IX CONFERÊNCIA INTERNACIONAL SOBRE BIBLIOTECAS E REPOSITÓRIOS DIGITAIS DA AMÉRICA LATINA
BIREDIAL-ISTEC 2019
SÃO PAULO · BRASIL 30·31 DE JULHO / 1·2 DE AGOSTO 2019

Los datos de investigación en las Humanidades – periódicos decimonónicos

Isabel Galina Russell[1], Miriam Peña Pimentel[2], Ernesto Priani Saisó[3]

[1] Instituto de Investigaciones Bibliográficas, Universidad Nacional Autónoma de México (UNAM). Correo: igalina@unam.mx
[2] Instituto de Investigaciones Bibliográficas, Universidad Nacional Autónoma de México (UNAM). Correo: miriampp@unam.mx
[3] Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México (UNAM). Correo: epriani@filos.unam.mx

Resumen

Introducción: El OCR de periódicos digitalizados del siglo XIX pueden entenderse como "datos crudos" para la realización de investigación histórica hemerográfica y ofrecemos una reflexión en torno a cómo deben integrarse estos datos al repositorio para poder ser utilizado para estos fines. Este trabajo forma parte de los resultados de “Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories”, un proyecto de minería de datos en repositorios de periódicos digitalizados. Materiales y metodología: OcEx está compuesto por 6 equipos de investigación en 9 países y busca modelar patrones de flujo de información en periódicos del XIX. México participa con el repositorio Hemeroteca Nacional Digital de México (HNDM) que contiene más de 7 millones de imágenes y sus correspondientes archivos en XML producto del proceso de OCR. A partir de utilizar este repositorio para realizar un proyecto de minería de datos humanístico hacemos una reflexión en cómo pueden ser concebidas las colecciones digitales y sus datos de tal forma que pueden ser utilizados como datos crudos para investigación hemerográfica. Resultados y conclusiones: El uso de repositorios para custodiar colecciones de datos crudos permite compartir y reutilizar esta información. Sin embargo, es necesario una descripción bibliográfica de la digitalización. Para lograr esto es necesario aproximarse a los datos resultantes de una digitalización, desde una perspectiva crítica y no únicamente tecnológica. Este trabajo representa un primer acercamiento y reflexión a cómo este tipo de datos deben de ser almacenados en repositorios para que sean adecuadamente interpretados y utilizados en la investigación.

Abstract

Introduction: OCR from digitised 19th century newspapers can be viewed as “raw data” for historical periodical researchers. In this talk we offer an approach on how this data can be integrated into a repository so that it can be useful for this type of work. This paper presents some of the results of “Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories (OcEx)”, an international data mining project focused on repositories of historical digitized newspapers. Materials and

methodology: OcEx is a project consisting of 6 research teams in 9 different countries that seeks to model information flow in 19th century newspapers using data mining techniques. Mexico participates with the national newspaper library, the Hemeroteca Nacional Digital de México (HNDM), with over 7 million images and their corresponding XML OCR files. Using the data mining work we did on the HNDM as a case study, we reflect on how these digital collections and the data can be designed in order to be used as raw data for periodical studies. Results and conclusions: The use of repositories for raw data collection allows for sharing and reuse. However, a bibliographical description of the digitization is necessary. In order to do this a critical and analytical and not just technological approach to the digitization process is necessary. This paper is an initial approximation and reflection on how historical newspaper collections can be stored in repositories in such a way that they can be interpreted and used adequately for humanistic research.

Introducción

Las publicaciones formales académicas, tales como artículos de revistas o capítulos de libro, han sido de interés primordial para muchos repositorios y el movimiento de acceso abierto. Los repositorios sin embargo, pueden albergar una vasta cantidad de materiales, tales como documentos fotográficos, manuscritos, planos arquitectónicos, y partituras musicales y bases de datos por mencionar algunos, ya que estos también contribuyen al ciclo de gestión, promoción y visibilidad del conocimiento. En la última década los repositorios tienden a convertirse en plataformas para los datos abiertos. Sin embargo, “los datos, los estándares de evidencia, las formas de representación y las prácticas de investigación están entrelazados. Las diferencias entre comunidades se vuelven aparentes sólo cuando se intenta hacer uso de ellos al combinarlos con datos externos para colaborar entre disciplinas” (Borgman 2015[1]).

