Título: *Trayectos V* (fragmento), de Héctor Miguel Guerrero Aburto

Tecnología OCR para la transcripción de registros de testimonios orales de la leyenda de Martín Toscano

Ignacio Moreno Nava
Universidad de La Ciénega del Estado de Michoacán de Ocampo

Resumen

El objetivo principal de esta investigación, fue digitalizar registros de testimonios orales que contuvieran datos de la leyenda de Martín Toscano provenientes del Archivo de Historia Oral (AHO) de la Unidad Académica de Estudios Regionales de la Coordinación de Humanidades de la Universidad Nacional Autónoma de México (UAER-COHU-UNAM) y aplicar tecnología de reconocimiento óptico de caracteres (OCR, por sus siglas en inglés) para generar un corpus textual de contenidos susceptible de ser explorado mediante búsquedas con cadenas de caracteres y herramientas de las humanidades digitales. Se localizaron los registros de interés a través de los catálogos del AHO, solicitando permiso para el registro fotográfico de las fojas de documentos. Mediante la herramienta *online JPGtoPDF* se integraron las imágenes seleccionadas en archivos PDF (Portable Document Format del software Adobe Acrobat) para su posterior procesamiento con la herramienta *OnlineOCR* para generar contenido transcrito. Entre los resultados obtenidos se digitalizaron y transcribieron 25 registros de testimonios orales que contenían datos de la leyenda de Martín Toscano. El corpus textual generado consta de 50 cuartillas, 11 144 palabras, 49 158 caracte-

Interpretextos
24/Otoño de 2020, pp. 209-228

res sin espacios, 580 párrafos y 1 235 líneas. Se reflexionó sobre la importancia de la literatura oral y las posibilidades integrativas con las humanidades digitales para su investigación y análisis. La misión principal del AHO es preservar memorias y recuerdos, la utilización de herramientas informáticas ofrece alternativas innovadoras para acceder a esta información.

Palabras clave

OCR, reconocimiento óptico de caracteres (OCR), literatura oral, humanidades digitales, Martín Toscano.

Título: *ha5* (fragmento), de Héctor Miguel Guerrero Aburto

Use of OCR Technology for the Transcription of Oral Testimony Records of Martín Toscano's Legend

Abstract

The main objective of this research was to digitize records of oral testimonies that contained data on the legend of Martín Toscano from the Oral History Archive (AHO) of the UAER CoHU UNAM and apply optical character recognition (OCR) technology to generate a textual corpus of content that can be explored through searches with character strings and digital humanities tools. Records of interest were located through the AHO catalogs and requesting permission for photograph the document sheets. Through the online tool JPGtoPDF, the selected images were integrated into PDF files for further processing with the OnlineOCR tool to generate transcribed content. Among the results obtained, 25 records of oral testimonies containing data from the legend of Martín Toscano were digitized and transcribed. The generated textual corpus consists of 50 pages, 11 144 words, 49 158 characters without spacing, 580 paragraphs, and 1 235 lines. Reflections on the importance of oral literature and the integrative possibilities with digital humanities for its research and analysis were made. The main goal of the AHO is to preserve memories. The use of computer tools offers innovative alternatives to access this information.

Interpretextos
24/Otoño de 2020, pp. 209-228

Keywords
OCR, optical character recognition, oral literature, digital humanities, Martín Toscano.

Introducción

La versión más popularizada de la leyenda de Martín Toscano lo menciona como un bandolero que dejó escondidos fabulosos tesoros encantados. Existen documentos escritos, conocidos como *relaciones*, para encontrar su ubicación exacta en muchos puntos de la geografía, comprendida desde el occidente del estado de Michoacán, pasando por Jalisco y hasta Colima.

Toscano fue un preinsurgente, capitán de gavillas y, a ojos de los españoles, un ladrón de lo peor; por el contrario, en opinión de muchos otros fue un valiente que encarnó el sentir de un pueblo cansado de vejaciones y abusos, tomando las armas y haciendo lo necesario para mantener un movimiento de resistencia que antecedió y sembró la semilla de la independencia. Parte de estos lejanos recuerdos quedaron presentes en la memoria de los habitantes de Jiquilpan.

El municipio de Jiquilpan forma parte de la región Ciénega de Chapala del Estado de Michoacán y colinda con los municipios de Sahuayo, Cojumatlan, Villamar, Marcos Castellanos, Cotija (en Michoacán) y con los municipios Quitupan y Valle de Juárez (en Jalisco). En el ámbito educativo, la región se caracteriza por tener presencia de múltiples universidades en un espacio geográfico relativamente pequeño. Tienen presencia la Universidad Autónoma de México, el Instituto Politécnico Nacional, el Tecnológico Nacional de México, la Universidad de La Ciénega del Estado de Michoacán de Ocampo, la Universidad de Formación Superior, la Universidad Interamericana para el Desarrollo, la Universidad Pedagógica Nacional y el Instituto Michoacano de Ciencias de la Educación.

