Journal of Open
Humanities Data

Reflections on Encoding Languages in Historical Data: Working With the Multilingual Dimension of the Dutch East India Company Archives

RESEARCH PAPER

K. W. PEPPING

]u[ubiquity press

ABSTRACT

This article investigates the challenges of encoding languages in historical data through the example of a reference dataset: a thesaurus in SKOS format of commodities traded by the Dutch East India Company (VOC).

The VOC archives, from which this thesaurus draws a lot of its data, are far from purely Dutch. The company's multilingual workforce and interactions across Asia resulted in records influenced by a multitude of languages, full of loanwords and citations. This is further complicated by the VOC's role in colonising regions and suppressing local languages, resulting in some languages potentially only surviving in these 'Dutch' archives.

This means that when working with a large corpus like the VOC archives, various challenges arise regarding historical language evolution, vocabulary borrowing, extinct languages, technical standards that are not geared towards historical context, and political sensitivities around identity-bound language.

The article demonstrates how the GLOBALISE project navigates these issues by prioritising transparency, flexibility, and iterative refinement. It argues that as long as researchers are aware of the challenges, language complexities are not a roadblock but offer opportunities for further research and critical engagement with the past, encouraging broader discussions and creative solutions for encoding historical multilingualism and development of language.

CORRESPONDING AUTHOR:

K. W. Pepping

Huygens Institute for the History of the Netherlands, Amsterdam, The Netherlands

kay.pepping@hugyens.knaw.nl

KEYWORDS:

Dutch East India Company (VOC); skos; htr; reference data; language tagging; code-switching

TO CITE THIS ARTICLE:

Pepping, K. W. (2024). Reflections on Encoding Languages in Historical Data: Working With the Multilingual Dimension of the Dutch East India Company Archives. *Journal of Open Humanities Data*, 10: 29, pp. 1–10. DOI: https://doi.org/10.5334/johd.176

INTRODUCTION

While historical research often takes a qualitative approach by constructing a narrative that is exemplary for a larger trend, the quantitative approach to history has gotten a boost in recent years through the developments in digital humanities. Creating historical, structured data comes with various challenges due to, amongst other things, historians' desire to embed data in context and the reality of (non)-survival of historical source material (Finnane et al. 2018). This article will focus on one specific issue when representing historical data in a structured format: language encoding. Proper language encoding is important, as it benefits the reuse of historical data (Thieberger & Tuohy 2017).

Through an example of a reference dataset on commodities traded by the Dutch East India Company (VOC), this article will reflect on the challenges encountered when encoding the language of terms sourced from archival material. This example dataset is formatted in SKOS (simple knowledge organisation system), which defines providing every piece of textual information with a language tag as best practice (Mader et al. 2012).

First, an overview of the GLOBALISE project and how it uses the reference datasets that inspired this article will be given. Then, the aforementioned commodities dataset will be outlined to explain why encoding languages became relevant to the project.

This is followed by the historical background of the VOC and its archive. The VOC (1602–1798) established long-distance trading between Asia and Europe. In Asia, the European VOC employees encountered a world full of new languages, flora, fauna, and cultures. This heavily encouraged both an interest in the local languages and the borrowing of terms from these languages to describe concepts the Europeans did not have a word for. However, the VOC was also a violent and proto-colonial enterprise. Not all languages they encountered and potentially borrowed from survived their interactions with the company.

The article will then lay out the problems encountered when trying to encode this complicated relationship with language into structured data, with these challenges being subdivided into historical, linguistic, technical, and political challenges. Finally, the article will explain how the GLOBALISE project has tried to deal with some of these issues.

The inherently multilingual dimension of (proto)colonial enterprise like that of the VOC makes the languages of its information imperative to creating datasets. Still, it is far from the only historical context where this matters. Recent examples from topics as varied as the medieval Venetian labour relations Garzoni-project (Ehrmann et al. 2016), which needed to encode the names of occupations in a specific dialect, to the interspersing of Greek, Arabic, and Hebrew in Latin scholarly writing from the Republic of Letters (van Miert 2010), show that many archives are multilingual. Although this article is inspired by the challenges of correctly adding language tags to a SKOS thesaurus, the reflections in this article apply to any use case where the language of terms sourced from historical sources needs to be encoded.

