LATIN AMERICAN DIGITAL INITIATIVES

INF385T, Fall 2017. Project by Itza Carbajal, Vivek Khetan, Emma Whittington

Overview

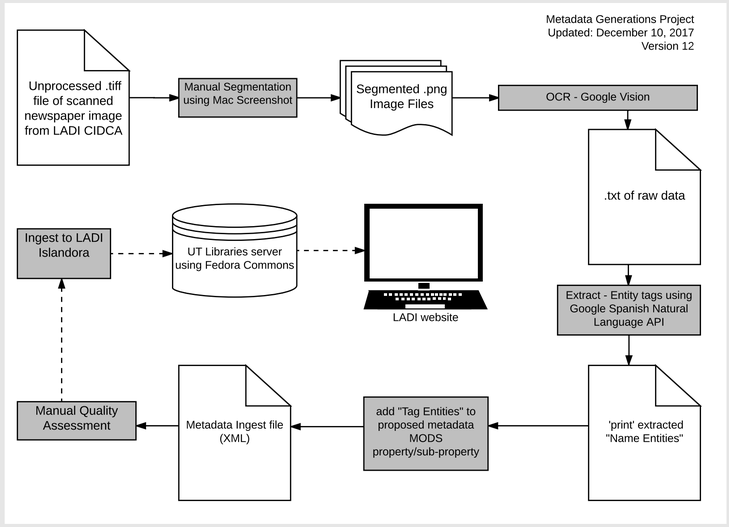
Our project uses Google's Vision and Language APIs to generate OCR text transcriptions and name entity values for LADI's corpus of *La Información* newspapers. The metadata we generate feeds into and enhances LADI's existing, MODS-compliant metadata schema. The added value of this project is that the name entity values generated for each transcript— stored as a list of "topics" alongside the object's other metadata — function as key words/search terms. These 'topics'/'key words' ultimately provide LADI database users with a more robust, searchable database.

Context

The Latin American Digital Initiative (LADI) is a collaboration between LLILAS Benson Latin American Studies and Collections at The University of Texas at Austin, the University of Texas Libraries, and Latin American partner institutions to "preserve and provide access to unique archival documentation from Latin America, with an emphasis on collections documenting human rights, race, ethnicity, and social exclusion in the region."

LADI hosts digitized collection material — provided to them by various partner sites — on their online repository platform (stored as .tiff images).

System Architecture



The final version of our system architecture. Solid arrows show our project's deliverables while dotted arrows show how LADI would integrate this metadata.

Metadata Schema

The OCR transcripts and Name Entities we generate expand LADI's existing metadata schema to include one new, MODS-compliant **<note>** property and series of multiple (depending on the script's output) **<topic>** sub-properties. LADI uses MODS as a standard for all of their collection material; our <note> property acts as a new, standalone field, while our <topic> sub-property functions as a sub-property of the more general <subject> entity.

LADI stores their metadata using Extensible Markup Language (XML), so here is an excerpt from a LADI metadata entry for a specific object (image) showing how the values generated by this project would be structured/tagged in XML:

NOTE (property used to store OCR transcript output for a given object):

```
<note type="content" displayLabel="Transcription" lang="spa">
INSERT OCR TRANSCRIPTION</note>
```

TOPIC (sub-property used to store NER value output for a given object):

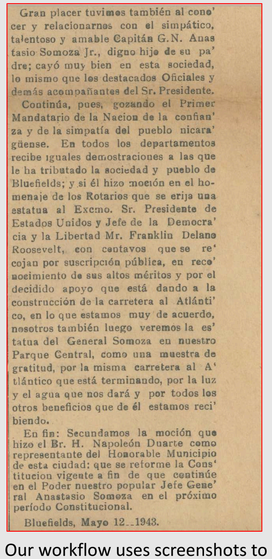
```
<subject lang="spa">
<topic>INSERT NAME ENTITY VALUE</topic>
```

Implementation/Workflow

"1943-05-13.tiff"
Ex. Original 'Composite' Image

To optimize output quality, newspaper articles must be manually **preprocessed** (segmented into individual chunks of text) before Google's Vision and Language APIs can process them.

