Una frágil episteme
Los datos como objetos de conocimiento en humanidades

Ernesto Priani Saisó
Universidad Nacional Autónoma de México

Resumen

En la investigación en humanidades mediada por un entorno digital aparecen los datos como fuente de nuevos objetos de conocimiento. Este artículo analiza cuál es la naturaleza de los datos, de qué forma representan los objetos y cuáles son las dificultades para que los datos sean inteligibles. Para ello tomamos como ejemplo el trabajo desarrollado dentro del proyecto Intercambios Oceánicos, con el fin de mostrar que los datos se construyen por capas y que en cada una de ellas se ve comprometida su inteligibilidad. Es por ello que consideramos los datos como epistemológicamente frágiles y su constitución como objeto de estudio aún insuficiente para dar lugar a una nueva forma de conocimiento.

Abstract

In humanities research mediated by a digital environment, data appears as a source of new objects of knowledge. This article discusses the nature of the data, how they represent objects, and what difficulties are for the data to be intelligible. To do this, we take as an example the work developed within the Ocean Exchanges project, in order to show that the data is built in layers and that in each of them its intelligibility is compromised. That is why we consider the data to be epistemologically fragile and its constitution as an object of study even insufficient to give rise to a new form of knowledge.

Palabras clave

Datos; Humanidades; Humanidades digitales; Epistemología; Algoritmos

Key words

Data; Humanities; Digital humanities; Epistemology; Algorithms

Fecha de recepción: Agosto de 2019
Fecha de aceptación: Noviembre de 2019

Introducción

Aunque en el imaginario de muchos humanistas la tecnología digital no cambia ni debería cambiar, la naturaleza de su trabajo como han escrito, por ejemplo, Marche[1], Kirsch[2] y Brennan[3], es muy difícil creer que la mediación tecnológica a través de la cual se realiza la mayor parte de la investigación en humanidades no tenga ninguna consecuencia en la construcción del conocimiento.

Una rápida mirada a la historia de la utilización de herramientas digitales por parte de humanistas -filólogos, historiadores, geógrafos, estudiosos de la literatura, pedagogos- nos permite observar dos cosas:

Una, que esta no es tan reciente como podría imaginarse. Han pasado setenta años desde que el padre jesuita Roberto Busa acudiera a IBM con la propuesta de automatizar la creación del léxico en las obras de Tomás de Aquino. Y, dos, que la comunidad formada alrededor de la aplicación de estas herramientas tecnológicas que se identifica a sí misma como Humanidades Digitales, suele definirse precisamente por la utilización de estas herramientas, por ejemplo, en Sneha,[4] Schreibman.[5] Es decir, se presenta más como una comunidad de práctica que como una forma de producción de conocimiento distinta, intentando así mantener un equilibrio entre los campos disciplinares tradicionales, por un lado, y la introducción de los métodos y herramientas digitales por el otro.

Podemos observar entones que, a pesar de las décadas de utilización de herramientas digitales para la investigación, la comunidad que de forma expresa las ocupa no ha superado la distinción entre disciplinas y métodos digitales. Mientras tanto, la investigación humanística en general se ha visto inmersa en la generalización de la mediación digital, sobre todo en las últimas dos décadas. A nivel

---

1 Stephen Marche, "Literature Is Not Data: Against Digital Humanities". *Los Angeles Review of Books*. (Octubre 28, 2012) https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/. 23 de septiembre 2019.

2 Adam Kirschi, "Technology Is Taking Over English Departments". *The New Republic* (Mayo 2, 2014) https://newrepublic.com/article/117428/limits-digital-humanities-adam-kirsch. 23 de septiembre 2019.

3 Timothy Brennan, "The Digital-Humanities Bus". *The Chronicle of Higher Education* (Octubre 15, 2017). https://www.chronicle.com/article/The-Digital-Humanities-Bust/241424 23 de septiembre 2019.

4 Puthiya Purayil Sneha. *Mapping Digital Humanities in India.* (India: The Centre for Internet and Society, 2016). 14 ss.

5 Susan Schreibman, Siemens, Unsworth. "The Digital Humanities and Humanities Computing: An Introduction" en *A Companion to Digital Humanities.* Blackwell Companions to Literature and Culture 26 (2004).http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml&chunk.id=ss1-1-3&toc.depth=1&toc.id=ss1-1-3&brand=default 23 de septiembre 2019.

institucional se ha impulsado la creación de infraestructura digital que, junto con otros factores, como la adopción de políticas públicas a favor del uso de la tecnología, la creciente disposición de nuevas tecnologías para el investigador promedio y la cada vez más amplia gama de materiales digitalizados, han hecho que el entorno digital sea el espacio dentro del cual se desarrollen las humanidades.

Dicho de otra forma, las humanidades han quedado sumergidas -como muchos otros aspectos de nuestra vida contemporánea- dentro de lo digital. Cualquier investigación incluye, al menos, la búsqueda de libros y artículos en bibliotecas digitales, su lectura -si es el caso- en formatos digitales y la producción, como resultado de la investigación, lo mismo de textos originalmente digitales como de otros objetos digitales: audios, videos, modelados 3D, visualizaciones, presentaciones.

