Building Indian Language Digital Library Collections: Some Experiences with Greenstone Software

B.S. Shivaram and T.B. Rajashekar

National Centre for Science Information (NCSI),
Indian Institute of Science,
Bangalore 560 012, India
{shivaram, raja}@ncsi.iisc.ernet.in

With its diverse cultural and linguistic heritage, India today produces significant volume of digital material in Indian languages. This has been facilitated by increasing availability of word processing systems supporting Indian languages and their use in various areas including e-governance; education and research; and mass media. There is need for digital library software for organizing and provision of access to this material. Such software has to meet two prime requirements: Indexing and searching of documents in Indian languages (full text and metadata), and customizing the collection user interface in Indian languages. Further the software should be able to handle Indian language material in different encoding formats and fonts. Majority of Indian language material available online today seem to follow one of the three encoding strategies: ISO 8859-1 and Windows 1252 series character sets, with custom fonts; ad-hoc (font-specific or user defined) encoding schemes; and Unicode character set. Search and retrieval requirements would include features such as word truncation and alphabetical sorting. Cross-language material searching is an advanced search requirement. Greenstone is a very popular open source software used today for creating digital library collections. Main objective of this study was to assess capabilities of this software in creating Indian language digital library collections with above mentioned requirements for indexing, searching and display. We gathered five sample collections in two Indian languages Hindi and Kannada, in different encoding formats, for this study. For each of these collections, we assessed the multilingual support of Greenstone with respect to collection building; search and retrieval; and interface design. We used the 'Collector' approach of Greenstone to build the five collections. We could successfully build the collections. Limitations were found in handling metadata in Indian languages using the 'GLI' approach. We present details of internal mapping of character sets carried out by Greenstone during collection building process. We could successfully carry out simple keyword and Boolean searches on these collections. We discuss details of search features. Viewing results requires installation of suitable fonts at the operating system level and configuration of the browser. We found limitations in sorting. Greenstone does not support cross-language searching. In terms of users interface, Greenstone has in-built support for customizing the interface for well known languages. It also supports designing customized interface for other languages. We could successfully design desired user interface for the test collections in Hindi and Kannada. Overall, Greenstone appears to be a versatile software for building Indian language digital library collections, with some limitations.

Z. Chen et al. (Eds.): ICADL 2004, LNCS 3334, p. 661, 2004.