Si bien se han estudiado los distintos procesos de comunicación científica entre las Humanidades y las Ciencias y el impacto que esto tiene en repositorios, existe menos trabajo en cómo los documentos digitales almacenados en repositorios pueden descomponerse en datos que son utilizados para realizar investigación en las Humanidades y por lo tanto, saber cómo se conforman esos datos y cómo se trabaja con ellos , para que ello sea tomado en consideración en la creación de repositorios (Gómez, Méndez, y Hernández-Pérez 2016).

Estamos viviendo un periodo en donde “los datos” se han convertido en un objeto de estudio de mucha importancia, en particular con la posibilidad del *big data* en donde podemos utilizar una gran cantidad de ellos para realizar análisis y proyecciones de nuevas formas (Borgman 2012). El énfasis en lo “abierto” y la amplia disponibilidad de datos, han generado un interés por la ciencia de datos. Sin embargo, podemos argumentar que los “datos” no son elementos aislados, independientes y objetivos sino que están constituidos a partir de decisiones e interpretaciones humanas y que estas deben contextualizarse para que otros hagan uso de ellos (Lynch 2002). Tanto en las

[1] Todas las traducciones son de los autores.

Ciencias como en las Humanidades, los metadatos acerca de los datos son claves y necesarios pues proveen contexto -en mínima expresión- sin la necesidad de desarrollar una narrativa (Fenner et al. 2019).

Para cada disciplina existen formas particulares de entender qué son datos válidos, su naturaleza y cómo deben ser recabados (Gómez, Méndez, y Hernández-Pérez 2016; Borgman, Wallis, y Mayernik 2012). Las disciplinas humanísticas generalmente no generan datos para investigación, sino que utilizan registros de actividad humana como fuente de estudios, tales como fotografías, cartas, periódicos, libros, artículos, archivos, registros civiles (Borgman 2009). Sin embargo, al digitalizarse estos objetos, tanto a partir de su contenido como de su descripción, pueden ser construidos conjuntos de datos que permiten hacer una aproximación distinta a estos, transformando el modo como pueden ser analizados tales registros. Para Christof Schöch (2013) “un dato en humanidades puede ser considerado como una abstracción digital, construida selectivamente, manipulable por una máquina, que representa algún aspecto de un objeto dado para el estudio humanístico”.

En este trabajo se plantea que el OCR (*Optical Character Recognition*) de periódicos digitalizados del siglo XIX pueden entenderse como "datos crudos" para la realización de investigación histórica hemerográfica y ofrecemos una reflexión en torno a cómo deben integrarse estos datos al repositorio para poder ser utilizado para estos fines. No pretendemos que se entienda el periódico histórico únicamente como datos, pues como se ha señalado antes los datos, dependiendo de la forma cómo se estructuran, sólo representan un aspecto de ese objeto, nunca su totalidad. Nuestra intención, en cambio, es realizar una reflexión en torno a cómo, al digitalizarse los periódicos y al generarse textos digitales a partir del OCR los investigadores pueden proponer diversas estructuras de datos -por ejemplo, entidades nombradas, palabras más frecuentes, fuente y destino, sólo por citar algunas- para ser procesados a través de una computadora. Esto implica una aproximación epistemológica diferente con respecto a los objetos de estudio, en este caso los periódicos del siglo XIX, pues utiliza representaciones parciales para ser conocidas a través de procesos computacionales. Por eso debemos analizar críticamente esta conversión para entonces contextualizar apropiadamente los resultados que obtenemos cuando realizamos desde cosas sencillas como búsquedas hasta temas más complejos como análisis lingüísticos. El término "datos crudos" puede entenderse como un oxímoron, ya que todos los datos “están cocinados" de alguna forma (Gitelman 2013). Cuando reducimos objetos físicos con cargas culturales e históricos a datos realizamos un proceso interpretativo, y en este trabajo argumentamos que es importante considerar que los “datos” son el resultado de procesos subjetivos y de decisiones humanas que las condiciona.