Esta inmersión de las humanidades en lo digital debe ser entendida en al menos tres grandes aspectos: a) a través de la aparición de nuevos objetos de conocimiento, b) la utilización de nuevos métodos de investigación y c) la utilización de otras formas de comunicación del conocimiento, ninguna de las cuales es neutral y, de manera separada y en conjunto modifican, en modos que aún no podemos apreciar en su totalidad, la naturaleza del conocimiento humanístico.

Es necesario, pues, como ya lo han sugerido Berry[6] y Hui,[7] entre otros, plantearse algunos problemas epistemológicos que subyacen a esta transformación de las humanidades. En este artículo me detendré a explorar el primero de los aspectos mencionados antes, la aparición de nuevos objetos de conocimiento y, en particular, el objeto de conocimiento que, en la última década, ha cobrado mayor relevancia dentro de las humanidades: los datos. Mi intención aquí es preguntar cuál es la naturaleza de los datos en tanto que objetos, y la forma como éstos se constituyen en objeto de conocimiento en el contexto de las humanidades.

Se trata sí, de preguntas teóricas a las que me propongo aproximar a partir de una experiencia concreta de trabajo con los datos dentro del proyecto "Intercambios oceánicos: Trazando redes de información global en repositorios de periódicos históricos, 1840-1914",[8] pues uno de los elementos más relevantes en

---

[6] David Berry "The Computational Turn: Thinking about the Digital Humanities". *Culture Machine* 12 (2011). https://sro.sussex.ac.uk/id/eprint/49813/1/BERRY_2011-THE_COMPUTATIONAL_TURN-_THINKING_ABOUT_THE_DIGITAL_HUMANITIES.pdf 23 de septiembre 2019.

[7] Yuk Hui. "A Phenomenological Inquiry on the Emergence of Digital Things". En *What Does a Chameleon Look Like?* Cologne: Herbert von Halem Verlag, 2011. 338 ss.

[8] Este trabajo forma parte de los resultados del proyecto de investigación "Intercambios Oceánicos: Trazando redes de información global en repositorios de periódicos históricos, 1840-1914", financiado por Conacyt (FONCICYT 274861).

la comprensión de los datos en las humanidades es cómo vienen a constituirse en objetos de conocimiento y esto, como se verá a continuación, se produce a través de una compleja red de acciones tecnológicas e institucionales que se entretejen con métodos de clasificación y representación, que es más claro comprender en un caso concreto y privilegiado como el de los datos utilizados en este proyecto.

Los datos y cómo encontrarlos

Intercambios oceánicos fue un proyecto internacional financiado en el marco de la convocatoria *Digging Into Data Challenge,* que tenía como objetivo utilizar herramientas digitales para estudiar el flujo de información entre periódicos históricos entre 1840 a 1914, a partir del acceso a las bases de datos de colecciones hemerográficas digtializadas en distintos países. A nivel internacional, se trabajó con colecciones como *The Times Digital Archive*, *Europeana Newspapers*, especialmente las colecciones de la: Austrian National Library, Berlin State Library, Hamburg State and University Library, Dr. Friedrich Tessman Library South-Tyrol, además *The Digital Collection of the National Library of Finland, British Newspaper Archive Library of Congress' Chronicling America Project* y, en nuestro caso, con la Hemeroteca Nacional Digital de México (HNDM) que consta de 7 millones de páginas de periódicos digitalizados.

En este contexto, el proyecto implicó adentrarse en la colección de periódicos decimonónicos de la Hemeroteca Nacional Digital de México (HNDM), ya no solo como una herramienta de búsqueda para apoyar al investigador a identificar diversas publicaciones históricas para su trabajo, sino como una fuente de datos para aquello que se quería estudiar. Dicho de otra manera, para este proyecto específico se dejó de tomar la HNDM como un reemplazo de los documentos originales, para tomar su contenido como objeto de estudio.

Hay que aclarar que el mayor valor no se encontraba en las imágenes digitalizadas de los periódicos, lo que el usuario común ve como resultado de su búsqueda, sino en unos archivos llamados ocr que el sistema utilizan para hacer las búsquedas -y de los que en última instancia depende el resultado-, así como de los metadatos con los que se clasifican las imágenes digitalizadas y que constituyen otro elemento para la representación de los datos de la hemeroteca. En suma, estábamos interesados en dos tipos de datos:

a) Los datos no estructurados

Son aquellos datos que carecen de una estructura interna identificable. Por ejemplo, un listado de números cualquiera, que no tendrán sentido hasta que no sepamos a que se refieren, precios, peso, medidas.

