

Scott Farrar, University of Washington
Using canonical typology to achieve e-Linguistics

With the rise of the Web as an important medium for displaying and curating typological data, the next generation of linguists will have un-envisioned access to thousands of data sets on many of the world's remaining languages. It is not out the question that many or most of the grammars ever published and possibly many sets of field notes will be digitized and stored in a Web accessible format (cf. Google Book Search). At the same time, we as a field are faced with the prospect of leaving the primary data of our field in disarray for the next generation. What are the intended meanings of annotation elements? How does one data set relate to another? How do we determine whether two data points are the same? As more and more resources are digitized, there is an ever-increasing need for transparency regarding these issue. Standardization of terminology is certainly one path towards achieving such transparency. Beyond standards and terminology, however, there are more basic challenges, including a clear articulation of a conceptual framework for specific areas of linguistics including, among many others, canonical typology. Put another way there is currently no broadly accepted logic or computer language that would allow arbitrary linguistics data to be automatically processed, i.e., no possibility of an 'e-linguistics'. In this talk, I explore laying the foundations for such an enterprise with the combined use of ontologies and recent Web technologies. In particular, I will address work within the GOLD Community of Practice whose aim is the creation of a General Ontology for Linguistic Description and to apply the resulting e-Linguistics framework towards issues in canonical typology.