En diciembre de 2005, a través de un comodato por 99 años, la Unidad Académica de Estudios Regionales de la Coordinación de Humanidades de la UNAM (UAER-COHU-UNAM), sede la Ciénega en Jiquilpan, Michoacán, recibió del Centro de Estudios de la Revolución Mexicana "Lázaro Cárdenas" A.C. (CERMLC), sus acervos docu-

mentales y bibliográficos, el museo centrado en la figura del general Lázaro Cárdenas y su infraestructura en general (UNAM, 2012).

El objetivo general de esta investigación fue digitalizar registros de testimonios orales que contuvieran datos de la leyenda de Martín Toscano, provenientes del Archivo de Historia Oral (AHO) de la ya mencionada UAER-COHU-UNAM, y aplicar tecnología de reconocimiento óptico de caracteres (OCR) para generar un corpus textual de contenidos que pueda ser explorado mediante búsquedas con cadenas de caracteres y herramientas de las humanidades digitales.

Es pertinente mencionar que el AHO es una relevante fuente de información para la generación de distintas investigaciones académicas en la región y fuera de ella, contiene gran cantidad de registros de literatura oral de habitantes de Jiquilpan y de la región Ciénega de Chapala; sin embargo, las consultas al archivo y sus contenidos se siguen realizando de manera poco eficiente y con prácticas de la época pre-computacional.

El uso de herramientas de las humanidades digitales puede contribuir a innovar en la forma de buscar esta información. La conciencia de un pasado histórico es necesaria en todas las comunidades, siendo este un factor determinante para entender quiénes somos y por qué nos diferenciamos de otras sociedades.

El proyecto de la generación del AHO nace en la década de 1980, a cargo del entonces coordinador Salvador Rueda Smithers y los investigadores Guadalupe García Torres, María de los Ángeles Manzano, Guillermo Ramos Arispe y Griselda Villegas; siendo su objetivo fundamental el rescatar la información histórica contenida en la memoria oral de la comunidad jiquilpense. Los informantes de los registros del archivo fueron personas de la tercera edad, quienes oscilaban en un rango de nacimiento que iba de 1888 a 1920. Se realizaron grabaciones de todas las entrevistas a profundidad y, posteriormente, se transcribieron en máquina de escribir. El acervo cuenta con alrededor de 300 entrevistas que describen los siguientes aspectos:

- Datos de los informantes pertenecientes a diferentes jerarquías sociales, desde donde se aprecian las diferentes subjetividades históricas.

- Formas de propiedad y unidades de producción (familiares y generales de la localidad, tanto de zonas rurales como del pueblo).
- Vida cotidiana. Los ritmos y los espacios del trabajo y del ocio. Los cambios y las permanencias. La modernidad como síntoma del paso del tiempo.
- Formación escolar, tendencias educativas y tipos de escuelas.
- El trabajo en la localidad, el hogar, el campo, el taller, la hacienda, la fábrica.
- El comercio local y exterior. La arriería.
- Vida política.
- Vida social.
- Instituciones de justicia y vigilancia.

Entre los hechos históricos que enmarcan y ubican la información de las entrevistas destacan:

- Los últimos años del porfiriato
- La Revolución de 1910-1920
- El levantamiento cristero
- El general Lázaro Cárdenas.

El contenido de estos registros es literatura oral; es decir, la literatura en su primigenia forma se usa especialmente para trasmitir las tradiciones y el folclore, pasa de boca en boca a lo largo de las generaciones. Es el primero y más extendido modo de comunicación humana, y comprende mitos, cuentos populares, leyendas, canciones y otros (Castillo, 2018). El concepto de oralidad está construido desde la cultura de la escritura, por ello al referirlo es necesario situarse en este ámbito (Dorra, 1997).

La tecnología OCR

Las nuevas herramientas tecnológicas digitales y sus posibilidades de utilización en el ámbito académico, especialmente las humanidades, toma impulso en nuestros días, tal es el caso de la implementación de la tecnología OCR aplicada en archivos históricos, misma que per-

mite convertir cualquier archivo digitalizado en cadenas de texto rastreables; esto es, el software transforma el texto digitalizado del documento en un formato comprensible para el ordenador, lo cual nos permite obtener un archivo de texto rastreable y editable.

Las humanidades digitales, en uno de sus múltiples sentidos, buscan integrar herramientas, procesos y metodologías del campo tecnológico para innovar y desarrollar nuevas prácticas en el campo de las humanidades, y constituyen un campo emergente y transdisciplinar. El texto es, sin lugar a dudas, uno de los conceptos más trascendentes para el conjunto de las humanidades y su relación con la nueva ecología digital, es espacio de reflexión académica que genera debates sobre el uso de métodos y sistemas computacionales (Priani, 2015).

Materiales y métodos

Para este proyecto se diseñó una estrategia práctica para la digitalización y aplicación de tecnología OCR a los registros del AHO que contenían información sobre la leyenda de Martín Toscano. Para ello, en primer lugar se dio lectura a los catálogos del AHO, ubicando de manera manual las claves correspondientes a las entrevistas que mencionaron la leyenda; en seguida, se procedió a solicitar el acceso al archivo para la consulta de la documentación seleccionada, junto con el permiso correspondiente para el registro fotográfico de las fojas de documentos, lo cual se realizó con una cámara fotográfica digital Canon T5i.