En esta categoría se encuentran los archivos ocr que son el resultado de procesar la imagen digitalizada de un texto en un sistema de reconocimiento óptico de caracteres. Los resultados varían notablemente dependiendo de si se trata de textos contemporáneos o históricos, del tipo de papel, de la tipografía, de si son libros, revistas o periódicos. En las mejores condiciones, se obtiene un texto altamente confiable respecto del original. Pero en otros, el resultado puede variar notablemente. Tomemos como ejemplo el archivo .txt resultante del proceso de ocr del diario *Bandera del Anáhuac*, publicado en Mérida Yucatán del 30 de octubre de 1827 es el siguiente:

> el p a trl0 ta sanjua nis ta pefo 3ic0 de mz2ri rad-e yu'catan z 7 m2'tcs 0 de ctubre t 7 o ds la iarflendcnclct s de lz 2-depzb2ca fedrczda y la qle el bten de lt patrm apeteeera os lt verdadero fit ltrcure 0s k'k 2 oficna a cargo cel cm cesureo aagts plaza de s 3t a v y ez sus vcnas el jdven ms ticrno deusnd igor circulur ya-la vcla per-tid sus encantos coro sbcn tods veneer y pelear e y aun it parca tembland0 respeta loot elero 6 as o"t cczztdilos un valor qu no tiene ejemlflar qe et dolores spiroii to2cfiar l6 tre e is filrd crtdzrzs todo cs arma del nuero guerrero y fftlm't"ttig ol r-o lb at aveededil'"del que osel oponcrse'el dttroi sumetnl'aliento r qtte-lo aguarda el horrorsepulcr51 tcnodado en los bronces tremendo esi'ofits qt cantar5 par'e dd caro altcrzzando cl 1 ahahuac lds t-ii0s valicntczs e depc'tado de-i empurtt'i las nrmas y cor-'em l'rt d cco terrible uarcinl arias once de luto v de sa're ue breaalo coil 1l'1 in'ttvtll libertid 6 1 muertc bttsclndo con tt'tuna ycn lucha no iual sti'e nochc sa'nda princpo de una litha qui no hade lclbar tcnga lcb us fuegooctltos quu ouos svn hs que td ltccs brillir a'2 [9]

Como puede verse, no todos los caracteres forman palabras reconocibles -a esto se le suele llamar un archivo ocr "sucio" que puede o no ser "limpiado" para eliminar de los caracteres sin sentido o corregir los "errores" de transcripción. Lo que constituye una intervención al archivo original. Adicionalmente, el archivo no tiene contenida su propia descripción. Es decir, no sabemos, en este caso concreto, por el archivo mismo a qué diario, a qué fecha y a qué página se corresponde el texto. Por lo tanto, no sabemos tampoco de que manera este resultado se ajusta a la imagen fuente del que se extrajo, pues el archivo tampoco indica a qué imagen se corresponde.

Ahora bien, pero con todas esas carencias, ¿se trata de datos? La respuesta es sí, si tomamos en cuenta que, como recoge Schöch[10] siguiendo a Fiori dato es la "ausencia de uniformidad".

---

[9] La versión de esta transcripción en la HNDM puede verse en http://www.hndm.unam.mx/consulta/publicacion/visualizar/558075bd7d1e63c9fea1a148?intPagina=0&tipo=publicacion

[10] Christof Schöch. "Big? Smart? Clean? Messy? Data in the Humanities". *Journal of Digital Humanities* 2. (Summer, 2013). http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/. 28 septiembre 2019.

Se trata, pues, de datos en un sentido muy genérico. Pero podemos observar también que se trata de datos que de manera inmediata no son inteligibles. Es decir, son insuficientes para hacer sentido de ellos y, por lo tanto, para constituirse en objetos de conocimiento. No satisfacen la definición de Borgman,[11] en tanto que los datos son "una representación reinterpretable de información en una manera formalizada que puede ser utilizada para comunicar, interpretar o procesar".

b) Datos estructurados

El segundo tipo de datos en el que estábamos interesados son los datos estructurados, que como el nombre lo indica, contienen las referencias suficientes para identificarlos. En el caso de la HDNM esos datos estructurados se corresponden con los metadatos que describen a qué ejemplar del periódico se corresponden los archivos y que están contenidos en los archivos .xml con que las imágenes son relacionadas.

Podemos ver aquí, entonces, el primer problema que se enfrenta al trabajar con datos en humanidades: que son insuficientes para constituirse como tales. Una palabra en un texto, un número en un documento, aun a pesar de ser un dato, solo lo será propiamente, en la medida en que sean completados con elementos que describen su naturaleza. Porque los datos no son inteligibles inmediatamente, requieren estar insertos en una estructura de clasificación, a la que se suele llamar metadatos. Esta está conformada por una familia de categorías, atributos y clases, pertenecientes a una ontología dentro de la cual los datos son descritos.

La estructuración de los datos y sus implicaciones

Para el caso de HNDM y el ejemplo tomado del diario *Bandera del Anáhuac*, los metadatos están preestablecidos siguiendo un criterio que ha sido decidido institucionalmente por la Hemeroteca Nacional, siguiendo a su vez estándares bibliográficos internacionales. De manera general, pero no exclusiva los metadatos consideran, entre otros, el título de la publicación, su carácter, la fecha, el lugar de publicación, la página, con lo cual provee una estructura de inteligibilidad para los datos. Es por ellos que podemos identificar el documento .txt obtenido del ocr como una representación de la edición de *Bandera del Anáhuac*, publicado en Mérida Yucatán el 30 de octubre de 1827.

