Best Practice Poster: Best Practices for Complex Diacritics Handling in CONTENTdm

Jason W. Dean
University of Arkansas Libraries, USA
jwdean@uark.edu

Deborah E. Kulczak
University of Arkansas Libraries, USA
dkulczak@uark.edu

Keywords: CONTENTdm, diacritical marks, indexing, UTF, encoding

In order to ensure the best possible access for materials held by libraries and archives, these institutions must employ special accent and punctuation marks when transcribing or transliterating languages other than English. These marks are called diacritical marks by the library community. Their use in MARC cataloging is widespread, as is their use in library catalogs. However, users of digital content management systems (CMS), such as CONTENTdm encounter difficulty in ensuring appropriate diacritical marks are read by the CMS when metadata is imported or migrated into such a system. These problems further compound searching issues for the user, as noted by Bar-Ilan and Gutman (2005) in the *Journal of Information Science*. Little literature and few instructions exist to assist users in working with these diacritical marks. However, some pertinent literature exists on the subject.

In Hongyan Jing's essay for an IEEE symposium on speech synthesis (2002), the author in discussing Italian highlights the ubiquity of all types of diacritical marks. His work states that of 445,626 entries in a dictionary, 4.9% of these entries include a diacritical mark. Though the number is not high, for libraries and archives this number represents a barrier to access and description that must be overcome. Tull and Straley's article in *Library Hi Tech* (2003) covers the issues presented in sorting and searching in relation to diacritical marks. Most literature discusses formatting text in UTF-8 or similar UTF standards however some literature discusses the use of ASCII. This poster focuses on the use of UTF-8, which is required by CONTENTdm to ingest diacritical marks correctly.

The research and work behind this poster came largely from a recently completed project at the University of Arkansas Libraries that dealt with metadata and items in a plethora of languages, from English and French to Quapaw, many of which required the use of unusual diacritical marks. The authors were responsible for the ingestion of metadata into CONTENTdm and encountered several issues with complex diacritical marks presented by the disparate languages in this project. What follows is the procedure arrived at and now codified in a metadata "cookbook."

The handling of these diacritical marks was primarily in three areas: controlled vocabularies in CONTENTdm, transcripts, and loading metadata spreadsheets.

Creating and importing a controlled vocabulary list is most easily done in Notepad++ and encoded as "UTF-8 without BOM." Using these settings, diacritical marks ingested into CONTENTdm will be maintained using this encoding setting and following the CONTENTdm instructions for loading a controlled vocabulary.

Transcripts are best handled using a similar procedure. Transcripts are created in Notepad++, and saved as "UTF-8 without BOM" as the encoding setting. However, some transcripts might be loaded at the same time as metadata in a spreadsheet. In this case, if the spreadsheet is created in Excel, the user must use the "Arial Unicode MS" font for data entry. When data entry is complete, use the Save As command to save the spreadsheet as a tab-delimited text file. In the Save As dialog box, select "Unicode Text" from the "Save as type" menu. After selecting "Unicode Text", select the "Tools" box to the left of the "Save" button. Select the "Encoding" tab

in the "Web Options" dialog box. In the "Save this document as" box, select "Unicode (UTF-8)" from the drop-down menu. Select "OK" then "Save" in the Save As menu.

Metadata spreadsheets and tab-delimited files present a similar set of challenges for diacritical marks loading into CONTENTdm. In this case, if the spreadsheet is created in Excel, the user must use the "Arial Unicode MS" font for data entry. When data entry is complete, use the Save As command to save the spreadsheet as a tab-delimited text file. In the Save As dialog box, select "Unicode Text" from the "Save as type" menu. After selecting "Unicode Text", select the "Tools" box to the left of the "Save" button. Select the "Encoding" tab in the "Web Options" dialog box. In the "Save this document as" box, select "Unicode (UTF-8)" from the drop-down menu. Select "OK" then "Save" in the Save As menu.

References

Bar-Ilan, Judit, and Tatyana Gutman. (2005). How do search engines respond to some non-English queries? Journal of Information Science. 31(1), 2005, 13-28.

Jing, Hongyan. (2002). Identifying accents in Italian text: a preprocessing step in TTS. Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002, 151-154.

Tull, Laura, and Dona Straley. (2003). Unicode: Support for multiple languages at the Ohio State University Libraries. Library Hi Tech. 21(4), 2003, 440-450.