Materiales y metodología

Las colecciones de periódicos en hemerotecas son una importante fuente para el estudio de la historia, la literatura, la lingüística y otras disciplinas. El manejo de periódicos es notoriamente difícil en el ámbito bibliotecológico tanto en términos de organización documental como en su preservación. En términos de catalogación es

complicado identificar las temáticas de un periódico ya que las noticias y artículos que contiene un número son muchos y variados, por lo que generalmente se identifican solamente el título y la fecha, así como otros datos generales. Los periódicos suelen publicarse diariamente por lo que las hemerotecas generalmente se enfrentan a grandes volúmenes de documentos que son difíciles de almacenar, identificar y dar acceso. Aunado a esto los periódicos están impresos en papel de baja calidad lo que significa que se deterioran rápidamente.

La digitalización de periódicos se empezó a realizar hace ya varias décadas y esto ha permitido un renovado interés por los estudios hemerográficos (Latham & Scholes, 2006). Las colecciones digitales de periódicos permiten a los usuarios consultar los materiales y contribuye a la preservación de los mismos. En general se tiende a estudiar estos repositorios como si fueran sustitutos o suplentes de la versión impresa. La computadora se entiende como una simple ventana al archivo físico y no un sistema de remediación del archivo (Cordell, 2017). Así los investigadores realizan búsquedas y detectan periódicos relevantes pero interpretan los resultados como si hubieran consultado el archivo físico.

Algunos investigadores utilizan el sistema a través de algún tipo de catálogo y encuentran los periódicos por título o por fecha. Sin embargo, la mayoría de los repositorios de periódicos permiten también realizar búsquedas sobre el texto completo, algo que no es posible en el archivo físico, lo cual incrementa el alcance de los materiales que se consultan. Sin embargo, pocos investigadores están conscientes de que las búsquedas sobre el texto completo se hacen sobre archivos de texto plano extraído mediante OCR de las imágenes de los periódicos que ellos están viendo. Las palabras del periódico se convierten en "datos" a partir del momento en que un investigador introduce una palabra o una secuencia de palabras para ser identificadas en el texto del OCR. Sin embargo, salvo algunas excepciones, el investigador basa su marco interpretativo teniendo en mente el archivo físico y no el OCR. Y esa es una cuestión relevante, pues el texto OCR jugará un papel significativo en la obtención de los resultados, en este caso a partir de la búsqueda simple. Por ello es importante empezar a entender el OCR no solamente como un proceso técnico relegado a las máquinas, sino que debemos comprender que se trata de la transformación del objeto que incide directamente en la forma de obtener resultado, por lo que es fundamental saber qué ocurre para contextualizar los resultados desde una perspectiva humanística.

En nuestro caso hemos estudiado la Hemeroteca Nacional Digital de México (HNDM) que contiene más de 7 millones de imágenes y los archivos en XML producto del proceso de OCR correspondiente. El trabajo se realizó en el marco del proyecto de investigación “Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840-1914”, un proyecto de minería de datos en repositorios de periódicos digitalizados. El proyecto de investigación OcEx está compuesto por 6 equipos de investigación en 9 países y 11 proveedores de datos (colecciones hemerográficas) y busca identificar y modelar patrones de flujo de información en periódicos del siglo XIX a través de métodos computacionales. Es importante aclarar que OcEx no busca centralizar los repositorios digitales, sino establecer ontologías a partir de la interoperabilidad en las diferentes estructuras de metadatos para evitar las

complicaciones naturales a la multiplicidad de lenguajes participantes en este proyecto, para vincularlos como una red que ofrezca a los investigadores acceso a herramientas analíticas, conectadas a través de bases de datos remotas.

