

Text Analysis and Pattern Detection

Lewis LANCASTER

University of California, Berkeley, USA

buddhst@berkeley.edu

When we take a digital collection of texts such as the Chinese Buddhist canon composed of 52 million individual characters listed under 1500 titles, the great challenge is to empower the user. This particular set of data is of interest not only for its philosophical and religious thought but also for the appearance of place names, personal names, time periods and events. These elements of Where? When? Who? and What? are so ubiquitous that they cannot be placed into a single domain or a particular metadata vocabulary. We need a vocabulary that can apply to all of them. These four elements constitute the basis for finding references to any topic of research. They appear in specialized library resources: biographical dictionaries, subject catalogs, chronologies, and gazetteers. Such resources and procedures are the primary tools used by traditional library reference services. However, the digital environment is still weak in providing an effective counterpart. We lack the equivalent of a reference desk for the users of digital library material. The past efforts of the Electronic Cultural Atlas Initiative (www.ecai.org) and the future work are focused on how existing and emerging standards and protocols can be used or adapted to access data.

Using the Chinese Buddhist canon as a backdrop for exploring these issues, the use of 3-D and graphic display are shown for models of how users can be supported. These interfaces strive to offer a dynamic experience for canonic research by combining multiple modalities (text, images, maps, audio, video, 3D graphics, etc.) and contextualizing them in space and time.