Building a VOCabulary: the uses and challenges of thesauri for working with early modern recognized entities

Kay Pepping (ORCID), Huygens Institute for History and Culture of the Netherlands, Amsterdam, The Netherlands

Brecht Nijman (ORCID), Huygens Institute for History and Culture of the Netherlands, Amsterdam, The Netherlands

During its existence from 1602 to 1798, the Dutch East India Company (VOC) created a veritable mountain of documentation. Not only do these archives contain ledgers and details on trade, they also show us a detailed, though distorted version of the early modern Asian world. The GLOBALISE project[1] aims to unlock a key section of this archive called the *Overgekomen Brieven en Papieren* (OBP), which consists of documents that were sent from the company's Asian headquarters in Batavia to the Dutch Republic. Through a mixture of Handwritten Text Recognition (HTR) and Natural Language Processing (NLP) techniques like named entity recognition, the project aims to make these archives researchable rather than just searchable.

With an archive as expansive as that of the VOC, applying HTR and NLP is not enough to achieve this goal. The VOC activities spanned three continents and countless localities, each defined by its own traditions, culture and language. Its employees hailed from all over the world. While the company's lingua franca was Dutch, the company's records are peppered with influences from other tongues. When creating the archives the terms and phrases used to describe a single concept were not normalised. Within the documents, a traded commodity might therefore be referred to with local terms, terms used by traders of other Asian or European origins, and Dutch interpretations of these words. All three options feature all sorts of spelling variations. The same goes for references to weights, measures, professional titles and the other aspects of Asia described by the Company. Therefore, a method such as entity linking (linking an annotated entity to its representation in a knowledge graph)[2] is hampered by not knowing which annotations (might) refer to the same concept.

GLOBALISE hopes to solve this problem through the creation of a hierarchical reference thesaurus: the VOCabulary. By providing Unique Resource Identifiers (URIs) for terms, their variations, place in the taxonomy and location in the archives can be tethered to a single, digital reference point. A researcher would be able to then expand their query using our database, so that synonyms and semantically related concepts to a term would be

[1] https://globalise.huygens.knaw.nl/

[2] See for instance Delip Rao, Paul McNamee, and Mark Dredze, "Entity Linking: Finding Extracted Entities in a Knowledge Base," in *Multi-Source, Multilingual Information Extraction and Summarization*, ed. Thierry Poibeau et al., Theory and Applications of Natural Language Processing (Berlin, Heidelberg: Springer, 2013), 93–115, https://doi.org/10.1007/978-3-642-28569-1_5.

included in the results.[3] When looking for ‘Cloves’, the results would also include ‘Kruidnagel’, ‘Nagelen’ and the other terms used by the company. Moreover, due to the SKOS classification scheme used to structure the database, they could also choose to include subtypes of cloves such as ‘garioffelnagel’ in their results.[4] Besides commodities, the VOCabulary currently also includes measurement units and occupations with much more to come.

The benefit of creating our own structured vocabulary is that we can completely gear it towards our context: Asia as seen through early modern European eyes. Existing thesauri and datasets are valuable, but limited in its uses for our material. The structure of a source like the Getty's Art and Architecture Thesaurus (AAT)[5] is useful due to its linkability and expandability, but its contents are not geared towards early modern interactions with the Asian world. This is also the case for the Dutch Cultural Heritage Thesaurus (CHT).[6] On the other hand, a source such as the Huygens’ VOC-glossary[7] is geared specifically towards our context, but is static and lacks the hierarchy needed for optimal usability. A dedicated, expandable and linkable vocabulary geared towards our specific context is needed to provide additional, contextually appropriate information to the recognized named entities in the corpus. It is essential for the operability of the aforementioned query expansion. Furthermore, such a vocabulary could be of use to institutions which house VOC-related collections.

In this short paper, we aim to set out the aforementioned value of such thesauri and the challenges faced when building one. For the VOCabulary, there are currently three main challenges: time, reconciliation and curation. Where time is concerned: the task could take a lifetime, which means combining existing resources is key to getting it done within the allotted five year period of the project. While using existing reference material is essential for creating the VOCabulary within the allotted time, it creates the new time-intensive tasks of reconciliation and curation. There is a mountain of datasets and research to build on, but we need something geared towards our domain. We refer to existing thesauri (via SKOS mapping properties) where possible, but aim to provide an interpretation of the material that does justice to the structure of the source and the "creator" of the archive. Structuring the VOCabulary as Linked Open Data (LOD) will allow for future expansion from other perspectives. The large variety of occurring entities makes identifying, resolving and disambiguating terms encountered in the text an extremely labour intensive task.[8] Additional

[3] J. Bhogal, A. Macfarlane, and P. Smith, “A Review of Ontology Based Query Expansion,” *Information Processing & Management* 43, no. 4 (July 2007): 866–86, https://doi.org/10.1016/j.ipm.2006.09.003.

[4] Alistair Miles and Sean Bechhofer, “SKOS Simple Knowledge Organization System Reference,” W3C Recommendation, 2009, http://www.w3.org/TR/skos-reference/.

[5] https://vocab.getty.edu/aat

[6] https://thesaurus.cultureelerfgoed.nl/

[7] Marc Kooijmans and Judith Schooneveld-Oosterling, *VOC-Glossarium, Verklaring van Termen, Verzameld Uit de Rijks Geschiedkundige Publicatien Die Betrekking Hebben Op de Verenigde Oost Indische Compagnie* (The Hague: Instituut voor Nederlandse Geschiedenis, 2000).

[8] For other projects integrating multiple thesauri and datasets, see: A. Léon et al., “SILKNOW. Designing a Thesaurus about Historical Silk for Small and Medium-Sized Textile Museums,” in *Science and Digital Technology for Cultural Heritage*, ed. Pilar Ortiz Calderón et al., 1st ed. (CRC Press, 2019), 187–90, https://doi.org/10.1201/9780429345470-34; Helen Goulis, “The BBT Meta-Thesaurus Model: Building Interoperable Thesauri for Humanities Researchers,” in *Controlled Vocabularies and Knowledge Organisation*

difficulty is introduced by the needed curation of the existing definitions on top of reconciliation. Merely reconciling terms from the archive and existing reference works runs the risk of reproducing the biases of the VOC as well as those of the later colonial period. This ranges from the selection of which terms are included to the use of explicit derogatory language in the definitions of concepts. How do we deal with such insensitivities? It is key to present the information in such a way that the database does not contain offensive or outdated terms. At the same time, it should be kept transparent when a definition stems from, for example, a source from the 1920s. The line between collecting (merely connecting sources) and curating data (actively checking and editing their contents) is a fine one. As these issues show, thesaurus building, even with pre-existing material, remains a labour intensive task.

for the Digital Humanities, ed. Bruno Almeida, Rute Costa, and Filipa Medeiros (Lisbon: NOVA FCSH/CLUNL, 2021), 4–7, https://doi.org/10.34619/pgtp-upne.