En un principio se consideró que el establecimiento de las ontologías sería relativamente sencillo. Sin embargo, al principio del proyecto de investigación quedó claro que si bien teníamos los datos no estaba suficientemente claro mucha información acerca de ellos: cómo están compuestos, cómo fueron adquiridos, qué tan precisos son, a qué objetos físicos corresponden, qué información metatextual o de metadatos incluyen y qué tan precisa o confiable es. Se realizó una detallada revisión de la documentación disponible, acompañada de una serie de entrevistas a actores claves en la formación de la HNDM así como observaciones del mismo equipo de investigación del proyecto al utilizar los datos de la HNDM para el proyecto.

Lynch (2002) hace una importante diferencia entre "colecciones digitales" y "bibliotecas digitales" que nos interesa retomar aquí. Según Lynch las colecciones digitales son los materiales "en crudo" (*raw materials*) en donde el enfoque está en crear grandes cantidades de contenidos digitales con herramientas muy básica de acceso y descubrimiento. Sobre estos "datos crudos" se realizan algunos trabajos "interpretativos" o de "curaduría" para dar significado y contexto a los contenidos. En cambio, considera que las bibliotecas digitales son "sistemas que hacen que las colecciones digitales cobren vida, los hace accesibles de una forma útil, de tal suerte que sean útiles para nuestro trabajo y las conecta con sus comunidades" (Lynch, trad.propia). En este mismo sentido, para Voutssas (2006) las bibliotecas digitales se caracterizan por contener no solo los materiales digitales sino también ofrecer servicios bibliotecarios (Voutssás 2006). ¿Pero cuáles podrían ser algunos de estos servicios para el caso de la investigación que se interesa por los datos?

Nosotros consideramos la HNDM como un repositorio y no como una biblioteca/hemeroteca digital porque no ofrece servicios bibliotecarios, en especial aquellos dirigidos a la investigación. Solo contiene una colección y, adicionalmente, el grado de catalogación y clasificación es muy básico. Aunado a esto, no brinda acceso directamente a fuentes básicas de datos como los XML formados a partir del OCR. Nuestro equipo de investigación trabajó sólo de manera muy limitada a través de la interfaz de la HNDM para identificar las noticias que podían ser de interés para la investigación. En cambio, utilizó de forma aislada los XML y sus metadatos asociados a estos documentos para realizar la minería de datos que era el objetivo final del proyecto. Es decir, el OCR contenido en los archivos XML fue la fuente de los datos para la investigación tanto en la forma más básica de la búsqueda, como para la extracción, mediante procedimientos computacionales como *sentimental analysis*, entidades nombradas, frecuencias, de la información.

El problema mayor, desde el punto de vista de la formación de colecciones y bibliotecas digitales es el de repensar que todos sus componentes digitales -metadatos, XML, OCR, imágenes, etcétera- tienen valor como representación de aspectos de los objetos digitalizados y, en ese sentido, poseen un alto valor para los usuarios de la colección.

Resultados

La HNDM es un proyecto de digitalización que tiene más de quince años y por lo tanto ha pasado por diversas etapas. La estructura básica sin embargo se ha mantenido. Cada imagen corresponde a una página de un periódico y posteriormente se utilizó un sistema de reconocimiento de caracteres (OCR). El OCR está insertado en un XML que también contiene los metadatos del periódico así como las coordenadas de cada palabra. Esto permite que cuando un usuario realice una búsqueda en la interfaz de la HNDM, la palabra que busca sale resaltada en amarillo en la misma imagen. Si bien es una excelente funcionalidad, contribuye a que el usuario interprete que está haciendo sus búsquedas sobre el periódico digitalizado. El archivo XML en donde se hace la búsqueda está oculto al usuario.