Las tomas de los documentos se hicieron en un ángulo cenital, cuidando de no generar sombras sobre los documentos; posteriormente, con los materiales ya digitalizados, se inició un proceso de normalización de imágenes, realizando cortes a aquellas que excedían los márgenes de los documentos fotografiados e igualando sus características de contraste e iluminación para uniformizar las características del conjunto de imágenes.

Mediante la herramienta *online JPGtoPDF* se integraron las imágenes seleccionadas en un archivo con formato PDF. El servicio convierte las imágenes JPG a formato PDF, las rota automáticamente y optimiza y reduce proporcionalmente las imágenes para mantener la resolución original (imagen 1).

Interpretextos
24/Otoño de 2020, pp. 209-228

Imagen 1

Integración de imágenes en un archivo PDF con la herramienta online JPGtoPDF

Fuente: Elaboración propia.

Para lo anterior, se accedió al sitio web https://jpg2pdf.com/es/, seleccionando la opción de *Subir archivo*, es permitido seleccionar un máximo de 20 imágenes a convertir y esperar a que los procesos de carga y de conversión se completen. Posteriormente se seleccionó la función de *Combinado* para descargar todos los archivos combinados en un solo documento PDF (imagen 2) (Media4x, 2020).

La técnica de OCR fue aplicada haciendo uso de los servicios de una herramienta en línea denominada *OnlineOCR.net*, la cual es una web gratuita que permite convertir documentos PDF escaneados (incluyendo archivos de múltiples páginas), faxes, fotografías o imágenes capturadas de una cámara digital, en documentos electrónicos editables y de búsqueda, incluyendo formatos PDF, Docx, Xlsx, RTF, Html y Txt.

La calidad de la imagen es uno de los factores más importantes para mejorar el reconocimiento: una resolución de 200 a 400 DPI para imágenes de entrada es ideal. El tamaño máximo de archivo de entrada es de 200 MB (OnlineOCR, 2020). Al registrarse en la página se accede a opciones extra.

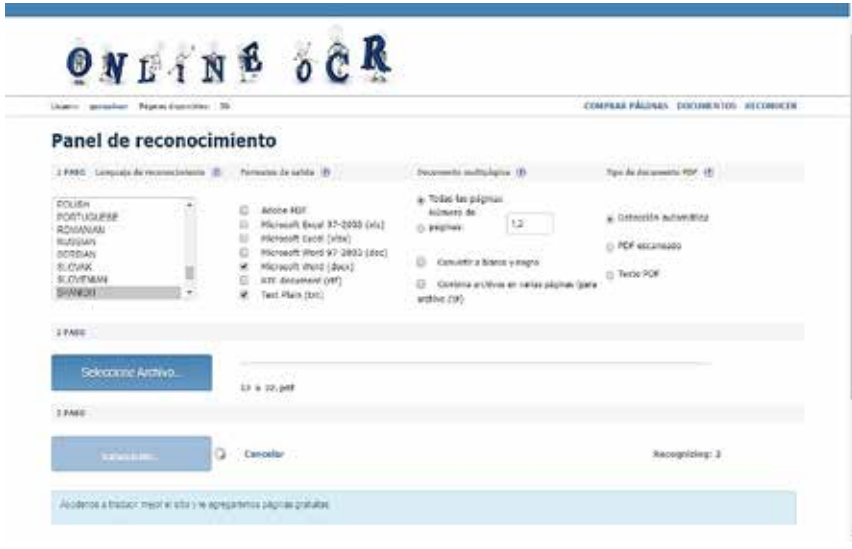
Imagen 2

Interfaz de carga de archivos de OnlineOCR.net

Fuente: Elaboración propia.

Imagen 3

Panel de reconocimiento en cuenta de usuario registrado en el sitio OnlineOCR.net



Fuente: Elaboración propia.

Finalmente, se procedió a descargar los resultados en formato Docx y TXT, seleccionando los textos deseados y trasladándolos a un procesador de texto mediante los comandos de copiado y pegado, para la conformación de contenidos.

El proceso de conversión de imagen a texto no fue totalmente preciso, por lo que un porcentaje de los textos procesados presentaba errores de codificación de caracteres; por tal motivo se procedió a ingresar de manera manual las secciones que presentaron esta característica. El porcentaje de precisión osciló entre 90 y 80%, siendo de gran ayuda para la transcripción de las secciones de los registros seleccionados.

Resultados

A partir del proceso realizado, se digitalizaron y transcribieron 25 registros de testimonios orales que contenían datos de la leyenda de Martín Toscano, provenientes del AHO de la UAER-COHU-UNAM y se aplicó tecnología OCR.

A continuación, y a modo de ejemplo, se muestra un fragmento del registro de la entrevista correspondiente a Petra Méndez Abad (AHOCLC-Z1-E: 95/153 pp.). Tal como puede observarse en la imagen 4, el archivo cuenta con la característica de tener una buena calidad y adecuada definición.