La HNDM funciona no a partir del.txt original, sino de un archivo .xml en el que, como podemos ver a continuación, se incluyen los metadatos de fecha,

[11] Christine L. Borgman, "The Digital Future Is Now: A Call to Action for the Humanities". *Digital Humanities Quarterly 3*, no. 4 (January 2, 2010). https://escholarship.org/uc/item/0fp9n05s. 28 de septiembre 2019.

título, ciudad, estado, país, categoría, colección e idioma, y las coordenadas de cada palabra según la sección que ocupa en la imagen escaneada de la página. Con ello se logran dos cosas: hacer inteligible los datos, y hacerlos localizables en la imagen, que es lo que el usuario final verá tras hacer una búsqueda, pues el sistema señalará en amarillo sobre la imagen la palabra encontrada.

```
<!-- OriginFramework METS implementation 1.0 (c)2003 -->
<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/ele-
ments/1.1/" xsi:schemaLocation="http://www.loc.gov/
standards/METS/" OBJID="6BE4D33B1338C2E3A99CC70139EE80
0F" TYPE="image/tiff" LABEL="METS IMAGE METADATA">
<METS:metsHdr CREATEDATE="11/21/2003 10:54:26 AM" LAST-
MODDATE="11/21/2003 10:54:26 AM">
<METS:agent ROLE="CREATOR">
<METS:name>ORIGINFRAMEWORK1.0</METS:name>
</METS:agent>
</METS:metsHdr>
<METS:dmdSec ID="METADATA">
<METS:mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="DUB-
LIN CORE METADATA">
<METS:xmlData>
<DC:date>18271030</DC:date>
<DC:title>La Bandera de Anahuac</DC:title>
<DC:city>Merida</DC:city>
<DC:state>Yucatan</DC:state>
<DC:country>Mexico</DC:country>
<DC:category>Irregular</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>
<!--
OriginFramework OCR implementation v0.1
 Each word element has coordinates and a word value.
 The coordinates are xy based from the bottom left
 corner of the image.
```

```
-->
<hiddentext object="00002-01.tif">
<pagecolumn>
<region>
<paragraph>
<line>
<word coords="170,2188,314,67">
<![CDATA[ EL ]]>
</word>
<word coords="342,2235,346,10">
<![CDATA[ : ]]>
</word>
<word coords="382,2191,454,67">
<![CDATA[ P ]]>
</word>
<word coords="458,2191,761,69">
<![CDATA[ A_TR1,0 ]]>
</word>
```

Puede verse con claridad cómo este archivo autocontiene su descripción, describe cual es la fuente y cómo está organizado. Por supuesto, estos metadatos no son los únicos que se pueden utilizar. Para una investigación se pueden agregar categorías para la estructuración de los datos, ya sea en documentos xml, como categorías de una base de datos, como etiquetas para la clasificación de objetos, etcétera. Para el caso de los diarios del siglo XIX, algunas categorías posibles serían, por ejemplo, reportaje, nota, telegrama, pieza de opinión, como también puede ser calificada por la temática del contenido de la nota, por su posición política, por su estilo de escritura, entre un número muy amplio de otras posibilidades que dependerán, por supuesto, las intenciones del investigador.

De hecho, aunque el término "ontología" es utilizado en un sentido técnico para referirse a un sistema de categorías de clasificación, no oculta el problema al que se enfrenta la ciencia de datos en general, y de manera específica, la que se utiliza en humanidades: el problema ontológico que subyace a los datos, carentes de naturaleza propia en tanto no adquieran inteligibilidad conforme se construye su existencia como representación de algún objeto.

Por poner un ejemplo concreto: cuando al texto transcrito arriba se dice que pertenece a un grupo de categorías tales como periódico, nota, autor, tema, etcétera, se van formando, mediante la denotación de sus propiedades la naturaleza de los datos ante los cuales nos encontramos. De esta forma, los datos van emergiendo como representación de un objeto, en este caso una página de un periódico.

Ahora bien, la segunda cuestión con la que nos enfrentamos al trabajar con datos es la que señala Hui:[12] con qué precisión podemos capturar, en un sistema de clases, relaciones y atributos la naturaleza del objeto y el mundo al que pertenecen los datos, pero, sobre todo, "¿Cómo podemos abordar una cosa en las proposiciones que están sujetas a inferencia lógica sin perder objetividad?"

La dificultad de trabajar con datos cuya característica es de carecer de una naturaleza propia, es precisamente la confiabilidad de la representación de sus propiedades. Cuando a una nota de una publicación semanal del siglo XIX, le asignamos la categoría de "periódico", por ejemplo, ¿estamos siendo precisos? Es decir, cuál es el sentido de la palabra usada: periódico en tanto que es aparece con una periodicidad o, en un sentido más contemporáneo, periódico como diario. Con los datos en humanidades, el problema de su "objetividad" se traslada a las categorías que lo denotan y lo califican, que son en última instancia formas de representación conceptual del mundo.

Esto nos lleva al tema de dónde se original las categorías y por qué son estas y no otras las que utilizamos para describir ciertos datos. Sabemos que la mayoría de los sistemas de clasificación están establecidos por criterios institucionales y estándares internacionales, muchos de los cuales, especialmente los utilizados para la archivística y la clasificación bibliográfica, responden a una larga tradición que responde a objetivos e intenciones, propios de esas áreas.

