Linguistic Survey of India and Polyglotta Africana: Two Retrostandardized Digital Editions of Large Historical Collections of Multilingual Wordlists

Robert Forkel[1], **Johann-Mattis List**[1,2], **Christoph Rzymski**[1], **Guillaume Segerer**[3]
Max Planck Institute for Evolutionary Anthropology[1], University of Passau[2], CNRS, LLACAN[3],
Leipzig, Germany[1], Passau, Germany[2], Paris, France[3]
{robert_forkel, mattis_list, christoph_rzymski}@eva.mpg.de, guillaume.segerer@cnrs.fr,

Abstract

The Linguistic Survey of India (LSI) and the Polyglotta Africana (PA) are two of the largest historical collections of multilingual wordlists. While the originally printed editions have long since been digitized and shared in various forms, no editions in which the original data is presented in standardized form, comparable with contemporary wordlist collections, have been produced so far. Here we present digital retro-standardized editions of both sources. For maximal interoperability with datasets such as Lexibank the two datasets have been converted to CLDF, the standard proposed by the Cross-Linguistic Data Formats initiative. In this way, an unambiguous identification of the three main constituents of wordlist data – language, concept and segments used for transcription – is ensured through links to the respective reference catalogs, Glottolog, Concepticon and CLTS. At this level of interoperability, legacy material such as LSI and PA may provide a reasonable complementary source for language documentation, filling in gaps where original documentation is not possible anymore.

Keywords: Cross-Linguistic Data Formats, African languages, languages of India, retrostandardization, historical wordlist collections

1. Introduction

Along with the rise of Indo-European studies as a scientific discipline in the early 19th century, there has been an increased interest in documenting linguistic diversity through comparative wordlist collections. While seen critically by some researchers, wordlists and wordlist collections have played a major role in comparative linguistics and quite a few important insights have been made with their help, not only in historical linguistics, but also in linguistic typology. The past years have seen a drastic increase not only in large unified wordlist collections (Key and Comrie, 2016; Dellert et al., 2020), but also in new attempts to standardize the linguistic data represented in wordlist collections in a consistent way (Forkel et al., 2018). While these efforts have resulted in large cross-linguistic databases that can be investigated in various ways (see, for example Dediu 2023 or Jackson et al. 2019), many historical wordlist collections that were compiled long before the digital age have still not been integrated in the emerging web of connected wordlist datasets.

While digitization of historical collections can be seen as a purely technical question, recent attempts of retrostandardizing dialectal data (Geisler et al., 2021) have clearly shown that the process and the resulting comparable data can be very useful for linguistic investigations. In this study, we try to illustrate the benefits of retrostandardization by introducing two newly retrostandardized resources, the comparative vocabularies of the Linguistic Survey of India by Grierson (1928) and the wordlists published in Polyglotta Africana by Koelle (1854).

2. Materials

The starting point of our retrostandardization efforts are digitized versions of both the Linguistic Survey of India (LSI, Grierson 1928) and Polyglotta Africana (PA, Koelle 1854). The LSI is a large, multi-volume collection of linguistic information on various language varieties spoken in India, including languages from three major families (Indo-European, Dravidian, and Sino-Tibetan) and several isolates. Compiled in the beginning of the 20th century, the second part of the second volume provides comparative vocabularies for more than 350 language varieties, collected using a questionnaire of 168 concepts. Polyglotta Africana is a large collection of vocabularies of 200 varieties of African languages, collected by Sigismund Koelle in the 19th century, based on a questionnaire of 319 concepts.

Both resources were independently digitized, i.e. transcribed from (scans of) the original books by earlier projects. Representing digital versions of the historical collections, these digitization efforts can be seen as valuable resources in their own right. Considering the crucial role that standard representations of major aspects in wordlists play,

LREC-COLING 2024, pages 10578–10583
20-25 May, 2024.

such as the representation of concepts, linguistic forms, and language varieties, it becomes clear that creating these resources in a retrostandardized form is the next logical step. In this format, languages can be resolved to geolocations, identical concepts are linked across resources, and phonetic transcriptions are harmonized with the help of metadata from reference catalogs.