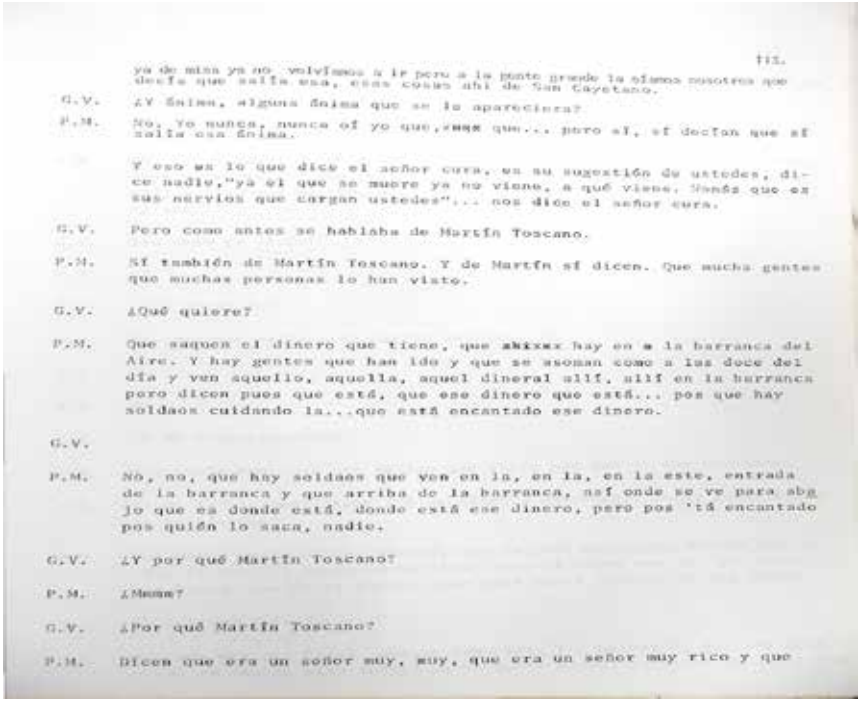
Debido a que la tecnología OCR está basado en la emulación de la visión humana, es también susceptible a no poder distinguir de manera adecuada los caracteres que se encuentren en una imagen, lo cual es primordial para conseguir buenos resultados y debe ser un factor especialmente controlado para reproducir la técnica.

Aplicando la tecnología de OCR y descargando el archivo generado en formato Docx, se obtuvo el siguiente resultado: se dejaron las marcas de corrección ortográfica y gramatical que señala el navegador utilizado para la visualización; el documento generado respetó las características de organización de texto de la fuente original; el reconocimiento de caracteres mediante tecnología OCR fue sumamente preciso; el documento de entrada incluía un conjunto de caracteres sobrepuestos y un carácter aislado también sobrepuesto, a manera de tachadura, en la quinta línea del texto, la tecnología OCR trató de interpretar estas anomalías convirtiéndolas

a caracteres legibles y produciendo, en el primer caso, una palabra fuera de contexto y, en el segundo caso, el caracter correspondiente a la letra "a" (imagen 5).

Imagen 4

Fragmento del registro de la entrevista correspondiente a Méndez Abad, Petra (AHOCLC-Z1-E: 95/153 pp.)



ya de misa ya no volvíamos a ir pero a la noche grande la oíamos nosotros que decía que salía esa, esas cosas ahí de San Cayetano.

G.V. ¿Y ánima, alguna ánima que se le apareciera?

P.M. No, Yo nunca, nunca oí yo que,xxxx que... pero sí, sí decían que sí salía esa ánima.

Y eso es lo que dice el señor cura, es su sugestión de ustedes, dice nadie,"ya el que se muere ya no viene, a qué viene. Vendá que es sus nervios que cargan ustedes"... nos dice el señor cura.

G.V. Pero como antes se hablaba de Martín Toscano.

P.M. Sí también de Martín Toscano. Y de Martín sí dicen. Que muchas gentes que muchas personas lo han visto.

G.V. ¿Qué quiere?

P.M. Que saquen el dinero que tiene, que xxxxx hay en a la barranca del Aire. Y hay gentes que han ido y que se asoman como a las doce del día y ven aquello, aquella, aquel dineral allí, allí en la barranca pero dicen pues que está, que ese dinero que está... pos que hay soldaos cuidando la...que está encantado ese dinero.

G.V.

P.M. No, no, que hay soldaos que ven en la, en la, en la este, entrada de la barranca y que arriba de la barranca, así onde se ve para abajo que es donde está, donde está ese dinero, pero pos 'tá encantado pos quién lo saca, nadie.

G.V. ¿Y por qué Martín Toscano?

P.M. ¿Mmmm?

G.V. ¿Por qué Martín Toscano?

P.M. Dicen que era un señor muy, muy, que era un señor muy rico y que

Fuente: Elaboración propia.

Interpretextos
24/Otoño de 2020, pp. 209-228

Imagen 5

Resultado de la conversión con tecnología OCR a formato Docx con la herramienta OnlineOCR.net

G.V. Pero como antes se hablaba de Martín Toscano.

