Aug 17, 2007

Outsourcing in India

by Teri Tan

Content services jobs heading for India used to be 100% in English. But with globalization and localization on every astute publisher's mind, releasing co-editions of core products and ancillaries is becoming commonplace. At the same time, English- language publishers aren't the only ones seeking out Indian vendors: Asian, Middle Eastern, Eastern European and Scandinavian publishers are following in their footsteps. As a result, languages with double- byte character sets (Chinese, Japanese), elaborate hyphenation and diacritics (Thai, Croatian), and right- to- left or bidirectional text (Arabic, Hebrew) have ceased to be as perplexing as they were a decade or so ago. For vendors, anything that is recognizable by computers and can be tagged in XML is fair game. Language- specific rendering scripts have made it possible.

Non- English newspaper conversion/ digitization has long been a niche segment. V- p for international sales Amit Vohra recalled, “One of our earliest projects involved producing 50,000 pages of Arabic text, and we hired 30 linguists to ensure that both language and content were accurate. More recently, one 200,000- page Danish news agency project arrived at our door with a six- month deadline. We extracted the text using OCR technology and integrated Danish- language dictionary and hyphenation support into the QC workflow. The content was then converted into XML.”

Planman has also been digitizing Norwegian newspapers-- the daily Aftenposten as well as Adresseavisen and Trondades-- for the National Library of Norway. “We usually receive microfiches, which are scanned at 300 dpi grayscale. We process the scanned pages according and deliver the final output in several formats: issue- level PDF, page- level PDF, JPEG2000 compressed image and XML,” explained director Sourav Chatterjee.

Every month, Planman processes about five non-- English- language projects-- two newspaper digitizations and three general content conversions-- which make up roughly 40% of its output. Besides the languages mentioned above, it has also tackled French, German, Welsh, Swedish and Hebrew projects. “Text extraction and quality assurance are the biggest challenges in non- English projects, especially when special characters and diacritics are involved,” added Vohra, whose team has just completed two Spanish- language el- hi projects comprising a total of 2,500 pages that require the full treatment: composition, formatting and proofreading.