ABOUT GLOBALISE

The use of Handwritten Text Recognition (HTR) has become increasingly prevalent in scholarly research as a method to unlock large archival corpora effectively. By training a computer model to undertake the labour-intensive task of transcribing documents, researchers can make bodies of work digitally available in ways that were previously far too expensive or laborious to attempt to unlock in full. A standout example of an archive that was too big to manually transcribe is the Dutch East India Company (VOC) archive housed in the National Archives in The Hague.

Throughout its existence from 1602 to 1798, the VOC created a staggering five kilometres of documentation. Within the archive, one can unearth not merely records on finance and trade but also encounter a meticulously detailed, albeit heavily biased, view of the Early Modern Asian world. The unique potential for research of the archive led to its designation as UNESCO World Heritage in 2003.

GLOBALISE is an infrastructure project that aims to unlock a specific segment of the VOC archive, known as the *overgekomen brieven en papieren* (OBP). The OBP, which can be translated as the 'letters and papers received' comprises the documents dispatched to the company management in the Dutch Republic from various Dutch East India Company settlements across the Indian Ocean world. By combining Handwritten Text Recognition and Natural Language Processing (NLP) techniques such as named entity recognition with a large amount of researched reference data, the project aims to make the archives not just searchable but researchable (Petram & van Rossum, 2022).

The first version of the HTR-generated transcriptions of the OBP was released in July 2023.[1] They were generated using HTR-engine LOGHI[2] and include roughly five million pages. Future iterations are planned to improve the quality of the data.

Rather than limiting its scope to merely providing transcriptions of the documents for online text search, GLOBALISE seeks to give researchers a robust set of tools to navigate the overwhelming amount of material generated by the HTR process. To help researchers work with those five million pages, GLOBALISE creates reference datasets to provide context for the entities found in the archive. One of these reference datasets will be discussed in the next section.

DATASET DESCRIPTION

OBJECT NAME

GLOBALISE Thesaurus – Commodities

FORMAT NAMES AND VERSIONS

RDF Trig

CREATION DATES

March 2022–February 2023

DATASET CREATORS

K. W. Pepping (Investigation), H. Vellinga (Investigation), M. Kuruppath (Investigation, Supervision), L. van Wissen (Data curation), M. van Rossum (Conceptualization, Supervision)

LANGUAGE

Dutch, German, English, French, Indonesian, Malay, Portuguese, Tamil (more to be added in the future)

LICENSE

CC-BY SA 4.0

REPOSITORY NAME

Dataverse (https://hdl.handle.net/10622/YAWDOV)

PUBLICATION DATE

2023-11-28

ABOUT REFERENCE DATASETS AND THE COMMODITIES THESAURUS

To help researchers conduct precise and focused searches in the archive, GLOBALISE is creating reference data in the Linked Open Data (LOD) format. This reference data compiles available information about a specific entity, such as a definition or temporal and spatial dimensions: in a sense, an encyclopedia-like information base. Identified occurrences of an entity in the

[1] https://hdl.handle.net/10622/JCTCJ2.

[2] https://github.com/knaw-huc/loghi.

HTR transcriptions of the archive can then be linked to the corresponding information in the reference data (Petram & van Rossum 2022).

One of the primary forms this reference data takes is that of taxonomical reference thesauri. Thesauri currently in development cover entities such as commodities, weights, measures, occupations, and document types. The primary data model used to create these thesauri is SKOS (Simple Knowledge Organization System) and its extension SKOS-XL. SKOS allows organising representations of knowledge organisation systems such as thesauri and controlled vocabularies as structured, linkable data (Allemang & Hendler 2011).