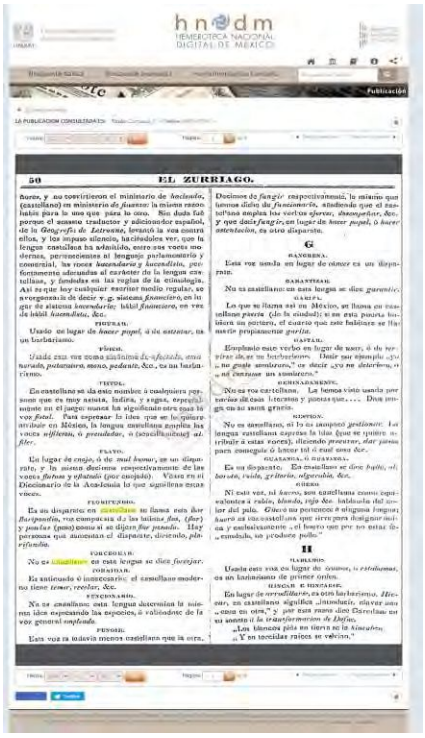

Fig. 1. Búsqueda por palabra en la HNDM. La palabra se resalta en amarillo.

El usuario común no tiene acceso a los archivos XML. Por lo tanto, no tiene la posibilidad de descargar o utilizar los datos sobre los cuales finalmente está haciendo sus búsquedas. Todos los periódicos que se encuentran en el dominio público están disponibles en acceso abierto. La HNDM que está resguardada por la Universidad Nacional Autónoma de México, tiene una política de apoyo a acceso abierto (Gaceta 2015). Los periódicos con acceso restringido solamente se pueden consultar en los equipos de cómputo que se encuentran en el edificio de la Hemeroteca Nacional.

Cuando el equipo de investigación solicitó acceso a los XML de las imágenes que están en acceso abierto se produjo cierto desconcierto ya que no existía una política o un procedimiento para esto. El acceso abierto se había entendido hasta ese momento como acceso a los periódicos, es decir, a las imágenes. El OCR de los periódicos no es el periódico, aunque es aparentemente el contenido del periódico. Una de las características de los periódicos es que están en columnas y algunas notas pueden empezar en una página pero continuar en otra. Los programas de reconocimiento óptico de caracteres pueden funcionar bastante bien con textos lineales de página completa y continua, como son los libros. El OCR aplicado a periódicos resulta complicado y es prácticamente imposible leer como humano el texto de un OCR como si fuera el objeto material o su versión digitalizada. Aunado a esto la calidad del OCR de periódicos es bastante deficiente, comparada con otros tales como libros modernos. Esto se debe por un lado a una compleja distribución de la página periodística y también a que se utiliza un papel delgado, amarillo y tinta de menor calidad, y tipografías que ya no son comunes en nuestros días.