3. Methods

Creating an efficient pipeline to retrostandardize legacy data like LSI and PA has become a lot easier with the introduction of standardized linguistic data formats which can serve as the target of such a pipeline. Thus, the CLDF Wordlist format, first introduced by Forkel et al. (2018) and later expanded to provide additional levels of standardization List et al. (2022), allows us to skip any upfront design regarding serialization formats and jump right into data modeling.

CLDF provides interoperability on multiple levels: CLDF data consists mainly of tabular data in CSV files, thus can easily be accessed from computing platforms such as R or even spreadsheet programs. On a higher level of *semantic* interoperability, CLDF data – by virtue of being CSVW compliant – can be interpreted as typed data, rather than just tabular text data. Furthermore, CLDF data is described by an ontology associating tables and columns with specific cross-linguistic semantics. Lastly, since there is a well-defined conversion of CSVW data to RDF[1], CLDF data can be integrated into the Semantic Web.

One of the main tools by which CLDF supports interoperability across datasets are *reference catalogs*, i.e. catalogs holding metadata of basic objects like language varieties, concepts or transcription systems. Given that, among the first tasks of standardization is linking objects in a dataset to reference catalogs.

3.1. Language Identification

At the time when LSI and PA were compiled no standard codes for languages were available, so language varieties in these works are referenced by name. Fortunately, the additional information given in both works such as language classification and location allows identification with *languoids* in Glottolog in most cases (see Forkel and Hammarström 2022 for details on Glottolog and its structure). Having become the standard for the identification of historical and comparative languages in the past years, Glottolog has been the reference catalog of choice for the handling of language-related information in datasets provided in Cross-Linguistic Data Formats (Forkel et al., 2018).

3.2. Concept Mapping

Concept mapping is performed on the basis of the Concepticon reference catalog (List et al., 2023). Concepticon provides a comprehensive inventory of concepts used to elicit lexical data, each with a unique identifier. Concepts in a wordlist, e.g. in LSI or PA, can be linked by mapping a given form's gloss to Concepticon's respective concept ID. This, in turn, allows for a straightforward cross-linguistic comparison of concepts.

The mapping process can be done in a computer-assisted way (Tjuka, 2020) and the tools provided by the Concepticon project make this process transparent and effortless (Forkel et al., 2021).

3.3. Orthography Conversion

Phonetic transcriptions were harmonized with the help of orthography profiles (Moran and Cysouw, 2018). Orthography profiles are simple replacement tables that translate from one orthographic representation to the other while segmenting transcriptions into distinct sounds at the same time. For the conversion, the Lexibank workflow was used (List et al., 2022), by which target transcriptions follow the Cross-Linguistic Transcription Systems reference catalog, providing a standardized version of the IPA (List et al., 2021).

3.4. CLDF Creation

Converting the raw (digitized) data to CLDF is aided by `cldfbench` (Forkel and List, 2020), a framework for the creation, curation and retrostandardization of (cross-)linguistic datasets. `cldfbench` provides a consistent and reproducible workflow to manage linguistic data in a version-controlled environment. Additionally, all the principle components used in the creation of a CLDF dataset (e.g. mappings to reference catalogs) are tracked, thereby allowing for easy updates of datapoints or updates of the reference catalogs. Since reference catalogs are subject to change over time themselves, having a pipeline to recreate a dataset given new versions of catalogs is essential to future-proof resources.

The CLDF components used to model the data of our two resources are straightforward, representing the three major constituents of wordlist data: A `LanguageTable` holds metadata about the documented varieties, a `ParameterTable` lists the concepts and a `FormTable` provides the (segmented) word forms. The only difference to born-digital wordlists are scans of book pages of

[1] See `https://www.w3.org/TR/csv2rdf/`.

the original works, which are linked transparently using CLDF's `MediaTable` component, thereby providing fully traceable provenance for each form. CLDF's linking mechanisms are flexible enough to allow linking scans to concepts in the case of LSI or forms for PA, respectively (see Figure 1).

