Title: The SEAlang Archives: Preservation, Discovery, and Access for the 'Scattered Literature' of Southeast Asian Linguistics

Sub-Field: Library (*not really sure*)

Authors / Affiliations / e-mails:

Doug Cooper, Center for Research in Computational Linguistics / doug.cooper.thailand@gmail.com

Saowapha Viravong, Australian National Library / sviravong@nla.gov.au

The *SEAlang Archives* collects, scans, indexes, and disseminates scholarly publication on Southeast Asian language and linguistics, and devises and tests innovative approaches to aggregating and exposing the field's scattered literature. Archived texts include festschrifts, conference proceedings, working papers, special collections, and regular and itinerant journals, as well as the extensive unpublished materials – field notes, theses, and unfinished lexicons – that have particular importance to the linguistics community.

Publication of Southeast Asian linguistics literature is broadly distributed, with substantial contributions from the U.S., Australia, Europe, and Southeast Asia itself. These texts form a well-defined and coherent central body of work, roughly split between journals, regular conference proceedings, and special publications. But few of these publications have strong institutional sponsorship – even long-established journals change homes regularly – and collection in libraries is increasingly hit-and-miss. Nor does traditional book-oriented cataloging practice always guarantee that individual articles are listed, meaning that an article cannot necessarily be discovered even if the text it is in might be available.

Thus, even as interest in studying and documenting the endangered languages that abound in Southeast Asia is on the rise, the field's mainstream literature is increasingly inaccessible, and in some cases unknown, to a new generation of students and field researchers.

The SEAlang Archives directly addresses the dual problems of discovery and access by assisting in resource discovery, then making texts available immediately in a readable image format such as PDF or DjVu. The Archives system architecture rests on a specification of the TEI-compliant XML markup conventions used to tag individual index entries, and an implementation of software that builds a browsable Web-based 'archive of archives.'

Within the Archives, each source publication retains its unique identity as a journal, conference report, monograph, or other collection. Because each source's branch of the full Archives tree can be navigated independently, the complete branch can be incorporated directly and transparently into the publication's own website if desired. At the same time, the Archives generates sets of aggregated content pages that group articles by publication year, author(s), subject(s), and language(s) or region(s) of concern, and which are readily found and indexed by Google, Yahoo, and other discovery engines.

We will describe the design and implementation of the SEAlang Archives, discuss the contributions it makes to preservation and discovery of the scattered literature of Southeast Asian linguistics, and invite broader participation in the archiving program.