P.M. Sí también de Martín Toscano. Y de Martín sí dicen. Que mucha gentes que muchas personas lo han visto.

G.V. ¿Qué quiere?

P.M. Que saquen el dinero que tiene, que Akimax hay en a la barranca del Aire. Y hay gentes que han ido y que se asoman como a las doce del día y ven aquello, aquella, aquel dineral allí, allí en la barranca pero dicen pues que está, que ese dinero que está... pos que hay soldaos cuidando la. —que está encantado ese dinero.

G. V.

P.M. No, no, que hay soldaos que ven en la, en la, en la este, entrada de la barranca y que arriba de la barranca, así onde se ve para abajo que es donde está, donde está ese dinero, pero pos 'tá encantado pos quién lo saca, nadie.

G.V. ¿Y por qué Martín Toscano?

P.M. ¿Mmmm?

G.V. ¿Por qué Martín Toscano?

Fuente: Elaboración propia.

En la imagen 6 se muestra el resultado generado en la caja de texto de la herramienta *online* al momento simultáneo de la conversión. Se han dejado a propósito las marcas de corrección ortográfica que señala el navegador utilizado para la visualización.

La generación de contenido directamente en la caja de texto resulta de gran utilidad para brindar rapidez al flujo de datos de conversión de caracteres cuando se trabaja por pequeños fragmentos de texto o cuando no se requiere preservar la estructura de origen del texto original, ya que su accesibilidad es inmediata y no requiere de la descarga de un archivo, ya que es posible manipularla mediante comandos simples de copiado o cortado.

Imagen 6

Resultado de conversión OCR
generado en la caja de texto de la herramienta


G.V. Pero como antes se hablaba de Martín Toscano.
P.M. Sí también de Martín Toscano. Y de Martín sí dicen. Que mucha gentes que muchas personas lo han visto.
G.V. ¿Qué quiere?
P.M. Que saquen el dinero que tiene, que Akimax hay en a la barranca del Aire. Y hay gentes que han ido y que se asoman como a las doce del día y ven aquello, aquella, aquel dineral allí, allí en la barranca pero dicen pues que está, que ese dinero que está... pos que hay soldaos cuidando la. —que está encantado ese dinero.
G. V.
P.M. No, no, que hay soldaos que ven en la, en la, en la este, entrada de la barranca y que arriba de la barranca, así onde se ve para aba jo que es donde está, donde está ese dinero, pero pos tá encantado pos quién lo saca, nadie.
G.V. ¿Y por qué Martín Toscano? P.M. ¿Mmmm? G.V. ¿Por qué Martín Toscano?

Fuente: Elaboración propia.

A continuación se muestra el resultado de la transcripción, con un criterio ecdótico, centrado en el sentido del texto y expresión oral, con la eliminación de redundancias para facilitar una legibilidad fluida.

Méndez Abad, Petra

(AHOCLC-Z1-E: 95/153 pp.)
Entrevistas con la señora Petra Méndez Abad, realizadas por Griselda Villegas M. en tres sesiones: los días 30 de enero y 11 y 13 de febrero de 1984 en Jiquilpan, Michoacán.
Petra nació en 1900 y originaria de Jiquilpan, Michoacán.
La leyenda de Martín Toscano (pp. 113-114).

G.V. Pero cómo antes se hablaba de Martín Toscano
P.M. Sí de Martín Toscano, se les apareció a muchas personas.
G.V. ¿Qué quiere?
P.M. Que saquen el dinero de la barranca del Aire. Muchas personas han ido como a las 12 del día y ven el tesoro allí en la barranca, pero cuentan que sus soldados lo están cuidando porque está encantado.
G.V. Y ¿Dónde están los soldados?
P.M. Están a la entrada de la barranca. En la cima viendo hacia abajo es donde está el dinero, pero como le digo, lo dejó encantado ¿Así quien lo saca? Pues nadie.
G.V. ¿Y por qué Martín Toscano?
P.M. Era un señor de mucho dinero y cuando lo ven está montado en su caballo.
G.V. ¿Aparece montado en un caballo?

Interpretextos
24/Otoño de 2020, pp. 209-228

P.M. Sí, va vestido con un sombrero de charro y con un pantalón de botonadura de plata y siempre sale del mismo lugar así es como me han contado.
G.V. ¿Su papá decía algo?
P.M. Sí.
G.V. ¿Qué decía él?
P.M. Que salía Martín Toscano.
P.M. Sí.
G.V. ¿Y de otro aparecido?
P.M. No, solo de él, decían que salía por San Cayetano.