The first thesaurus the project set out to create was one on commodities. The first released version of this dataset contains 1400 of the products that were most shipped and traded by the VOC in early modern Asia. The project created this dataset using the large body of existing source publications, datasets, literature, and other research works available on the VOC and its activities, particularly the *Boekhouder Generaal Batavia* (BGB). The BGB consists of trade ledgers detailing the goods carried by 18th-century VOC ships, meaning every commodity was sourced from the VOC archives (Schooneveld-Oosterling & Knaap 2013).

The commodities in this list were then given definitions by looking them up in relevant literature, such as the VOC Glossary (Kooijmans & Schooneveld-Oosterling 2000) and the *Woordenboek der Nederlandsche Taal*.[3] Every unique commodity from these sources was modelled as a skos:Concept and given a nondescript unique resource identifier (URI). Each concept was then given one or more 'labels'. These labels represent words, or sequences of words, that were used by either the VOC archive or the literature to refer to the concept. The most commonly used word was used as the standard label (skos:prefLabel), with other words referring to the same concept modelled as 'alternate labels' (skos:altLabel). They were then classified within a wider hierarchical schema of commodities.

For illustration, one could take the concept of a *kris*. A kris has as its skos:definition 'a dagger often found on Java' and is also known as a *keris*. Based on this definition, this object should be classified as a skos:narrower (sub-type) of the concept of 'dagger,' which in itself could be classified as being a skos:narrower of the more general skos:Concept of 'weaponry.' This sub-categorization will result in a 'tree' of data, going from more general concepts to increasingly specific definitions. Concepts can have multiple labels: here 'Kris' and 'Keris' would both be the skos:prefLabel and skos:altLabel for the concept.

By structuring the information in the SKOS format, the dataset allows researchers to be much more precise in their queries. When interested in 'daggers,' they could reference the relevant URI within the data and include all relevant labels and narrower concepts within their query. This way, they would not just get the plain-text search result of *dolk* (the Dutch word for dagger) but also the highly relevant results that include a 'kris' or 'keris' (Nijman & Pepping 2023).

As stated in the introduction, best practice for SKOS recommends encoding the language of information, including labels, into the data. This recommendation led GLOBALISE to identify a number of challenges when encoding languages in a historical dataset like this.

THE MULTILINGUAL CORPUS OF THE DUTCH EAST INDIA COMPANY

To understand why encoding languages in this dataset can be difficult, it is necessary to move away from thinking of the VOC archive as composed entirely in Dutch. The archival material produced by transnational trading companies, be it the VOC or its contemporary, the English East India Company (EIC), was inherently multilingual. These records, in the case of the VOC, were born out of intercultural encounters that spanned two centuries.

Merchants are traditionally multilingual. Within Early Modern trading companies, being multilingual was often a prerequisite for employment and a successful career (Kaislaniemi 2017). As a result, the VOC sourced its employees from all over Europe, including many Danish, Swedish, English, and German speakers among its personnel (Gaastra 2007). These employees would create records in Dutch while trading and working in an Asian environment. Locally,

[3] https://gtb.ivdnt.org/search/?owner=wnt.

there may have been one or multiple languages of preference. Additionally, in many locations, VOC employees also had to communicate with merchant communities and diaspora from other regions who brought their languages with them (Gaastra 1992).

This cosmopolitan context of interaction resulted in company correspondence that was influenced by other languages. One of the ways in which this is visible is the VOC employees' interest in foreign languages. A famous example would be the director of Surat J. J. Ketelaar (1659–1716), an employee who used his experience working for the VOC to write and publish the first book on Hindustani grammar (Vogel 1936). The OBP itself contains similar clear indications of interest in the local language, such as a collection of papers written to aid the translation of the Bible into Malay.[4]

Sometimes, employees would include documents in other languages in their archives, expecting the recipient to be able to understand the original language.[5] In other cases, they would copy text in non-Latin alphabets for reference.[6] Looking at the large amount of borrowed words and texts in other languages, one could argue that the archives are only partially Dutch archives.