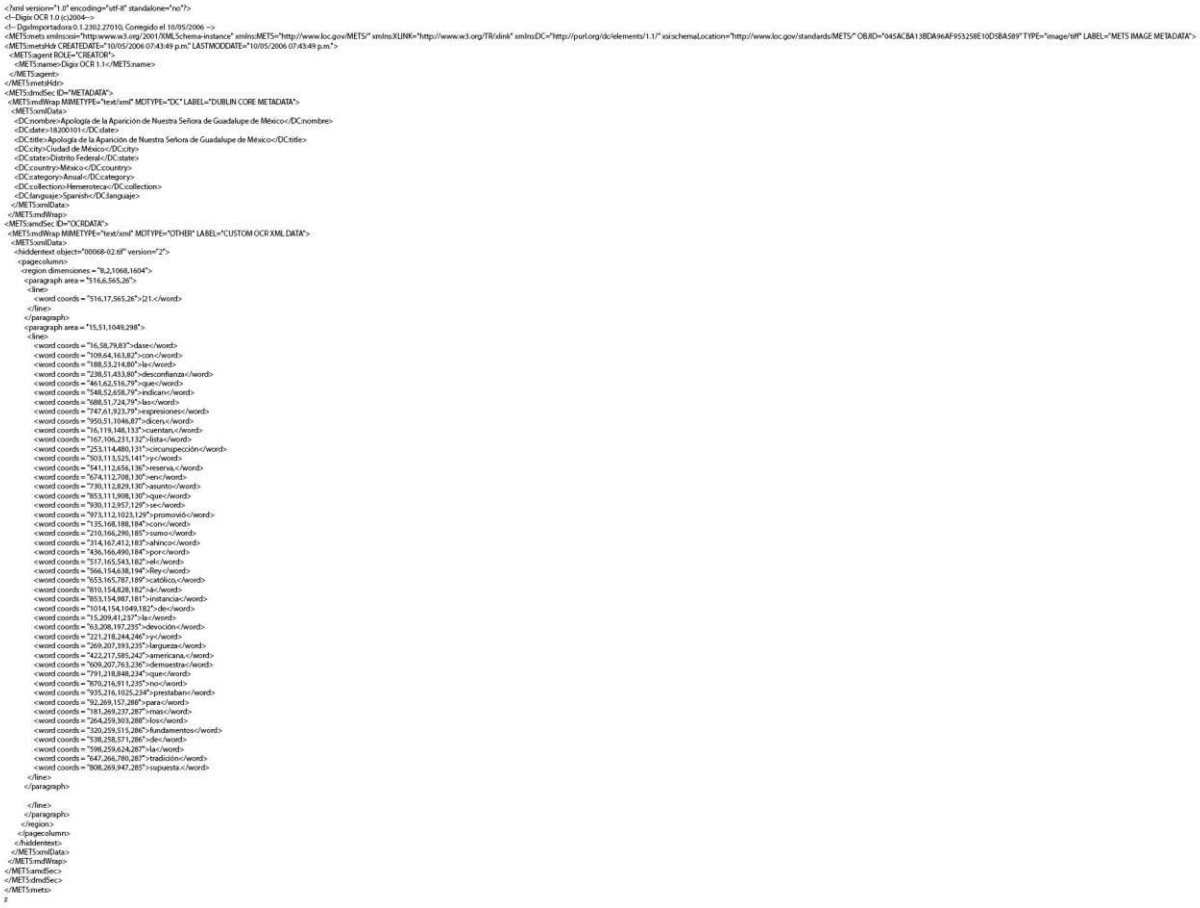
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<!--Digix OCR 1.0 (c)2004-->
<!-- DigiImportadora 0.1.2302.27010, Corregido el 10/05/2006 -->
<METS:mets xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:METS="http://www.loc.gov/METS/" xmlns:XLINK="http://www.w3.org/TR/xlink" xmlns:DC="http://purl.org/dc/elements/1.1/" xsi:schemaLocation="http://www.loc.gov/standards/METS/" OBJID="045ACBA138DA96AF95325BE10D5BA589" TYPE="image/tiff" LABEL="METS IMAGE METADATA">
<METS:metsHdr CREATEDATE="10/05/2006 07:43:49 p.m." LASTMODDATE="10/05/2006 07:43:49 p.m.">
<METS:agent ROLE="CREATOR">
<METS:name>Digix OCR 1.1</METS:name>
</METS:agent>
</METS:metsHdr>
<METS:dmdSec ID="METADATA">
<METS:mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="DUBLIN CORE METADATA">
<METS:xmlData>
<DC:nombre>Apología de la Aparición de Nuestra Señora de Guadalupe de México</DC:nombre>
<DC:date>18200101</DC:date>
<DC:title>Apología de la Aparición de Nuestra Señora de Guadalupe de México</DC:title>
<DC:city>Ciudad de México</DC:city>
<DC:state>Distrito Federal</DC:state>
<DC:country>México</DC:country>
<DC:category>Anual</DC:category>
<DC:collection>Hemeroteca</DC:collection>
<DC:language>Spanish</DC:language>
</METS:xmlData>
</METS:mdWrap>
<METS:amdSec ID="OCRDATA">
<METS:mdWrap MIMETYPE="text/xml" MDTYPE="OTHER" LABEL="CUSTOM OCR XML DATA">
<METS:xmlData>
<hiddentext object="00068-02.tif" version="2">
<pagecolumn>
<region dimensiones = "8,2,1068,1604">
<paragraph area = "516,6,565,26">
<line>
<word coords = "516,17,565,26">21.</word>
</line>
</paragraph>
<paragraph area = "15,51,1049,298">
<line>
<word coords = "16,58,79,83">dase</word>
<word coords = "109,64,163,82">con</word>
<word coords = "188,53,214,80">la</word>
<word coords = "238,51,433,80">desconfianza</word>
<word coords = "461,62,516,79">que</word>
<word coords = "548,52,658,79">indican</word>
<word coords = "688,51,724,79">las</word>
<word coords = "747,61,923,79">expresiones</word>
<word coords = "950,51,1046,87">dcen,</word>
<word coords = "16,119,148,133">cuentan,</word>
<word coords = "167,106,231,132">lista</word>
<word coords = "253,114,480,131">circunspección</word>
<word coords = "503,113,525,141">y</word>
<word coords = "541,112,656,136">reserva,</word>
<word coords = "674,112,708,130">en</word>
<word coords = "730,112,829,130">asunto</word>
<word coords = "853,111,908,130">que</word>
<word coords = "930,112,957,129">se</word>
<word coords = "973,112,1023,129">promovió</word>
<word coords = "135,168,188,184">con</word>
<word coords = "210,166,290,185">sumo</word>
<word coords = "314,167,412,183">ahinco</word>
<word coords = "436,166,490,184">por</word>
<word coords = "517,165,543,182">el</word>
<word coords = "566,154,638,194">Rey</word>
<word coords = "653,165,787,189">católico,</word>
<word coords = "810,154,828,182">a</word>
<word coords = "853,154,987,181">instancia</word>
<word coords = "1014,154,1049,182">de</word>
<word coords = "15,209,41,237">la</word>
<word coords = "63,208,197,235">devoción</word>
<word coords = "221,218,244,246">y</word>
<word coords = "269,207,393,235">largueza</word>
<word coords = "422,217,585,242">americana,</word>
<word coords = "609,207,763,236">demuestra</word>
<word coords = "791,218,848,234">que</word>
<word coords = "870,216,911,235">no</word>
<word coords = "935,216,1025,234">prestaban</word>
<word coords = "62,269,157,289">para</word>
<word coords = "181,269,237,287">mas</word>
<word coords = "264,259,303,288">los</word>
<word coords = "320,259,515,286">fundamentos</word>
<word coords = "538,258,571,286">de</word>
<word coords = "598,259,624,287">la</word>
<word coords = "647,266,780,287">tradición</word>
<word coords = "808,269,947,285">supuesta.</word>
</line>
</paragraph>
</line>
</paragraph>
</region>
</pagecolumn>
</hiddentext>
</METS:xmlData>
</METS:mdWrap>
</METS:amdSec>
</METS:dmdSec>
</METS:mets>
z