El corpus textual generado consta de 50 cuartillas, 11 144 palabras, 49 158 caracteres sin espacios, 580 párrafos y 1 235 líneas. Se nutrió con los contenidos de los siguientes registros de entrevistas del AHO:

1. Díaz Madrigal, Nicolás (AHOCLC-Z1-E46/94 pp.)
2. Hernández Pulido, Francisco (AHOCLC-Z1-E27/50 pp.)
3. Herrera Macías, Melitón (AHOCLC-Z1-E1/128 pp.)
4. Aguilera Flores, Sabas (AHOCLC-Z1-E: 93/65 pp.)
5. Aguilera Ordaz, Irene (AHOCLC-Z1-E: 77/164 pp.)
6. Álvarez Martínez, Lilia (AHOCLC-Z1-E: 84/66 pp.)
7. Álvarez Santillán, Ma. De Jesús (AHOCLC-Z1-E: 87/80 pp.)
8. Macías Sánchez, Ma. Elena (AHOCLC-Z1-E: 88/33 pp.)
9. Méndez Abad, Petra (AHOCLC-Z1-E: 95/153 pp.)
10. Olloqui Rosas Juan (AHOCLC-Z1-E: 75/176 pp.)
11. Orozco Espinoza, María (AHOCLC-Z1-E: 103/102 pp.)
12. Orozco Espinoza, Ma. Trinidad (AHOCLC-Z1-E: 92/120 pp.)
13. Rodríguez Martínez, María (AHOCLC-Z1-E: 101/118 pp.)
14. Valdovinos Zapien, Fidelia (AHOCLC-Z1-E: 99/73 pp.)
15. Valencia Muratalla, Julia (AHOCLC-Z1-E: 94/81 pp.)
16. Vázquez Rocha, Catalina (AHOCLC-Z1-E: 107/51 pp.)
17. Gálvez Arceo, Vicente (AHOCLC-Z1-E: 139/62 pp.)
18. Gómez Martínez, Manuel (AHOCLC-Z1-E: 142/138 pp.)
19. González Cisneros, José (AHOCLC-Z1-E: 11/95 pp.)
20. Maciel Zuno, Ramona (AHOCLC-Z1-E: 145/55 pp.)
21. Martínez Magaña, Ma. De Jesús (AHOCLC-Z1-E: 153/167 pp.)
22. Mejía García, Juan (AHOCLC-Z1-E: 164/72 pp.)
23. Torres Martínez, Maclovia (AHOCLC-Z1-E: 143/70 pp.)
24. García Gálvez, Rafael (AHOCLC-Z1-E: 179/118 pp.)
25. Rodríguez Núñez, Carmen (AHOCLC-Z2-E: 184/49 pp).

Discusión

La aplicación de tecnología a procesos de investigación de manifestaciones del patrimonio cultural inmaterial, como es en este caso la oralidad, es una línea esencial del trabajo que se realiza desde el Laboratorio de Gestión Cultural y Humanidades Digitales de la Universidad de La Ciénega del Estado de Michoacán de Ocampo.

En primer término, al contar con transcripciones de los registros de testimonios orales se pueden realizar búsquedas de cadenas de caracteres (palabras o términos clave) dentro de sus contenidos, facilitando la ubicación de información en el corpus textual generado. Por otra parte, los registros de testimonios orales con los que cuenta el AHO son una fuente extensa e importante para el estudio de la memoria de la colectividad y constituyen una ventana a la literatura oral local y regional.

> La literatura oral está conformada por discursos que tienen como soporte la voz, la expresión corporal y la memoria. Su ejecución, sucede en momentos y espacios consensuados de forma única e irrepetible, determinada por su contexto de producción. Esta literatura cumple, además, una serie de funciones sociales: sanar, festejar, recordar, entretener, enseñar, reforzar la identidad de una comunidad, etcétera. Se genera en actos comunicativos en los que siempre están presentes un emisor y un receptor, aunque a veces sólo de forma simbólica, como en el caso de las oraciones, los conjuros las maldiciones. Transmite de forma eficaz el sistema de conocimientos, valores, normas y creencias compartidos por una colectividad y sirve, como acción, para configurar el mundo que habita (Granados y Cortés, 2018: s.p.).

En tercer lugar, la reflexión sobre los aspectos técnicos de esta investigación conduce a una mirada retrospectiva. La investigación científica se encuentra siempre sujeta a un paradigma, como un conjunto de conceptos y prácticas que definen una comunidad científica en un momento histórico concreto (Kuhn, 1962). El primer paradigma científico está basado en la percepción sensorial y la explicación del mundo desde una visión mágico-religiosa; el segundo, implicó la elaboración de teorías, abandonando el pensamiento causal simple y propiciando el surgimiento de la ciencia teórica; el

tercer paradigma fue el de la ciencia computacional, donde con el uso de la tecnología fue posible acceder a niveles de cálculo y detalles sin precedentes (Vallverdu, 2018).

Diversos teóricos afirman que el cuarto paradigma ha llegado (Gahegan, 2020; Hey y Trefethen, 2020; Granshaw, 2019; Ceri, 2018). Después del auge de las computadoras, la tendencia a la disminución de tamaño de sus componentes y sus capacidades de interconexión en múltiples ámbitos de nuestra vida, uno de los resultados ha sido el aumento exponencial de datos, lo que ha requerido nuevas aproximaciones a la noción de conocimiento y a los métodos con los que se gestiona el entorno (Vallverdu, 2018). Conocer y gestionar la información disponible en archivos, bibliotecas, repositorios digitales y bases de datos es sumamente importante para la investigación.