Pero una cuestión adicional es que los metadatos utilizados en los sistemas, como por ejemplo en la HDNM, no son públicos. Eso implica que en ocasiones no sabemos qué criterios se siguieron para estructurar los datos y si estos no tienen algunas implicaciones específicas -favorables o no la investigación que se realiza-, o incluso implicaciones políticas y raciales. A fin de cuentas, las categorías que se usan en humanidades -pero por extensión también, en muchas aplicaciones del cómputo- tienen una historia y una política de tras de ellas, que determinan la construcción de los datos como objetos de conocimiento, y que acarrean tendencias que, vistas críticamente, condicionan o distorsionan los datos. Al respecto, Julia Flanders[13] ha discutido el problema de la representación del género dentro de las categorías de la Text Encoding Iniciative[14] (TEI) que parte del estándar ISO, mostrando las dificultades e implicación de seguir y no seguir ese estándar.

---

[12] Yuk Hui. "A Phenomenological Inquiry on the Emergence of Digital Things". En *What Does a Chameleon Look Like?* Cologne: Herbert von Halem Verlag, 2011. 342 ss.

[13] Julia Flanders, "Building Otherwise". En *Bodies of Information : Intersectional Feminism and Digital Humanities. Debates in the Digital Humanities.* Minnesota: University of Minnesota Press, 2018. 290-295.

[14] Las Categorías TEI son un estandar de marcado de texto desarrollado por un consorcio formado por univeridades de Estados Unidos y Europa para la representación digital de los textos https://tei-c.org/

Algo semejante pasa, como señalan Sculley y Pasanek,[15] cuando para los fines de una investigación formulamos categorías para estructurar los datos. Imaginemos que queremos organizar un archivo de documentos del siglo XX alrededor del debate sobre los movimientos guerrilleros en México. ¿Hasta qué punto las categorías que usemos reflejarán nuestra posición teórica respecto de ese debate? Es decir, ¿qué tan objetivas son nuestras categorías? ¿Qué tanto representan prejuicios que no son discutidos?

Aplicar y formular categorías para estructurar los datos conforma, como puede verse, un momento epistemológicamente crítico en el trabajo con datos. No sólo porque de ello dependen la inteligibilidad de los datos, sino porque nos confronta con la cuestión de que la representación del objeto está siempre atravesada por la institucionalidad, la historia, la identidad, la posición política de quienes hacen la representación. De modo que los datos en humanidades no pueden ser tomados como evidencia, como ocurre con muchos datos de las ciencias y de las ciencias sociales pues, como apunta Borgman:[16] "A falta de una prospectiva externa, los investigadores en humanidades necesitan estar particularmente atentos a las asunciones no declaradas sobre sus datos, fuentes de evidencia y epistemología".

Procesar los datos y la emergencia de nuevos objetos

Las dificultades de trabajar con datos en humanidades no terminan, sin embargo, aquí. En la definición de datos de Borgman citada antes, se indica que estos deben poder ser utilizados, entre otras cosas, para ser procesados. Algo que Schöch, en su propia definición de datos también señala: "los datos en humanidades pueden ser considerados como una *abstracción digital, selectivamente construida, accionable por una máquina que representa algún aspecto de un objeto dado de la investigación humanística*".[17]

Lo que esto quiere decir es que el archivo .xml que contiene el texto obtenido del ocr, descrito adecuadamente mediante sus metadatos, tiene que ser procesable por una computadora, lo que es posible, por supuesto, al tratarse de un archivo resultado de un proceso computacional. Pero no se trata sólo de la posibilidad técnica del procesamiento a lo que se refieren, sino al hecho de

---

[15] Sculley, D., and Bradley M. Pasanek. "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities". *Digital Scholarship in the Humanities* 23, no. 4 (September 12, 2008): 409–24. https://doi.org/10.1093/llc/fqn019. 28 septiembre 2019.

[16] Christine L. Borgman, "The Digital Future Is Now: A Call to Action for the Humanities". *Digital Humanities Quarterly 3*, no. 4 (January 2, 2010). https://escholarship.org/uc/item/0fp9n05s. 28 de septiembre 2019.

[17] Christof Schöch. "Big? Smart? Clean? Messy? Data in the Humanities". *Journal of Digital Humanities* 2. (Summer, 2013). http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/. 28 de septiembre 2019.

que mediante el procesamiento de los datos éstos adquieren una inteligibilidad distinta a la que posee a partir de su sola descripción. Por decirlo de otra manera, el procesamiento puede descubrir -porque no siempre lo logra- otras propiedades de los datos.

Los humanistas ocupan muchos tipos de datos y en distinto volumen. Para una investigación puede ser suficiente con la datación de un solo documento y, por tanto, la identificación de un único dato. Para otras, puede ser suficiente con los datos relativos a las publicaciones hechas por un editor en el siglo XVIII, tal vez unos 50 títulos distintos. Pero en ninguno de estos casos tiene sentido ocupar un proceso computacional.[18]

Pero pensemos ahora, por ejemplo, en la fecha y lugar de envío de los cables telegráficos sobre un acontecimiento específico, digamos la Explosión de buque Maine en la bahía de la Habana el 15 de febrero de 1898 hacia México, uno de los casos estudiados dentro de Intercambios Oceánicos. Al utilizar el buscador de la HNDM se identificaron 657 noticias relativas al acontecimiento en el lapso del 15 de febrero al 10 de marzo de 1889, pertenecientes a 13 publicaciones periódicas de la ciudad de México.