4. Results

The two language resources presented in this paper (Koelle 2023 and Grierson 2023) contain the raw (digitized) data, the ancillary data facilitating reference linking as well as the CLDF datasets.

4.1. Workflow Artefacts

While the main result of the retrostandardization process are the CLDF datasets, it should be noted that some intermediate results of the curation process are re-usable on their own: Concept lists can be included in Concepticon, thereby increasing the semantic variety of the catalog. Since orthographies used for transcriptions are often based on regional language documentation traditions, orthography profiles can be re-usable to segment lexical data with similar provenance (Anderson et al., 2018).

The CLDF datasets themselves are re-usable and interoperable in a number of ways described in the following sections.

4.2. Validation

Thanks to the transparent, enforceable semantics of CLDF, validating a CLDF dataset can be done automatically. This validation includes checking referential integrity, which is crucial for a multi-table dataset. Validity of cross-dataset references such as Glottocodes, Concepticon links and IPA transcriptions can also be done in an automated way, provided the software has access to the reference catalogs.

4.3. Visualization

Visualization plays a critical role in exploratory data analysis. Thus, we can view visualization as a good indicator for overall data re-usability and also interoperability, because visualizations are typically mediated through third-party software.

A first set of visualizations is already possible with off-the-shelf tools run on the CLDF data: The machine-readable, rich JSON metadata describing each dataset can be transformed to a human-readable data description in Markdown – which in turn can be rendered by services such as GitHub as HTML.

Every CLDF dataset can be converted automatically to a standard SQLite database. Since relational databases enjoy rich tool support, the schema of the resulting database can be easily visualized as entity-relationship diagram created via the `cldfviz.erd` command (see Figure 1).

A third standard visualization for cross-linguistic datasets are coverage maps, i.e. geographic maps depicting the locations of the languages covered in a dataset. Again, CLDF provides enough semantics to do this with off-the-shelf tools such as `cldfviz.map` (see Figure 2).

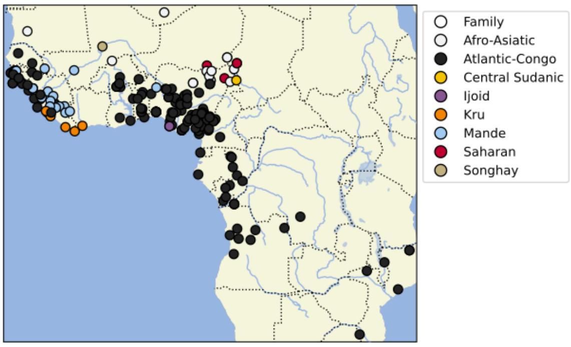
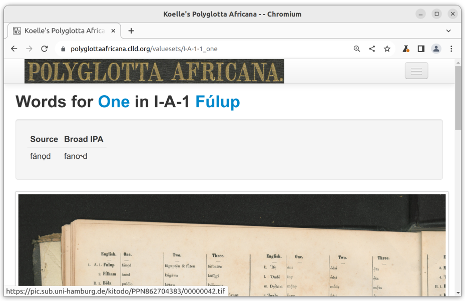


Figure 2: Coverage map of the PA dataset.

With `clld` (Forkel and Bank 2014), a web framework for the publication of CLDF data, datasets can be turned into websites. By employing the `MediaTable` component of CLDF, the original source data from scanned PDF files can be presented side-by-side with the converted CLDF version of the data (see Figure 3[2]).


Koelle's Polyglotta Africana - - Chromium
Koelle's Polyglotta Africa
polyglottaafricana.clld.org/valuesets/I-A-1-1_one
POLYGLOTTA AFRICANA.

Words for One in I-A-1 Fúlup

Source	Broad IPA
fánọd	fano˞d

https://pic.sub.uni-hamburg.de/kitodo/PPN862704383/00000042.tif

Figure 3: Word for "one" in Fúlup in the PA `clld` app.