La tecnología que se utilizó para el proceso presentado en este texto se encuentra entre los límites del tercer y cuarto paradigmas, vinculándolo con la textualidad digital, la cual es uno de los problemas en que se ha centrado el debate dentro de las humanidades digitales (Eggert, 2005; Caton, 2013; Buzzetti y Thaller, 2012). Desde el campo de las humanidades digitales, se cuenta con múltiples recursos y herramientas digitales pensadas específicamente para el estudio filológico. Eggert (2005), respecto a la producción de ediciones académicas electrónicas, reflexiona sobre la naturaleza del texto y explora las implicaciones para la codificación del texto.

De acuerdo con Boadas (2018), es evidente que los nuevos enfoques y el análisis de textos con instrumentos digitales ofrecen posibilidades de estudio que eran inimaginables hace algunos años. En las humanidades digitales, dentro de un ámbito semántico central, el término *texto* aparece de manera ubicua, tanto en sentido de masa como de sustantivo contable (Caton, 2013).

A partir del corpus textual generado, es posible la utilización de técnicas y herramientas de análisis del texto. Rockwell y Sinclair (2016) propusieron en *Hermeneutica: Computer-assisted interpretation in the humanities*, varias posibilidades a través del análisis de texto asistido por computadora y usando herramientas de libre acceso. Las prácticas académicas también están cambiando sus formas antiguas de investigación al combinarse con métodos modernos habilitados por Internet, informática accesible, disponibili-

dad de datos y nuevos medios. Gracias a la unión del conocimiento humanístico con la tecnología y las nuevas herramientas computacionales, podemos cambiar la manera de estudiar e investigar las humanidades proporcionando nuevas lecturas e interpretaciones. Mirar al pasado, desde el nuevo presente, preparándonos para el futuro (Fornes, 2018).

Para complementar sobre la leyenda de Martín Toscano, es preciso señalar que esta forma parte del universo simbólico de los habitantes de la región Ciénega de Chapala, tanto en Jalisco como en Michoacán. La existencia de Martín Toscano está documentada y nos remite al siglo XVIII. Nacido en el año de 1754 en Atoyac, Jalisco, y ejecutado el día 12 de enero de 1803 en Guadalajara, Jalisco. Sus correrías quedaron impresas en la memoria de los pobladores de las regiones que acostumbraba recorrer, tan fue así que la novela decimonónica *La hija del bandido*, escrita por Refugio Barragán de Toscano, tomó a Martín Toscano como base para crear el personaje ficticio de Vicente Colombo (Sedano y Moreno, 2019). También en muchas ocasiones al recuerdo de Toscano le confieren un toque de revolucionario, sin dejar de lado el aspecto sobrenatural.

Tal como lo mencionan Granados y Cortés (2018), en las tradiciones orales existen personajes que aparecen de manera recurrente y que forman parte del sistema de pensamiento y códigos comunes de un grupo social, de los cuales —en ocasiones— es posible rastrear literaria e históricamente a estos personajes, además de encontrar datos sobre su caracterización, sus implicaciones simbólicas y su distribución geográfica, así como los contextos de producción y ejemplos de sus diversas apariciones. En el caso específico del personaje de Martín Toscano, tales condiciones se cumplen, resultando los testimonios registrados en fuentes invaluables para su análisis hermenéutico y profundización investigativa.

Conclusiones

El objetivo planteado de digitalizar registros de testimonios orales que contuvieran datos de la leyenda de Martín Toscano, provenientes del AHO de la UAER-COHU-UNAM, aplicando tecnología OCR para generar un corpus textual que permitiera ser explorado con cadenas de caracteres y herramientas de las humanidades digitales,

se cumplió a cabalidad y abre un amplio panorama de posibilidades para ampliar la investigación y proponer la aplicación de la metodología empleada a mayor escala.

La transformación y la hibridación de técnicas lleva a la innovación disruptiva, en algunos casos la implementación tecnológica genera incertidumbre. La consulta de información de archivos no va a desaparecer, lo que cambia es la manera de realizarlas; en este caso, al modo de acercarse a esos contenidos y la forma de obtener información de los mismos.

En un mundo híbrido es importante repensar la misión de las instituciones u organizaciones. La misión principal del AHO es preservar memorias y recuerdos, así como ofrecer formas para acceder a esta información. La adaptabilidad es una habilidad sumamente necesaria en el paradigma actual de acceso a la información y puede resumirse en hacerlo exitosamente o aprender de la experiencia.

En términos de la interpretación de los contenidos, si bien se ha abordado un proceso tecnológico para la generación del *corpus*, la intervención humana sigue siendo fundamental para identificar el sentido de los contenidos, el uso de la hermenéutica interpretativa en conjunto con técnicas informáticas resulta de esencial valor para conocer el significado de los conjuntos de caracteres que, en términos técnicos, han sido creados con el procedimiento mostrado.