Ahora, comencemos por tratar de obtener un dato específico. ¿Cuál fue el primer telegrama enviado al México sobre la explosión en la Habana?

Este es un dato que puede ser identificado con cierta facilidad. Se trata de la nota publicada por *The Mexican Herald* del 16 de febrero de 1889. La nota dice así:

> THE CRUISER MAINE".
> Terrible Explosion on Board the Warship.
> KILLED AND WOUNDED.
> Believed that the Cruiser is Totally Destroyed by the Shock.
> (Associatted Press).
>
> LITTLE ROCK, Ark., Jan. 16, 1 a.m.- Night chief operator Shell at the St. Louis Western Union office has reported that he has received a report from New Orleans to the effect that the United States battle ship “Maine” was blown up in Havana harbor at 11:30.

Notemos que la noticia, en el original, tiene equivocada la fecha, pues en lugar de decir 16 de febrero, dice 16 de enero. La sabemos equivocada porque se refiere a un hecho acontecido el 15 de febrero, y a que aparece en la edición del 16 de febrero del *The Mexican Herard,* algo que se puede comprobar tanto recu-

[18] Salvo en el caso que se utilice un buscador para ello, en cuyo caso volveríamos al problema inicial: ¿están todos los datos que se buscan adecuadamente descritos para ser encontrados?

rriendo a la imagen del diario, como al ejemplar resguardado en la hemeroteca. También tiene mal escrita el nombre de la agencia de prensa, y ambos son un buen buen ejemplo de un dato equivocado desde su origen.

De manera individual, los datos de fecha y lugar de envío nos permiten identificar a éste como el primer telegrama sobre el evento que llega a la ciudad de México durante la madrugada de ese sábado 16. El telegrama además informa sobre el trayecto seguido por el mensaje para llegar a su destino y la persona, el jefe de operaciones nocturnas Shell, que lo transmitió.

Estos datos que son valiosos en el contexto en que se presentan, cambian cuando pasan a formar parte de la serie de todas las fechas y lugar desde los cuales fueron expedidas las noticias referentes a la Maine entre este día y hasta el 10 de marzo.

En principio, se modifica la relación del dato con el texto al que pertenece. Si en el ejemplo anterior servía para clasificar la noticia como la primera en una larga secuencia, en el momento en que es integrado al resto de las fechas y de los lugares del conjunto de las 657 noticias, y es relevante saber quién y a través de qué medio llegó, estos datos pierden conexión con el texto en el que se encuentran y a partir del cual cobran inteligibilidad. En su lugar, pasa a ser sólo un dato en una secuencia, que adquieren inteligibilidad a partir de que son procesados computacionalmente. En cuyo caso, si no se corrige el error de datación, aparecerá como una anomalía entre datos que deberían corresponder a unas fechas determinadas.

Aquí entran en juego un nuevo elemento a considerar: los algoritmos. Baste decir por ahora, que son uno de los instrumentos principales para el trabajo con datos. Son concebidos como una serie ordenada de operaciones matemáticas para obtener un resultado y se pueden desarrollar con fines de procesamiento específico, pero también encontramos hoy algoritmos de uso muy amplio tanto por las distintas comunidades de investigación, como por sistemas de cómputo que usamos con frecuencia. En el contexto de proyecto Intercambios Oceánicos, hicimos uso de diversos algoritmos: desde el sistema de búsqueda de la HNDM, a algoritmos de entidades nombradas, análisis de sentimientos, unigramas y bigramas, entre otras herramientas.

En humanidades, se utilizan los algoritmos para dar acceso a propiedades de los datos que solo son visibles cuando se trabaja con un grupo de ellos. Es decir, permite acceder a aquellas propiedades que son compartidas por un grupo de datos dentro de un marco específico y, en esa medida, cambia la perspectiva por la cual los datos son significativos y pueden constituirse en un objeto de conocimiento.

En el ejemplo de la datación de un documento, un solo dato bastó para que se produjera conocimiento. Pero si nuestra pregunta no es cuándo se escribió

este texto o cuándo y desde dónde se envió el primer telegrama del estallido del Maine, sino de dónde provenían las noticias que llegaron a México sobre el estallido del Maine, podemos darnos cuentas que buscamos patrones de frecuencia que nos permitan saber algo más acerca de los datos que tenemos y por lo tanto formular hipótesis sobre lo que esos patrones representan.

Dentro de los resultados que se obtuvieron para el caso del Maine en el proyecto fue el poder determinar que el 35% de los telegramas relativos al evento estaba fechados en Washington y tardaban en ser publicados en la ciudad de México un promedio de dos días. Mientras que los telegramas fechados en la Habana -lugar del evento- no rebasaban el 15% y tardaban los mismos dos días en ser publicados.