Fig.2. Archivo XML con OCR de página de un periódico de la HNDM

También identificamos que la HNDM contiene dos versiones de los archivos XML, los "sucios" que son el resultado del OCR y los "limpios" en donde se realizaron procesos computacionales para eliminar "basura" generada por el OCR, en particular caracteres problemáticos y frecuentes como %, &, *, por mencionar algunos. Estoy incluyó también los acentos. Las búsquedas se realizan sobre los archivos XML "limpios" por lo que las búsquedas ignoran palabras acentuadas o con estos caracteres problemáticos. Si bien, esto no necesariamente es crítico para todas las búsquedas, con algunos términos en donde el acento o algún carácter particular es importante los resultados arrojados por la búsqueda pueden ser problemáticos o deficientes. Esto no está indicado para los usuarios.

También se realizó un análisis de cómo están vinculados los títulos de los periódicos y sus metadatos con las imágenes correspondientes y el XML que lo debe acompañar. En las distintas etapas de la HNDM se migró el sistema en varias ocasiones a nuevas plataformas, tanto en la base de datos como las interfaces para acceder a ellas. Encontramos discrepancias y XMLs faltantes. En ocasiones entonces un usuario puede encontrar un periódico y su imagen y asume incorrectamente que el sistema de búsqueda de texto completo las incorpora. Sin embargo, como la búsqueda se realiza sobre el OCR, si el XML correspondiente no existe en el sistema, entonces no los incluye. Esto es algo que pasa desapercibido por el usuario.