Referencias bibliográficas

Boadas, S. y Vallverdu, J. (2018). *La textualidad digital. MOOC Humanidades digitales*. Barcelona: Universitat Autònoma de Barcelona, Coursera Inc. Disponible en: https://www.coursera.org/lecture/humanidades-digitales/la-textualidad-digital-TCUxy (Consultado el 28 de marzo de 2020).

Buzzetti, D. y Thaller, M. (2012). *Beyond embedded markup*. Journal of Computing in Higher Education, 1(2), 3-2. Disponible en http://web.dfc.unibo.it/buzzetti/dbuzzetti/pubblicazioni/hamburg_dh2012_ed.pdf (Consultado el 15 de marzo de 2020).

Castillo, I. (2018). *Literatura oral: Origen e historia, características y ejemplos*. Lifeder. Disponible en: https://www.lifeder.com/literatura-oral/ (Consultado el 15 de marzo de 2020).

Caton, P. (2013). On the term text in digital humanities. *Literary and Linguistic Computing*, 28(2): 209-220. Disponible en: https://doi.org/10.1093/llc/fqt001

Ceri, S. (2018). On the role of statistics in the era of big data: A computer science perspective. *Statistics and Probability Letters*, 136(1): 68-72.

Dorra, R. (1997). *¿Grafocentrismo o fonocentrismo? Perspectivas para un estudio de la oralidad. Memorias. Jornadas Andinas de Literatura Latinoamericana*. Ed. R.J. Kaliman. Vol. I. Tucumán: Univ. Nacional de Tucumán. 56-73.

Eggert, P. (2005). Text-encoding, theories of the text, and the work-site. *Literary and Linguistic Computing*, 20(4): 425-435. Disponible en: https://doi.org/10.1093/llc/fqi050

Fornes, A. (2018). *Música digital. MOOC humanidades digitales*. Barcelona: Universitat Autònoma de Barcelona, Coursera Inc. Disponible en: https://www.coursera.org/lecture/humanidades-digitales/de-las-humanidades-a-las-humanidades-digitales-kWP9Z (Consultado el 27 de marzo de 2020).

Gahegan, M. (2020). Fourth paradigm GIScience? Prospects for automated discovery and explanation from data. *International Journal of Geographical Information Science*, 34(1): 1-21.

Granados, B. y Cortés, S. (2018). Literatura oral. Enciclopedia de la literatura en México. Secretaria de Cultura. Fundación para las Letras Mexicanas A.C. Laboratorio de Materiales Orales: ENES Morelia UNAM. Disponible en: http://www.elem.mx/literatura_oral (Consultado el 28 de marzo de 2020).

Granshaw, S.I. (2019). Open data, the fourth paradigm and Plan S. *The Photogrammetric Record*, 34(165): 6-10.

Hey, T. and Trefethen, A. (2020). The Fourth Paradigm 10 Years On. *Informatik-Spektrum*, 42(6): 441-447.

Kuhn, T.S. (1962). *The structure of scientific revolutions*. University of Chicago press. Original edition.

Media4x. (2020). *JPT to PDF*. Disponible en: https://jpg2pdf.com/es/ 15 de marzo de 2020.

OnlineOCR (2020). *OnlineOCR*. Disponible en: https://www.onlineocr.net/es/ (Consultado el 15 de marzo de 2020).

Universidad Nacional Autónoma de México (2012). *Antecedentes UAER*. México: UNAM - Departamento de Sistemas de la Coordinación de Humanidades. Disponible en: http://uaer.humanidades.unam.mx/uaer/creacion/ (Consultado el 15 de marzo de 2020).

Priani, E. (2015). El texto digital y la disyuntiva de las humanidades digitales. *Palabra Clave*, 18(4): 1215-1234. DOI: 10.5294/pacla.2015.18.4.11

Rockwell, G. y Sinclair, S. (2016). Hermenéutica: Computer-assisted interpretation in the humanities. EU: MIT Press.

Sedano, D. y Moreno, I. (2019). De bandidos y literatura. *La Gaceta del CUSur*, 12(138): 10. Disponible en: http://www.cusur.udg.mx/es/sites/default/files/gaceta138marzo2019.pdf (Consultado el 28 de marzo de 2020).

Vallverdu, J. (2018). *El cuarto paradigma. MOOC Humanidades digitales*. Barcelona: Universitat Autònoma de Barcelona - Coursera Inc. Disponible en: https://www.coursera.org/lecture/humanidades-digitales/el-cuarto-paradigma-tVo4t (Consultado el 25 de marzo de 2020).

Interpretextos
24/Otoño de 2020, pp. 209-228

Recepción: Marzo 2 de 2020.
Aceptación: Mayo 30 de 2020.

Ignacio Moreno Nava
Correo electrónico: imoreno@ucienegam.edu.mx

Nacionalidad: mexicano. Doctor en pensamiento complejo en la Multidiversidad Mundo Real "Edgar Morin". Profesor-investigador de tiempo completo de la licenciatura en estudios multiculturales de la Universidad de La Ciénega del Estado de Michoacán de Ocampo. Áreas de interés y líneas de investigación: gestión cultural, humanidades digitales, patrimonio cultural y natural, pensamiento complejo y transdisciplina.