of

Concept' on which the prototype was unveiled to members of HugyensING staff in a team presentation.

What I took away with me from this Sprint was the conviction that structuring of topic metadata is something that would be extremely useful in EMLO and, if possible, should be implemented in the next round of development. Whenever subjects and keywords are collated for an early modern correspondence, these should be standardised and linked across the union catalogue to those recorded in other correspondences. 'User-added' keywords have to be linked to established ontologies and attached to the relevant branches from where, at a higher level, subject terms may be extracted and stored. Keywords, and the schemas into which they feed, could be visualised both over multiple correspondences and chronologically, but only if scholars use the agreed set of authority files. This could be extremely exciting and revealing work and I headed home to Oxford dreaming of visualizations of trees of knowledge with letters and keywords hanging off as virtual leaves. It was an exceptionally productive week and I am now an enthusiastic advocate of the Design Sprint format for concentrated collaborative work. One challenge remains, however: to ensure that all that was learned and demonstrated in The Hague can be picked up and carried forward to fruition.

Miranda Lewis
Digital Editor, EMLO

11 May 2016