"1943-05-13-1.png"
Ex. Segmented Image:



Our workflow uses screenshots to generate .png images of individual sections and columns of text. Collections of segmented files are stored in the same folder path according to issue date.

"1943-05-13-ocr.txt"
Ex. OCR Output:

"Gran placer tuvimos también al cono'|cer y relacionarnos con el simpático, talentoso y amable Capitán G. N. Anastasio Somoza Jr., digno hijo de su padre; cayó muy bien en esta sociedad, lo mismo que los destacados Oficiales y de más acompañante del Sr. Presidente,Continúa, pues, gozando el Primer Mandatario de la Nacion de la confianza y de la simpatía del pueblo nicaragüense. En todos los departamentos recibe iguales demostraciones a las que le ha tributado la sociedad y pueblo de Bluefields; y si él hizo noción en el homenaje de los Rotarios que se erija un aestatua al Excmo. Sr. Presidente de Estados Unidos y Jefe de la Democracia y la Libertad Mr. Franklin Delano Roosevelt, con centavos que se recojan por suscripción pública, en reco'nocimiento de sus altos méritos y por el decidido apoyo que está dando a la construcción de la carretera al Atlántico, en lo que estamos muy de acuerdo, nosotros también luego veremos la estatua del General Somoza en nuestro Parque Central, como una muestra degratitud, por la misma carretera al Atlántico que está terminando, por la luz y el agua que nos dará y por todos los otros beneficios que de él estamos recibiendo. En fin: Secundamos la moción que hizo el Br. H. Napoleón Duarte como representante del Honorable Municipio de esta ciudad: que se reforme la Constitucion vigente a fin de que continúe en el Poder nuestro popular Jefe General Anastasio Somoza en el próximo período Constitucional. Bluefields, Mayo 12.-1943."

Our **ladi.ocr.py** script feeds the group of files in a give folder path to Google's Vision API, prints the OCR transcript for each file, and stores the aggregated output in a new .txt file

"1943-05-13-nlp.txt"
Ex. Name Entity Output:

Presidente, sociedad, hijo, pueblos, Rotarios Bluefields, Franklin Delano Roosevelt, General Somoza, Oficiales, Anastasio Somoza, Jr., Nación, Primer Mandatorio, Estados Unidos, Honorable Municipio, departamentos, Censo, Br. H., Anastasio Somoza, Atlántico, Parque Central, Napoleón Duarte

Our **ladi.nlp.py** script feeds the previously-generated .txt OCR output to Google's Language API, processes this text, and generates a list of name entities it identifies as either "PERSON," "ORGANIZATION," or "LOCATION." It then stores these values in a new .txt file. All file titles link them back to the original image the values were generated off of.

NB: Entities are returned in the order (highest to lowest) of their salience scores, which reflects their relevance to the overall text.

Final Metadata File

File name: "1943-05-13-1.xml"

As a final step, our **ladi.mods.py** script takes the values stored in the OCR and NLP text files generated for a given image and populates them into a pre-structured XML file. This new file can be ingested by LADI staff and added to the existing metadata hosted by their online platform. Users will be able to search and filter based on these topics and notes just as they currently do for other metadata fields.

Ex. Of Transcript Entry:

<note type="content" displayLabel="Transcription" lang="spa">
Gran placer tuvimos también al conocer y relacionarnos con el simpático, talentoso y amable Capitán G. N. Anastasio Somoza Jr., digno hijo de su padre; cayó muy bien en esta sociedad, lo mismo que los destacados Oficiales y de más acompañantes del Sr. Presidente… **</note>.**

Ex. Of Topic Entries for Given Transcript:

<subject lang="spa">
<topic> Presidente **</topic>.**

<subject lang="spa">
<topic> sociedad **</topic>.**

<subject lang="spa">
<topic> hijo **</topic>.**