Portraying the Dutch East India Company as a peaceful trading enterprise has long been proven to be a misrepresentation. The rise of the company within Asia was leveraged by numerous violent and repressive undertakings, the full extent of which is still being discovered. In recent years, the company has been characterised as not just a trading endeavour but as a state-like actor. It colonised regions like Java, where it taxed and regulated the daily life of its large populations of local subjects and conducted diplomacy and war against foreign states (Stern 2011).

One of the most infamous examples of the atrocities committed by the VOC is the near extinction of the people of Banda in 1621, which shows that violence had an impact on language from the very start. The Dutch conquest of the Banda Islands from 1609 to 1621 involved the Dutch East India Company's efforts to control nutmeg production on the islands. After failed negotiations and the murder of a Dutch representative, Governor General Jan Pieterszoon Coen (1587–1629) led a brutal military campaign, resulting in mass killings and deportations. In their quest for nutmeg, the Dutch East India Company was consequently also responsible for the near extinction of the Bandanese language (Collins 2003).

As alluded to earlier when referencing the documents on Bible translation, the Dutch, functioning as a state-like actor, exhibited a distinct interest in promoting Malay as a language suitable for educational and religious purposes. Combined with their forceful displacements and killings of speakers of specific languages, the VOC had a serious impact on the linguistic landscape of Indonesia. Some languages, like Kelang, became extinct. Others, like Manipa, are endangered to this day (Collins 2003). As the next section will show, this makes it difficult to identify and encode these languages in the archive. However, it also makes it all the more important, as the VOC archive is one of the last places where these languages can potentially be found.

PROBLEMS ENCOUNTERED WHEN ENCODING

HISTORICAL/LINGUISTIC PROBLEMS

As shown in the previous section, the VOC had a complicated relationship with other languages. Borrowing a word or phrase from another language is not instantaneous; it has a strong temporal dimension. Words are often first adopted by a small set of specialists who try to describe a phenomenon outside of their 'regular' world of experience (which applies fully to the case of the VOC) and might then slowly spread to other speakers of the language (Sijs 1996). Terms borrowed from other languages to describe concepts within the VOC records survive there to this day.

A framework to think about this inclusion of foreign language in the records is the idea of 'code-switching,' particularly as expressed in the work of Kaislaniemi (2017) on the EIC. Simply put, there is a distinction to be made between two 'forms' of borrowing:

[4] Nationaal Archief, Verenigde Oost-Indische Compagnie (VOC) 1.04.02, inv. nr. 1902 fo. 1610–1677.

[5] Nationaal Archief, Verenigde Oost-Indische Compagnie (VOC) 1.04.02, inv. nr. 3790 fo. 2043.

[6] Nationaal Archief, Verenigde Oost-Indische Compagnie (VOC) 1.04.02, inv. nr. 3477 fo. 2360.

1. *Conventionalisation* is when a borrowing becomes conventionalised without changing. Obvious examples in the VOC corpus include phrases borrowed from Latin such as *folio, recto* and *verso*. Conventionalisation also includes the adoption of terms like *kris*.
2. *Free switching* refers to full-scale code-switching, where the author has some skill in both languages. This includes any creative use of a foreign language, including freely citing from foreign sources or making a play on words. Examples from the VOC corpus are the full citations from foreign texts and documents included in the archive.

The conventionalisation of words is of particular interest here. It describes a process, a changing understanding of a word. When encoding such a process into a dataset, there is a risk of reducing this process to a single, 'flattened' understanding (Van Erp 2023).

Take the earlier example of 'kris.' It was sufficiently conventionalised to warrant an entry in the modern-day Dutch dictionary *Van Dale*.[7] Whether that is a result of the Dutch East India Company's involvement with the modern-day region of Indonesia or of the later colonial ties between The Netherlands and the Dutch East Indies is not specified. Does that mean it can simply be tagged as a Dutch language term? Or should it be treated as a code-switch, as the early modern Dutch dictionary *Woordenboek der Nederlandsche Taal* (WNT) defines it as a word of Javanese origin?[8] In that case, arguments could be made that it should be tagged as a Javanese term, or as a Malay term, as the word has origins in both languages (Reid 2000).