Additionally, using a `clld` plugin for IPA charts, phoneme inventories can be computed from the segmented word forms and viewed with a browser (see Figure 4[3]).

[2] https://polyglottaafricana.clld.org/valuesets/I-A-1-1_one

[3] https://lsi.clld.org/languages/MALAYALAM#tipa

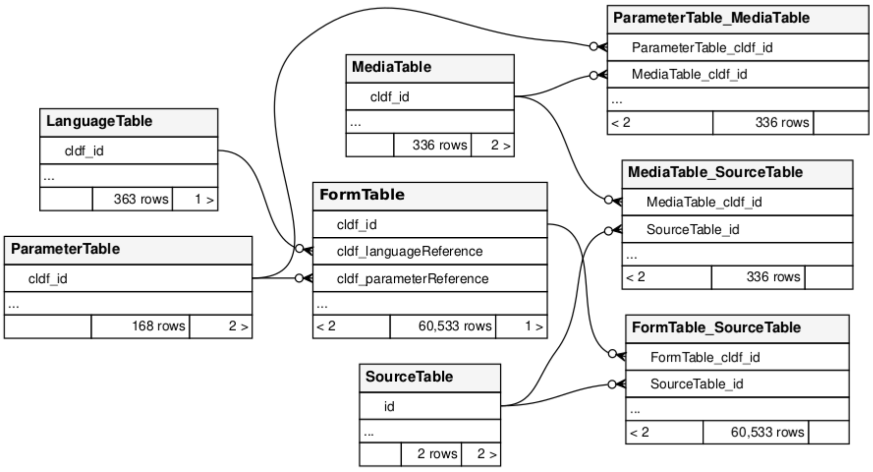
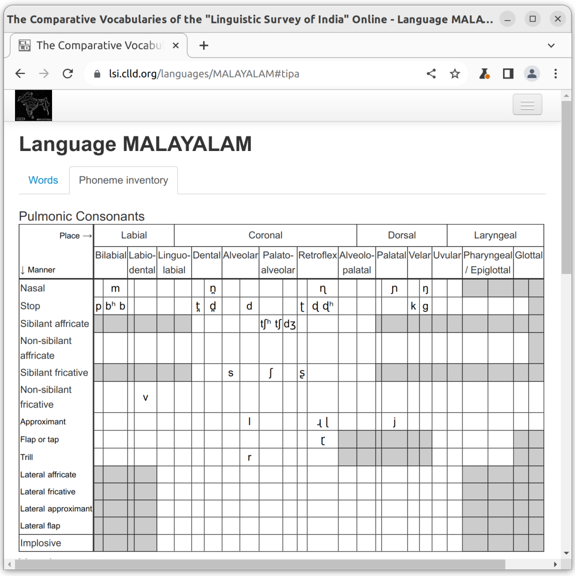


Figure 1: Entity-relationship diagram of the LSI data model.


The Comparative Vocabularies of the "Linguistic Survey of India" Online - Language MALA...
The Comparative Vocabu
lsi.clld.org/languages/MALAYALAM#tipa
Words
Phoneme inventory

Language MALAYALAM

Pulmonic Consonants

Place →	Labial			Coronal					Dorsal			Laryngeal	
↓ Manner	Bilabial	Labio-dental	Linguo-labial	Dental	Alveolar	Palato-alveolar	Retroflex	Alveolo-palatal	Palatal	Velar	Uvular	Pharyngeal / Epiglottal	Glottal
Nasal	m			n̪			ɳ		ɲ	ŋ			
Stop	p bʰ b			t̪ d̪	d		ʈ ɖ ɖʰ			k g			
Sibilant affricate						tʃʰ tʃ dʒ							
Non-sibilant affricate													
Sibilant fricative					s	ʃ	ʂ						
Non-sibilant fricative		v											
Approximant					l		ɻ ɭ		j				
Flap or tap							ɽ						
Trill					r								
Lateral affricate													
Lateral fricative													
Lateral approximant													
Lateral flap													
Implosive													

Figure 4: Consonant inventory of Malayalam in the LSI `clld` app.