Se puede inferir con facilidad cómo el procesamiento de las noticias da lugar a la identificación de una nueva propiedad en los datos, que solo es visible al interior del conjunto total de datos: su valor dentro del conjunto. Pero cómo este nuevo dato, producto de toda la serie de procesos que se han descrito hasta aquí, ¿puede constituir un nuevo objeto de conocimiento para las humanidades?

Los datos obtenidos para la frecuencia de noticias provenientes de la Habana y de Washington, ese 35% y 15%, serán un objeto de conocimiento para un humanista en la medida en que cumplan con un requisito general: 1) sean interpretables dentro de un marco conceptual previamente definido, pues no todo resultado producto del procesamiento de datos, como muestran Sculley y Pasanek,[19] puede serlo a veces como consecuencia de un problema con los datos de origen, en otros casos con la herramienta utilizada, en ocasiones por insuficiencias en el marco conceptual en que se inserta.

En el caso del ejemplo del Maine, esos datos pueden ser interpretados a partir de comprender el estallido de Maine no sólo como un evento singular, acontecido en una fecha determinada, sino como un evento que se desenvuelve en el tiempo en planos distintos: el estallido, la investigación de lo ocurrido, el entierro de los marineros, la disputa diplomática. En esa medida, se comprende cómo la fuente y el centro de atención se desplaza de un lugar a otro, y muestra cómo la producción de información en lugares y por instituciones determinadas, tiene una mayor representación en la prensa que otras.[20]

En el contexto de la disputa diplomática entre España y Estados Unidos por Cuba, la mayor representación de los envíos de Washington que de la Habana y

---

[19] Sculley, D., and Bradley M. Pasanek. "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities". *Digital Scholarship in the Humanities* 23, no. 4 (September 12, 2008): 409–24. https://doi.org/10.1093/llc/fqn019. 28 de septiembre 2019.

[20] Se puede argumentar que estas ideas están reflejadas en la forma cómo se seleccionan los datos a trabajar, lo que por supuesto plantea el problema de la circularidad al que se ve expuesto el trabajo con datos. ¿Estas sus conclusiones inducidas desde un principio?

Madrid (11%), nos permiten conocer algo nuevo sobre la prensa en el siglo XIX y la forma como ésta no sólo refleja los conflictos, sino que juega un papel en la producción de un dominio informativo alrededor del conflicto diplomático.

Algunas consideraciones finales

Si algo debe sobresalir al recorrer todo este camino en la construcción de los datos como objeto del conocimiento en humanidades es su fragilidad. En el caso específico examinado, los datos comenzaron con la extracción computacional de un texto a partir de la imagen de una publicación. Esos datos fueron clasificados a partir de un conjunto de metadatos que denotaban sus propiedades y que incluían la descripción de la fuente, así como otros metadatos, por ejemplo, que eran noticias sobre el estallido del Maine en unas fechas determinadas, y estos datos fueron finalmente procesados por unos algoritmos que nos arrojaron resultados estadísticos.

Pero no hay un punto, en este proceso, en que la representación de las noticias extraídas de un diario del siglo XIX, no presenten una dificultad. Ya vimos que el *ocr* suele ser un archivo sucio con numerosos errores que pueden o no subsanarse. En nuestro caso lo hicimos, porque el corpus de trabajo no era demasiado extenso, sin embargo, para proyectos mucho más amplios, es imposible hacerlo. Los errores en el *ocr* se trasladan al xml que usamos para recuperar las noticias relativas al Maine y eso plantea preguntas específicas para nuestro caso, pero que son preguntas comunes en cualquier proyecto digital: ¿recuperamos todas las entradas que había sobre el estallido del Maine en la HNDM?

Pero pasando de largo por estas primeras dificultades, ¿las categorías que usamos fueron las adecuadas? ¿No introdujeron ninguna distorsión en la obtención de datos? ¿todos los datos se ajustan corresponden a las categorías designadas? Pasados estos problemas, aun enfrentamos uno más, ¿fue el algoritmo de entidades nombradas, por ejemplo, el más eficaz para obtener este resultado? ¿Contempló todas las veces que aparecía la Habana o Washington con fuente en las notas? ¿No se agregaron noticias que no tuvieran como origen esas ciudades, pero las mencionaran? ¿El resultado estadístico relevante y comprensible?

Lo que estas preguntas señalan es la incertidumbre que rodea cada uno de los pasos hasta llegar a los resultados obtenidos, aun cuando estos sean alcanzados en algunos casos por la aplicación de fórmulas matemáticas. De la incertidumbre que rodea cada una de las capas a través de las cuales los datos adquieren inteligibilidad, se desprende su fragilidad como objeto de conocimiento. Está condicionado y limitado por todas las dificultades epistémicas que encierra cada una de las capas en el proceso.

Los datos, como objetos de conocimiento son epistemológicamente frágiles. Su fuente, construcción y procesamiento exigen ser cautelosos en extremo. Los datos ofrecen una representación provisional, incompleta, parcial del objeto de estudio. En cada una de las diferentes capas de las que están formados pueden encontrarse carencias, imprecisiones, confusiones, sin que sea posible, además, saber cuándo se han subsanado por completo.

