Digitizing Multilingual Newspapers for NDNP Grants

National Digital Newspaper Program Awardee Conference
September 17, 2015

Sarah Lynn Fisher
University of North Texas Libraries
sarahlynn.fisher@unt.edu
What you might know…

- Language encoding at the TEXTBLOCK level
- One code per TEXTBLOCK
- ALTO TEXTBLOCK language attribute
- NDNP requirement for all non-English codes: three-digit codes (ISO 639-2 specification)
Prior to digitization...

• Review MARC record for accuracy (041 and 546)
• Check microfilm as quality will affect OCR accuracy
• Communicate with your vendor
Multilingual Titles

• OCR engines can be set to detect more than one language
• Full collation vs. AutoDetect
• AutoDetect may reveal unknown multilingual textblocks
Quality Review

• Single language: check language codes in DVV at beginning and end of reel and 10% in batch
• Multilingual titles: search OCR language codes
• Search for names, common terms in OCR
• Is microfilm quality an issue?
• Communicate with LC liaison
One Strategy

• Process reel for more than one language (e.g. English and Spanish)
• Review results
• If needed: find and replace language codes in .xml using vendor tool or program like Dreamweaver
Batch Ferret OCR: Notes for vendor

1) sn84020617
Title: The morning journal. : (Albuquerque, N.M.) 1884-1886
Reel(s): 00296025239
Language(s): English
-Removed "spa" codes

2) sn84020626
Title: Santa Fe new Mexican and review (Santa Fe, N.M.)
ADD SPANISH TO MARC RECORD
Reel(s): 00296025215
Language(s): English, some Spanish
*OCR accuracy 70-80%

3) sn84020631
Title: Santa Fe daily new Mexican (Santa Fe, N.M.) ADD SPANISH TO MARC RECORD?
Reel(s): 00296025227
Languages(s): English, tiny amount of Spanish
-Remove "spa" codes except:
-Spanish in 1895-11-30, 1895-12-02 pg. 3
- 1895-05-10
- 1895-05-09
- 1895-05-08

4) sn84031081
Title: Albuquerque morning journal (Albuquerque, N.M.)
Reel(s):
00296025252 - remove "spa" codes
00296025264 - remove "spa" codes
00296025276 - remove "spa" codes
00296026288 - remove "spa" codes
00296025240 - remove "spa" codes
0029602529A - remove "spa" codes
Language(s): English

5) sn86063567
Title: The Albuquerque morning journal (New Albuquerque, N.M.)
Reel(s): 00296025239
Language(s): English - remove "spa" codes

6) sn92070564
Title: Alamogordo news-advertiser (Alamogordo, Otero County, N.M.)
Reel(s): 00296024892
Language(s): English - remove "spa" codes

7) sn92070566
Title: The Alamogordo news (Alamogordo, N.M.)
Reel(s):
00296025203 - remove "spa" codes except below
00296024892 - remove "spa" codes
Language(s): English, tiny amount of Spanish
-Spanish in 1906-11-17, 1906-12-08, 1907-11-19, 1909-01-30
Not for LC:

8) sn93061397
Title: La Estrella mejicana (Albuquerque, N.M.)
Reel(s): 00296025239
Language(s): Spanish
Change "eng" to "spa" page 1

9) no LCCN
Title: La Opinion Publica (Albuquerque, N.M.)
Reel(s): 00296025239
Language(s): Spanish, small amount of English
Example: *El Nuevo Mexicano and Weekly Santa Fe New Mexican*

- Began as a Spanish-language newspaper
- English language supplement beginning in 1912
- Filmed together on reel with pages out of order
- Some research required!
Share...

• Your experiences
• Potential problems and successes
• Quality Review strategies


Las Vegas daily optic. (Las Vegas, N.M.), 18 October 1899, pg. 3. Chronicling America: Historic American Newspapers. Lib. of Congress. Citation forthcoming.