In this case, SKOS does offer a solution: the term 'kris' can serve as a label for the concept in three distinct languages, namely Malay, Javanese and potentially Dutch. However, this example illustrates the significant complexity involved in accurately determining the language of a label. A definition taken from any of these three sources would be adequate to define and classify the concept in the dataset. Only together do they start to indicate the multilingual dimensions of the word in question.

TECHNICAL PROBLEMS

In the case of the VOC archives, the first problem is one of scale. When large amounts of archival material are transcribed by HTR, it is impossible to know beforehand what languages might be included (Liu & Smith 2020). This makes it difficult to identify loanwords, as automated methods often require identifying the borrowing and loaning language beforehand (Zhang et al. 2021).

Moreover, the VOC's behaviour resulted in the endangering and extinction of several languages. This tragedy has two practical results: there is no reference data of the language to aid automated identification, and these languages are often not supported by existing infrastructure for language tagging. The two-letter language standard ISO 639–1 does not include many extinct or endangered languages.[9] The extended version, the three-letter ISO 639–3, does, but in this case, only provides a tag for the endangered Manipa, not the extinct Kelang.[10] As a result, the popular SKOS-work suite Poolparty does not offer native support for these languages. This means that if a researcher discovers a concept in the archive that has as a label conventionalised word, or even a free switch, in Kelang, they would need first to find an acceptable machine-readable standard that could communicate that fact to others before being able to tag the label.

POLITICAL PROBLEMS

A final layer of complexity are cases where encoding a word as being in a particular language might be politically controversial or charged. A particularly relevant example in the context of the Dutch East India Company is the languages in South Asia. In India, language has become increasingly politically charged (Gould 2018). This led to attempts to remove words in Urdu,

[7] https://www.vandale.nl/gratis-woordenboek/nederlands/betekenis/kris.

[8] https://gtb.ivdnt.org/iWDB/search?actie=article&wdb=WNT&id=M035566&lemma=kris.

[9] ISO/TC 37/SC 2 2002 *ISO 639–1:2002 Codes for the representation of names of languages—Part 1: Alpha-2 code* [dataset]. Available at https://www.iso.org/standard/39536.html.

[10] ISO/TC 37 2008 *ISO 639–5:2008: Codes for the representation of names of languages—Part 5: Alpha-3 code for language families and groups* [dataset]. Available at https://www.iso.org/standard/39536.html.

a language associated with Islam by some political movements, from the shared vocabulary in favour of words in Hindi.[11] In such circumstances, tagging languages based on historical context could be perceived as controversial.

REMEDIAL STRATEGIES

The previous section has shown the variety of challenges that can be encountered when encoding the languages of historical data. This section will explain how GLOBALISE hopes to start addressing these questions and will draw inspiration from sociolinguistics to show how the increased availability of data can also help resolve these challenges.

Ideally, created infrastructures are flexible enough to support future explorations and developments. Linked Open Data triples and SKOS are accommodating to being adopted by others, with the structure of SKOS actively encouraging reuse and linking (Allemang & Hendler 2011). However, if the creators of the original dataset do not give other researchers access to editorial tools, the dataset becomes static once the creators stop working on it. As such, it is good to consider ways in which other stakeholders can contribute to the data, either during dataset creation or after.

One way in which GLOBALISE tries to achieve this is through organising data sprints and welcoming data from other contributors for (curated) inclusion during the project's runtime. Another option that is being explored is to allow users to suggest improvements, edits, and other improvements to the data from within the research interface.

A second step is to be open and transparent about the potential shortcomings and gaps within the data concerning language. It is important to convey that language tags may sometimes serve as provisional markers or oversimplifications. The creators of a reference dataset are ideally positioned to leverage their familiarity with the corpus to point future researchers in the right direction.

