Diacritics and Special Characters: How to avoid gobbledygook in item displays

Charity K.M. Stokes, MLS, MA, Texas A&M University Libraries

Diacritics and Special Characters– What are they?

- **Diacritic** : a mark near or through an orthographic or phonetic character or combination of characters indicating a phonetic value different from that given the unmarked or otherwise marked element. Also known as an 'accent mark'
  - Examples
    - ❖ *Acute accent - café, cliché*
    - ❖ *Cedilla – façade*
    - ❖ *Grave accent- learnèd*
    - ❖ *Umlaut – Brontë, naïve*
    - ❖ *Tilde – canon, piña colada*

- **Special characters:** A character that is not an alphabetic or numeric characters. Punctuation Marks and other symbols are examples of special characters, as well as mathematical symbols. Superscript and subscript also falls into this category.

  - **Punctuation Examples:**
    ? . ' " : ; ! & - -- / \ { } [ ]

  - **Math symbol Examples:**
    ≈ ≠ ÷ √ √x ∉ ∞ ∃∅ ≡ ⊥ ∝ α β γ Σ θ ϕ π ε Δ ∘ ⊃ ⊂ ∀ ∴ ∵

- Superscript and subscript: Although these seem to be a combination of known alphanumberic symbols, in the behind the scenes coding each one is 'constructed' using code, so an item abstract may include something like this
  $(C_2H_5)_3SiH + O_3 → (C_2H_5)_3SiOOOH → (C_2H_5)_3SiOH + O_2(^1\Delta_g)$

Problem

Texas A&M University Libraries was charged with ingesting the digitized version of thesis and dissertations (electronic thesis and dissertations, AKA ETDs) into the institutional repository. Metadata was harvested from the MARC (Machine Readable Cataloging) records found in the OPAC. but we noticed that those records that had diacritics or special characteriss had 'gobbledygook' instead of the characters. So we began to explore why.

Where Is metadata Getting Corrupted?

In examining each step of the workflow, we discovered that there were three points where the diacritics and special characters were corrupted

- Source of metadata
  If the metadata came from an outside source, such as MARC or a website there was a high chance that the coding for diacritics and special characters would be corrupted.
- Opening metadata in Excel
  If the metadata was was just 'opened' in Excel, even if displayed correctly, Windows tends to impose the Western character encoding standard, which doesn't reconize many diacritcs and special characters, and can even see quotation marks as wrong
- Creating the Simple Archival File
  With some SAF programs, the Windows OS can over-ride the programing and impose Western, even though it gives no indication of it happening.

Best Practices

- **ALWAYS SAVE TO UTF-8**
  This is especially true when creating the metadata manually. But even when exporting MARC data to a .bib file, ensure that you are exporting in UTF-8, even if using MARCedit.
- **BE AWARE OF EXCEL AND WINDOWS OS**
  Opening a file directly in Excel automatically imposes Western and causes the diacritics and special characters to be corrupted. The only way to avoid this is to iport the.csv file after Excel is open each and ever time.
- OR **USE A DIFFERENT SPREADSHEET PROGRAM**
  GoogleSheets or Open Refine are good programs to use to avoid corrupting the data. They rely on UTF-8 and do not change the underlying coding for special characters

What to do if diacritics and/or special characters are corrupted

- Explore each step to discover where metadata is becoming corrupted
  This is a tedious process, be can be necessary to discover where the UTF-8 Coding is being overwritten. It's best to break down the workflow and check the coding at each step
- UseNnotepad++ to ID problem areas
  Sometimes the metadata looks 'right' but when ingested into DSpace, the diacritics are corrupted. Cut and past the metadata into Notepad++ - which will display as it would look in DSpace. Be aware, although that the different fonts used can also cause issues.

WARING! Sometimes It Can't be fixed

It may be that your metadata is corrupted from the beginning because your source does not use UTF-8. If you are harvesting metadata from database, or you are sent the metadata as a .xlsx file, you will have gobblegook. The best solution is to just clean up those records once they are in Dspace.

VS