Processing Non-English Content

Andrew Weidner
NDNP New Mexico
Overview

- Vendors
- Workflow
- QR Tools
- Alternatives
Vendors

- Communication: start early, ask questions
Vendors

- Communication: start early, ask questions
- One language vs. Multiple languages
Vendors

- Communication: start early, ask questions
  - One language vs. Multiple languages
  - Processing Level
Vendors

- Communication: start early, ask questions
- One language vs. Multiple languages
  - Processing Level
    One language = title
Vendors

• Communication: start early, ask questions

• One language vs. Multiple languages

• Processing Level

  One language = title
  Multiple languages: title, reel, issue, page, article
Vendors

• Communication: start early, ask questions

• One language vs. Multiple languages

• Processing Level
  One language = title
  Multiple languages: title, reel, issue, page, article

• Pricing / Rework
Workflow

- Know your content: MARC record, essay research
Workflow

- Know your content: MARC record, essay research
- Microfilm evaluation: confirmation / discovery
Workflow

- Know your content: MARC record, essay research

- Microfilm evaluation: confirmation / discovery

  Best to find new content during film eval
Workflow

- Know your content: MARC record, essay research
- Microfilm evaluation: confirmation / discovery
  Best to find new content during film eval
- Batch QR: characterize content / check OCR quality
Know your content: MARC record, essay research

Microfilm evaluation: confirmation / discovery

Best to find new content during film eval

Batch QR: characterize content / check OCR quality

QR discovery = OCR rework
QR Tools

- Command Line: discover new content
QR Tools

- Command Line: discover new content

```bash
find . -name "*.xml" -exec grep -HiL "aviso" {} \
```
• Command Line: discover new content

```bash
find . -name "*.xml" -exec grep -Hil "aviso" {} \;
```
• Command Line: discover new content

```bash
find . -name "*.xml" -exec grep -Hil "aviso" {} \;
```
QR Tools

- Command Line: locate & quantify encoded content
QR Tools

- Command Line: locate & quantify encoded content

```bash
find . -name "*.xml" -exec grep -Ho "language=""spa"""" {} "; | uniq -c
```
QR Tools

- Command Line: locate & quantify encoded content
  
  `find . -name "*.xml" -exec grep -Ho "language="spa"" {} \; | uniq -c`
QR Tools

- Web browser: check OCR accuracy
QR Tools

- Web browser: check OCR accuracy
QR Tools

- Web browser: check OCR accuracy
Alternatives

- ASCII Text Editor: edit pages
Alternatives

- ASCII Text Editor: edit pages
- Find & Replace: edit entire issues/reels
Alternatives

- ASCII Text Editor: edit pages
- Find & Replace: edit entire issues/reels

language="spa" → language="eng"
Alternatives

- ASCII Text Editor: edit pages
- Find & Replace: edit entire issues/reels
  
  language="spa"  →  language="eng"

- Unencoded non-English content already on ChronAm?
Alternatives

- ASCII Text Editor: edit pages

- Find & Replace: edit entire issues/reels
  \[
  \text{language} = "\text{spa}\" \rightarrow \text{language} = "\text{eng}\"
  \]

- Unencoded non-English content already on ChronAm?
  Reprocess OCR & deliver overwrite content
Alternatives

- ASCII Text Editor: edit pages
- Find & Replace: edit entire issues/reels

  language="spa"  →  language="eng"

- Unencoded non-English content already on ChronAm?

  Reprocess OCR & deliver overwrite content
  Unencoded content is discoverable in basic search
Alternatives

- ASCII Text Editor: edit pages
- Find & Replace: edit entire issues/reels
  
  language="spa"  →  language="eng"

- Unencoded non-English content already on ChronAm?
  Reprocess OCR & deliver overwrite content
  Unencoded content is discoverable in basic search
  Only encoded content is discoverable with
  language specific Advanced Search
Questions ?