4.4. Aggregation

The above examples of exploratory analysis exemplify interoperability of the CLDF data with software. In addition, CLDF also provides the basis for interoperability across datasets.

We can easily compare phoneme inventories by extracting segmented forms from our two datasets and matching them to the inventories documented in PHOIBLE (Moran and McCloy 2019). Since all datasets are available in CLDF, we can identify matching languages using Glottocodes and matching phonemes using CLTS.

In practice, this analysis can be done using CLDF SQL (see Figure 5), after the three relevant datasets have been converted to SQLite using the `pycldf` package. Note that while some knowledge of the schema of the CLTS and PHOIBLE datasets is necessary, the only piece of information necessary to specify the language Malayalam across all datasets is its Glottocode *mala1464*.

```
ATTACH DATABASE "phoible.sqlite" AS phoible;
ATTACH DATABASE "clts.sqlite" AS clts;
ATTACH DATABASE "lsi.sqlite" AS lsi;

SELECT DISTINCT s.grapheme, 'not in PHOIBLE', clts.name
FROM
  (
    WITH split(grapheme, segments) AS (
      SELECT '', f.cldf_segments || ' '
      FROM lsi.formtable AS f, lsi.languagetable AS l
      WHERE f.cldf_languagereference = l.cldf_id
        AND l.cldf_glottocode = 'mala1464'
      UNION ALL SELECT
        substr(segments, 0, instr(segments, ' ')),
        substr(segments, instr(segments, ' ') + 1)
      FROM split
      WHERE segments != ''
    ) SELECT grapheme FROM split
    WHERE grapheme != ''
  ) AS s
JOIN clts."data/sounds.tsv" AS clts ON clts.grapheme = s.grapheme
WHERE
  clts.name LIKE '%vowel'
  AND s.grapheme NOT IN (
    SELECT c.cltsgrapheme
    FROM (
      SELECT v.cldf_value AS grapheme
      FROM phoible.valuetable AS v
      WHERE cldf_languagereference = 'mala1464' and contribution_id = 1762
    ) as p
    JOIN (
      SELECT g.grapheme AS phoiblegrapheme, s.grapheme AS cltsgrapheme
      FROM clts."data/graphemes.tsv" AS g, clts."data/sounds.tsv" AS s
      WHERE g.dataset = 'phoible' and g.name = s.name
    ) AS c
    ON c.phoiblegrapheme = p.grapheme)
ORDER BY s.grapheme;
```

Figure 5: Aggregating data from three CLDF datasets via CLDF SQL.

ʌ	LSI	unrounded open-mid back vowel
ʌː	LSI	long unrounded open-mid back vowel
a	PHOIBLE	unrounded open front vowel
aː	PHOIBLE	long unrounded open front vowel
æ	PHOIBLE	unrounded near-open front vowel
ɨ	PHOIBLE	unrounded close central vowel
ʊ	PHOIBLE	rounded near-close near-back vowel

Table 1: Malayalam vowels in LSI vs. PHOIBLE.

Tabulating vowels that appear only in the LSI data or only in the PHOIBLE inventory we get the result shown in Table 1. Such an analysis might suggest refining the orthography profiles (and subsequently creating and releasing an updated version of the dataset).

4.5. Analysis

Thanks to advances in partly automated analysis methods, historical wordlists of the kind we presented here – if available in CLDF – can readily be fed into processing pipelines running automatic cognate judgments, and computing language phylogenies based on the cognate data (see Rzymski 2023). This means they can serve as one of the basic datatypes currently employed when it comes to phylogenetic approaches in historical language comparison.

5. Conclusion

The two resources presented in this paper – following in the footsteps of the dialect atlas by Geisler et al. (2021) – have established a well-specified workflow to retrodigitize legacy wordlists in a way that allows maximal computational re-use. Considering the rate at which languages fall into oblivion (Bromham et al. 2021) and comparing this to the potentially available legacy data[4], this may prove to be a pragmatic, yet promising way to complete our understanding of linguistic diversity.