Conclusiones

El OCR de periódicos digitalizados del siglo XIX presenta varios problemas de fidelidad con el original sin embargo, esto no significa que no puede ser utilizado para realizar estudios de minería de datos. Si pensamos en el acceso abierto, es importante que no sólo tengamos acceso a las imágenes de los periódicos pero si queremos poder procesarlos y entenderlos, también se requiere el acceso a los datos crudos, en este caso los archivos XML con el OCR y los metadatos.

Uno de los principales problemas para el uso de estos repositorios con fines de investigación es la falta de descripción adecuada de los datos, su contexto y la relación con los metadatos y la interfaz de búsqueda. Para lograr esto es necesario aproximarse a la digitalización, o mejor dicho, a los datos resultantes de una digitalización, desde una perspectiva crítica e interpretativa y no únicamente tecnológica. El OCR derivado de un texto histórico constituye una nueva edición del texto (Cordell, 2017, p.196). Por lo tanto consideramos que los repositorios que almacenan periódicos digitalizados deben de incluir una descripción hemerográfica no solo de la colección impresa sino también de la colección digitalizada, en donde se realice un trabajo de descripción bibliográfica para "dar cuenta de las fuentes, tecnologías y realidades sociales de su creación de manera que sus posibilidades y limitaciones sean más fácilmente visibles y estén disponibles para la crítica" (Cordell, 2017).

El uso de repositorios para custodiar colecciones de datos crudos permite compartir y reutilizar esta información. En este caso de estudio buscamos entender el OCR de periódicos digitalizados como datos crudos para la investigación histórica. Esto constituye un primer acercamiento y reflexión a cómo este tipo de datos deben de ser

almacenados en repositorios para que sean adecuadamente interpretados y utilizados en la investigación.

Referencias

Gaceta (2015). Lineamientos generales para la política de acceso abierto de la UNAM, 10 de septiembre 2015.

Latham, S., & Scholes, R. (2006). The Rise of Periodical Studies. *PMLA*, 121(2), 517-531. https://doi.org/10.1632/003081206X129693

Borgman, Christine L. 2009. «The Digital Future is Now: A Call to Action for the Humanities» 3 (4). http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html.

———. 2012. «The Conundrum of Sharing Research Data». *Journal of the American Society for Information Science and Technology* 63 (6): 1059-78. https://doi.org/10.1002/asi.22634.

Borgman, Christine L. 2015. *Big Data, Little Data, No Data Scholarship in the Networked World*. Cambridge, Massachusetts: The MIT Press.

Borgman, Christine L., Jilian Wallis, y Matthew Mayernik. 2012. «Who's Got the Data? Interdependencies in Science and Technology Collaborations». SSRN Scholarly Paper ID 2089165. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2089165.

Cordell, Ryan. 2017. «"Q i-Jtb the Raven": Taking Dirty OCR Seriously». *Book History* 20 (1): 188-225. https://doi.org/10.1353/bh.2017.0006.

Fenner, Martin, Mercè Crosas, Jeffrey S. Grethe, David Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, et al. 2019. «A data citation roadmap for scholarly data repositories». *Scientific Data* 6 (1). https://doi.org/10.1038/s41597-019-0031-8.

Gitelman, Lisa, ed. 2013. *«Raw data» is an oxymoron*. Infrastructures series. Cambridge, Massachusetts ; London, England: The MIT Press.

Gómez, Nancy-Diana, Eva Méndez, y Tony Hernández-Pérez. 2016. «Social Sciences and Humanities Research Data and Metadata: A Perspective from Thematic Data Repositories». *El Profesional de La Información* 25. http://eprints.rclis.org/30054/.

Lynch, Clifford. 2002. «Digital Collections, Digital Libraries and the Digitization of Cultural Heritage Information». *First Monday* 7 (5). https://doi.org/10.5210/fm.v7i5.949.

Schöch, Christof. 2013. «Big? Smart? Clean? Messy? Data in the Humanities». *Journal of Digital Humanities* 2 (Summer). http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

Voutssás, Juan. (2006) Bibliotecas y publicaciones digitales, Universidad Nacional Autónoma de México. Disponible en: http://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/L67/1/bibliotecas_y_publicaciones_digttales.pdf