Su fragilidad, sin embargo, no impide que se constituyan en objetos de conocimiento, y que estos sean distintos de los objetos tradicionales de conocimiento humanístico.

Aun queda trabajo por hacer para describir estos objetos. En algunos casos, se trata de objetos con una escala distinta a los que suelen ocupar a los humanistas y, por ello, permiten una perspectiva diferente y un estudio distinto (puede verse, por ejemplo, Franco Moretti, Distant Reading,[21] Michel y Kui Shen con su propuesta de análisis cuantitativo de la cultura[22] o Ryan Cordell[23] al investigar los textos virales en periódicos del siglo XIX).

Pero no sólo se trata sólo de objetos que tengan esa escala y que se diferencien por su transversalidad, volumen o duración. También el procesamiento de datos permite trabajar con datos en una escala menor, para construir otros objetos de conocimiento. Es el caso del trabajo con corpus limitados, a través del marcado de texto, para obtener lo que Schöch[24] llama *Smart Data*, datos descritos con mayor precisión y claridad, con una mayor comprensión de las implicaciones que la categorización tiene en la formación de los datos y que deben servir para producir conocimiento más sólido a partir de ellos.

Pero ¿la construcción de estos nuevos objetos es un indicio de la emergencia de una episteme distinta a la que es propia de las disciplinas humanísticas?

Debemos imponernos cierta cautela a la hora de responder a esta pregunta. Por un lado, podemos observar cómo aparecen nuevos objetos de estudio e interés para los humanistas. Pero su aparición no es suficiente para hablar de una forma distinta de producir conocimientos. Hay discrepancias en cuanto a si se trata de legítimos objetos para las humanidades. Ni siquiera el hecho de incorporar métodos novedosos para estudiarlos garantiza que el resultado sea

---

[21] Franco Moretti, *Distant Reading*. London ; New York: Verso, 2013.

[22] Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg. "Quantitative Analysis of Culture Using Millions of Digitized Books". *Science* 331, no. 6014 (January 14, 2011): 176. https://doi.org/10.1126/science.1199644. 28 de septiembre 2019.

[23] Ryan Cordell, "Q i-Jtb the Raven': Taking Dirty OCR Seriously". *Book History* 20, no. 1 (2017) https://doi.org/10.1353/bh.2017.0006. 28 de septiembre 2019 p. 188.

[24] Christof Schöch. "Big? Smart? Clean? Messy? Data in the Humanities". *Journal of Digital Humanities* 2. (Summer, 2013). http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/. 28 de septiembre 2019.

un conocimiento humanístico original y de interés. y si a partir de ellos se pueden hacer aportes significativos para el conocimiento.

Además, como se ha mostrado aquí, la conformación de esos nuevos objetos preludia una episteme frágil. Los resultados obtenidos se sustentan siempre en objetos siempre pendientes de ser robustecidos. De modo que es necesario profundizar aun más en su epistemología, en sus formas de validación, e integrar y depurar los instrumentos de estudio. De otra forma, en un entorno digital dentro del cual ya se produce la mayor parte del conocimiento humanístico seguirán piezas sueltas.

Hay un último problema que se debe afrontar y que es el término mismo de humanidades donde emergen estos nuevos objetos. Si bien se trata de un concepto que sintetiza bien a qué campo pertenecen esos objetos, implica una transversalidad que debe ser discutida. ¿Por qué hablar de los datos para las humanidades y no de manera específica para sus disciplinas? ¿Qué implicaciones tiene esta decisión en la conformación de una nueva episteme para las humanidades?

Es difícil saber si estamos, efectivamente, frente a la formación de una nueva manera de conocer entre los humanistas. Pero si, en este campo hay todavía mucho que reflexionar, no hay duda, sin embargo, de que se está trabajando de otra manera, como otros instrumentos, sobre unos objetos complejos que solo se van haciendo comprensibles, en la medida en que son reconocidos y estudiados.

Bibliografía

Berry, David. "The Computational Turn: Thinking about the Digital Humanities". *Culture Machine* 12 (2011). https://sro.sussex.ac.uk/id/eprint/49813/1/BERRY_2011-THE_COMPUTATIONAL_TURN-_THINKING_ABOUT_THE_DIGITAL_HUMANITIES.pdf 23 de septiembre 2019.

Borgman, Christine L. "The Digital Future Is Now: A Call to Action for the Humanities". *Digital Humanities Quarterly* 3, no. 4 (January 2, 2010). https://escholarship.org/uc/item/0fp9n05s. 28 de septiembre 2019.

Brennan, Timothy. "The Digital-Humanities Bus". *The Chronicle of Higher Education*, (Octubre 15, 2017). https://www.chronicle.com/article/The-Digital-Humanities-Bust/241424 23 de septiembre 2019.

Cordell, Ryan. "'Q i-Jtb the Raven': Taking Dirty OCR Seriously". *Book History* 20, no. 1 (2017): 188–225. https://doi.org/10.1353/bh.2017.0006. 28 de septiembre 2019.

Flanders, Julia. "Building Otherwise". En *Bodies of Information : Intersectional Feminism and Digital Humanities. Debates in the Digital Humanities.* Minnesota: University of Minnesota Press, 2018. 289-304.