6. Bibliographical References

Cormac Anderson, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A cross-linguistic database of phonetic transcription systems. *Yearbook of the Poznań Linguistic Meeting*, 4(1):21–53.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2021. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 6(2):163–173.

Dan Dediu. 2023. Ultraviolet light affects the color vocabulary: evidence from 834 languages. *Frontiers in Psychology*, 14.

Johannes Dellert, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel 1 · Hizniye Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Johannes Wahle, and Gerhard Jäger. 2020. Northeuralex: a wide-coverage lexical database of northern eurasia. *Language Resources & Evaluation*, 54:273–301.

Robert Forkel and Sebastian Bank. 2014. The clld toolkit. Workshop presentation given at "Language Comparison with Linguistic Databases: RefLex and Typological Databases".

Robert Forkel and Harald Hammarström. 2022. Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web*, 13(6):917–924.

Robert Forkel and Johann-Mattis List. 2020. Cldfbench. give your cross-linguistic data a lift. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, pages 6997–7004, Luxembourg. European Language Resources Association (ELRA).

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(180205):1–10.

Robert Forkel, Christoph Rzymski, and Johann-Mattis List. 2021. *PyConcepticon [Python library, Version 2.8.0]*. Zenodo, Geneva.

Hans Geisler, Robert Forkel, and Johann-Mattis List. 2021. A digital, retro-standardized edition of the tableaux phonétiques des patois suisses romands (tppsr). In M. Avanzi, N. LoVecchio, A. Millour, and A. Thibault, editors, *Nouveaux regards sur la variation dialectale*, pages 13–36. Éditions de Linguistique et de Philologie, Strasbourg.

George Abraham Grierson. 1928. *Linguistic Survey of India: Comparative Vocabulary*, volume 1. Office of the Superintendent of Government Printing.

Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Peter J. Mucha, Robert Forkel, Simon J. Greenhill, Russell D. Gray, and Kristen Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Mary Ritchie Key and Bernard Comrie. 2016. *The Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

[4] Glottolog lists more than 8000 bibliographic references tagged as containing wordlists, see `https://glottolog.org/langdoc?sEcho=2&sSearch_7=wordlist`

Sigismund W. Koelle. 1854. *Polyglotta Africana or Comparative Vocabulary of Nearly Three Hundred Words and Phrases in more than One Hundred Distinct African Languages*. Church Missionary House, London.

Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. *Cross-Linguistic Transcription Systems. Version 2.1.0*. Max Planck Institute for the Science of Human History, Jena.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(316):1–31.

Johann Mattis List, Annika Tjuka, Mathilda van Zantwijk, Frederic Blum, Carlos Barrientos Ugarte, Christoph Rzymski, Simon Greenhill, and Robert Forkel, editors. 2023. *CLLD Concepticon 3.1.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Steven Moran and Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press, Berlin.

Steven Moran and Daniel McCloy. 2019. cldf-datasets/phoible: Phoible 2.0.1 as cldf dataset.

Christoph Rzymski. 2023. From Old Data to Fresh Phylogenies - A Linguistic Data Journey in the Times of CLDF. Workshop presentation given at ICHL 26.

Annika Tjuka. 2020. Adding concept lists to Concepticon: A guide for beginners. *Computer-Assisted Language Comparison in Practice*, 3(1).

7. Language Resource References

Grierson, George Abraham. 2023. *CLDF dataset derived from Grierson's "Linguistic Survey of India" from 1928*. Zenodo. [link].

Koelle, Sigismund W. 2023. *CLDF dataset derived from Koelle's "Polyglotta Africana" from 1854*. Zenodo. [link].

Acknowledgements

This research was supported by the Max Planck Society Research Grant *CALC³* (JML, https://digling.org), the ERC Consolidator Grant *ProduSemy* (JML, Grant No. 101044282, see https://doi.org/10.3030/101044282) and the ERC Synergy Grant *QUANTA* (CR, Grant No. 951388, see https://doi.org/10.3030/951388). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them. We thank Taraka Rama for encouragement and support.