Opening up the large corpora through HTR, structured reference data and other new techniques also offer opportunities. Previously, large-scale research into language use required a large amount of manual work to provide an adequately utilisable corpus and to locate relevant words within that text. Some of that effort can now be delegated to (semi-)automated processes.

Take as an example the work that Kaislaniemi (2017) has done on the Corpus of Early English Correspondence Supplement (CEECSU). The CEECSU is an expansion of the CEEC, a large English language corpus for testing the applicability of sociolinguistic methods painstakingly built from thousands of transcribed pages. This includes material from the English EIC factory in Japan.[12] To gauge how comfortable EIC employees were with the words they borrowed from the Japanese language, Kaislaniemi looked into a form of glossing where the author used 'or' to explain the term (Kaislaniemi 2018). A gloss is an interpretive aid that helps the reader understand what an author understands a term to mean (Stenner 2020).

When the author of a document explains a code switch to the perceived audience, it is possible to measure three things. First, it allows assessing how the author understood the term: the *kris,* for example, is glossed in the archive as simply 'a short dagger'.[13] This allows comparison to how it matches up with existing definitions, the 'glosses' of other authors in the corpus and how these understandings change over time.

Secondly, it indicates what terms were considered 'foreign,' even if they have since become conventionalised like the *kris*. Finally, it indicates how well the author thought the recipient understood the code-switch. If a term is frequently glossed, the authors' confidence that the audience has conventionalised the code switch is low (Kaislaniemi 2017).

This sort of research, which can illuminate ways in which people wielded foreign terms, primarily needs a way of locating glosses in the corpus. With archives that are unlocked through HTR, this is increasingly simple, as a lot of heavy lifting can be done through a simple bit of code that searches for glossing indicators like 'or' for English and *of*/*ofwel* for Dutch.

[11] For more recent newspaper coverage of this trend, see (McCool & Hussain 2019).

[12] Corpus of Early English Correspondence Supplement, 2006.

[13] Nationaal Archief, Verenigde Oost-Indische Compagnie (VOC) 1.04.02, inv. nr. 1577 fo. 0229.

An infrastructure project like GLOBALISE can support such research by including metadata on document types, authors/senders, and the perceived audience of documents to give context to glosses and the instances of the entities. By providing such metadata, projects cannot completely shirk their responsibilities for proper language tagging, but it can help encourage future improvements.

CONCLUSION

The vast and multilingual nature of the Dutch East India Company archives brings the challenges of encoding languages in historical data to the fore. Although the incentive for this article was adding language tags to a SKOS thesaurus about an early modern trading company, the problems highlighted could apply to any use case where a historical context featuring multilingualism and code-switching needs to be encoded. The most important problems included:

- The dynamic nature of borrowing words and code-switching being both difficult to represent in structured data, and time-consuming to research.
- The lack of reference data for extinct languages complicates the identification of important voices.
- Existing standards like ISO often do not include extinct or endangered languages, hindering accurate representation from the perspective of a historian.
- The political sensitivity of encoding languages which are heavily intertwined with identity, like Urdu and Hindi in South Asia.

Handling all these problems simultaneously is difficult, and the rigidity of structured data can risk creating a false sense of certainty. As such, transparency regarding uncertainties and a flexible infrastructure for ongoing refinement are crucial. The languages in these archives should not be seen as 'just a roadblock' to creating structured data but as a starting point for further research and scholarly discourse. Encouraging open discussions about language use and inventing creative solutions for iterative improvement can incentivise future projects to engage with the linguistic dimensions of existing corpora.

In the case of GLOBALISE, acknowledging the linguistic complexities of the corpus and actively working on ongoing refinement can help turn the foreign languages and loanwords in the VOC archive from a challenge into a way to uncover hidden voices and critically examine the colonial past.

ACKNOWLEDGEMENTS

The author would like to thank Manjusha Kuruppath and Lodewijk Petram for their valuable feedback regarding this article and Henrike Vellinga, Manjusha Kuruppath, Leon van Wissen, and Matthias van Rossum for their work on the dataset that was the primary inspiration for the article.

FUNDING INFORMATION

This publication is part of the project GLOBALISE (with project number 175.2019.003) of the research programme Research Infrastructure, which is (partly) financed by the Dutch Research Council (NWO).

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR CONTRIBUTIONS

K. W. Pepping: Conceptualization, Writing original draft.

AUTHOR AFFILIATIONS

K. W. Pepping orcid.org/0000-0002-3747-706X
Huygens Institute for the History of the Netherlands, Amsterdam, The Netherlands

REFERENCES

Allemang, D., & **Hendler, J. A.** (2011). *Semantic Web for the working ontologist: Effective modeling in RDFS and OWL.* 2nd ed. Morgan Kaufmann/Elsevier. DOI: https://doi.org/10.1016/C2010-0-68657-3

Collins, J. T. (2003). Language death in Maluku: The impact of the VOC. *Bijdragen Tot de Taal-, Land- En Volkenkunde, 159*(2/3), 247–289. JSTOR. Available at https://www.jstor.org/stable/27868034. DOI: https://doi.org/10.1163/22134379-90003745

Corpus of Early English Correspondence Supplement. (2006). [dataset]. Available at https://www.helsinki.fi/en/researchgroups/varieng/research/corpus-of-early-english-correspondence

Ehrmann, M., Colavizza, G., Topalov, O., Cella, R., Drago, D., Erboso, A., Zugno, F., Bellavitis, A., Sapienza, V., & **Kaplan, F.** (2016). *From Documents to Structured Data: First Milestones of the Garzoni Project*. DHCommons, 2. Available at http://infoscience.epfl.ch/record/252904

Finnane, M., Kaladelfos, A., & **Piper, A.** (2018). Sharing the archive: Using web technologies for accessing, storing and re-using historical data. *Methodological Innovations*, *11*(2), 1–11. DOI: https://doi.org/10.1177/2059799118787749

Gaastra, F. S. (1992). The Organization of the VOC. In R. Raben & H. Spijkerman (Eds.), *De archieven van de Verenigde Oostindische Compagnie: The Archives of the Dutch East India Company (1602–1795)*. (pp. 11–25). Sdu uitgeverij.

Gaastra, F. S. (2007). *De geschiedenis van de VOC* (10. dr). Walburg Pers.

Gould, W. (2018). Rethinking Religion and Language in North India: The Hindi-Urdu Dispute and the Rise of Right-wing Populism. *Revista Canaria de Estudios Ingleses*, *76*, 29–44. DOI: https://doi.org/10.25145/j.recaesin.2018.76.03

Kaislaniemi, S. (2017). 7. The early English East India Company as a community of practice: Evidence of multilingualism. In E.-M. Wagner, B. Beinhoff & B. Outhwaite (Eds.), *Merchants of Innovation*. (pp. 132–157). De Gruyter. DOI: https://doi.org/10.1515/9781501503542-007

Kaislaniemi, S. (2018). Chapter 4. The Corpus of Early English Correspondence Extension (CEECE). In T. Nevalainen, M. Palander-Collin & T. Säily (Eds.), *Advances in Historical Sociolinguistics* (vol. 8, pp. 45–60). John Benjamins Publishing Company. DOI: https://doi.org/10.1075/ahs.8.04kai

Kooijmans, M., & **Schooneveld-Oosterling, J.** (2000). *VOC-glossarium, Verklaring van Termen, Verzameld uit de Rijks Geschiedkundige Publicatien die Betrekking hebben op de Verenigde Oost Indische Compagnie*. Instituut voor Nederlandse Geschiedenis.

Liu, S., & **Smith, D.** (2020). Detecting de minimis Code-Switching in Historical German Books. *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1808–1814). DOI: https://doi.org/10.18653/v1/2020.coling-main.163

Mader, C., Haslhofer, B., & **Isaac, A.** (2012). Finding Quality Issues in SKOS Vocabularies. In P. Zaphiris, G. Buchanan, E. Rasmussen & F. Loizides (Eds.), *Theory and Practice of Digital Libraries* (vol. 7489, pp. 222–233). Springer Berlin Heidelberg. DOI: https://doi.org/10.1007/978-3-642-33290-6_25

McCool, A., & **Hussain, T,** (2019). Poetry is the antidote': In fight against Hindu nationalism, India turns to verse. *The Guardian*, 1 November. Available at https://www.theguardian.com/books/2019/jan/11/poetry-is-the-antidote-in-fight-against-hindu-nationalism-india-turns-to-verse

Nijman, B., & **Pepping, K.** (2023). *Building a VOCabulary: The uses and challenges of thesauri for working with early modern recognized entities*. Brussels: DH Benelux. DOI: https://doi.org/10.5281/ZENODO.7973694

Petram, L., & **van Rossum, M.** (2022). Transforming historical research practices – a digital infrastructure for the VOC archives (GLOBALISE). *International Journal of Maritime History*, *34*(3), 494–502. DOI: https://doi.org/10.1177/08438714221112873

Reid, A. (2000). *Charting the shape of early modern Southeast Asia*. Institute of Southeast Asian Studies/Trasvin Publications L P (Silkworm Books).

Schooneveld-Oosterling, J., & **Knaap, G.** (2013). *Introduction to the database Boekhouder-Generaal Batavia; het goederenvervoer van de VOC in de achttiende eeuw (BGB) Bookkeeper-General Batavia; the circulation of commodities of the VOC in the eighteenth century (BGB)*. Huygens ING. Available at https://bgb.huygens.knaw.nl/?page_id=40

Sijs, N van der. (1996). *Leenwoordenboek*. Sdu uitgevers.

Stenner, R. (2020). Gloss. In R. Stenner (Ed.), *Oxford Research Encyclopedia of Literature*. Oxford University Press. DOI: https://doi.org/10.1093/acrefore/9780190201098.013.1066

Stern, P. J. (2011). *The company-state: Corporate sovereignty and the early modern foundations of the British Empire in India*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780195393736.001.0001

Thieberger, N., & **Tuohy, C.** (2017). From Small to Big Data: Paper manuscripts to RDF triples of Australian Indigenous Vocabularies. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 19–23). DOI: https://doi.org/10.18653/v1/W17-0103

Van Erp, M. (2023). Unflattening Knowledge Graphs. *Proceedings of the 12th Knowledge Capture Conference 2023* (pp. 223–224). DOI: https://doi.org/10.1145/3587259.3630082

van Miert, D. (2010). Language and Communication in the Republic of Letters: The Uses of Latin and French in the Correspondence of Joseph Scaliger'. *Bibliothèque d'Humanisme et Renaissance*, *72*(1), 7–34. JSTOR. Available at https://www.jstor.org/stable/20680039

Vogel, J Ph. (1936). Joan Josua Ketelaar of Elbing, author of the First Hindūstānī Grammar. *Bulletin of the School of Oriental and African Studies*, *8*(2–3), 817–822. Cambridge Core. DOI: https://doi.org/10.1017/S0041977X00141473

Zhang, L., Fabri, R., Nerbonne, J., & **Nerbonne, J.** (2021). Detecting loan words computationally. In E. O. Aboh & C. B. Vigouroux (Eds.), *Contact Language Library* (Vol.59, pp. 269–288). John Benjamins Publishing Company. DOI: https://doi.org/10.1075/coll.59.11zha

TO CITE THIS ARTICLE:
Pepping, K. W. (2024). Reflections on Encoding Languages in Historical Data: Working With the Multilingual Dimension of the Dutch East India Company Archives. *Journal of Open Humanities Data*, 10: 29, pp. 1–10. DOI: https://doi.org/10.5334/johd.176

Submitted: 16 October 2023
Accepted: 21 February 2024
Published: 12 April 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.