Revista Cubana de Información y Comunicación alcance ISSN 2411-9970

Los metadatos de la Hemeroteca Nacional Digital de México. Aproximaciones para una ontología

Metadata for Digital National Newspaper Library of México. Approximations for an ontology.

Miriam Peña Pimentel[1] https://orcid.org/0000-0003-3457-8700
Isabel Galina Rusell[1*] https://orcid.org/0000-0001-5286-5733

[1]Universidad Nacional Autónoma de México (UNAM), Instituto de Investigaciones Bibliográficas. México.

[*]Autor para la correspondencia: igalina@unam.mx

RESUMEN

Cada vez es más común que se realicen investigaciones con la utilización de colecciones digitalizadas. Los resultados que arrojan los sistemas de búsqueda dependen de los metadatos y organización documental que por lo general, son invisibles para el usuario. Existe poco trabajo en torno a estudiar cómo están estructurados los metadatos y de qué formas permiten pero también condicionan los resultados de las búsquedas, actuando como una especie de mediador entre el usuario y la colección. El objetivo de este artículo es reflexionar en torno a los sistemas de organización documental y su impacto en los resultados de una investigación, a partir de un ejemplo concreto, el de la Hemeroteca Nacional Digital de México, a través del ejercicio de crear una ontología basada en el esquema de estructuración de metadatos de la Hemeroteca Digital a partir de JSON.

Palabras clave: ontología; metadatos; hemeroteca digital.

ABSTRACT

Digital collections are used more and more frequently by researchers. Search results depend on metadata and document organization which is generally invisible to the users. There is little research on digital collection metadata structure and how this enables but

also conditions search results, acting as a mediator between the user and the collection. The aim of this article is to reflect on document organization for digital platforms and its impact on research results, using Mexico's national newspaper library, the Hemeroteca Nacional Digital de México (HNDM) as a case study by creating an ontology based on the metadata schema in JSON of the digital collection.

Keywords: Ontology, metadata, digital newspaper collection

Recibido: 02/01/2021
Aceptado: 20/01/2021

Introducción

Las colecciones de periódicos digitalizados han generado en las últimas décadas un renovado interés por las investigaciones hemerográficas (Latham y Scholes, 2006). Las colecciones hemerográficas en línea permiten a los usuarios no solo acceder de forma remota a una cantidad importante de materiales, sino que también ofrecen nuevas formas de consultar las colecciones a través de sofisticadas interfaces de consulta y recuperación. En particular, el uso del *Optical Character Recognition* (OCR) para las búsquedas permite recuperar ejemplares buscando en el texto completo, algo que no es posible con las colecciones impresas. Para esto, los metadatos y sistemas de clasificación y catalogación que utilizan estas colecciones son de gran importancia. Las aproximaciones a estas colecciones digitales sin embargo, suelen enfatizar los aspectos de la visualización de las páginas de los periódicos y existe menos trabajo en torno a estudiar cómo están estructurados los metadatos y de qué formas permiten pero también condicionan los resultados de las búsquedas, actuando como una especie de mediador entre el usuario y la colección.

Los sistemas de organización documental son claves para los sistemas de búsqueda y recuperación de estas colecciones. Sin embargo, los metadatos y cómo están integrados al sistema de búsqueda generalmente son invisibles para los usuarios y el buscador es utilizado "a ciegas", asumiendo que estamos recuperando toda la información relevante.

El punto de partida de esta reflexión deriva del proyecto de investigación *Intercambios oceánicos: Trazando redes de información global en repositorios de periódicos*

históricos, 1840-1914, en donde utilizamos herramientas digitales para estudiar el flujo de información entre periódicos históricos digitalizados. El proyecto global involucró el trabajo de personas de once instituciones en seis países y la utilización de diversas colecciones digitalizadas nacionales 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 la Hemeroteca Nacional Digital de México.

A diferencia de otros proyectos en donde se busca concentrar todas las colecciones a estudiar en un solo repositorio, aquí se planteó que a través del establecimiento de ontologías sería posible integrar los datos para nuestros fines de investigación. Las ontologías se crean estableciendo conceptos y categorías con sus propiedades y las relaciones que existen entre ellas. Para crear ontologías complejas y significativas es necesario tener conocimiento sobre los datos. Como comentan Galina & Priani (En prensa) investigadores principales del proyecto: "uno de los primeros descubrimientos fue la poca información disponible acerca de cómo estaba compuesta la colección, no tanto en términos de contenidos (qué periódicos), sino en su estructura y su impacto en las herramientas de búsqueda y recuperación". En este trabajo describimos los resultados en torno al trabajo para la creación de la ontología, usando el mismo caso de estudio, el de la Hemeroteca Nacional Digital de México (HNDM).

En primera instancia es necesario conocer los diferentes formatos de organización disponibles a colecciones de este tipo de recursos para después entrar en materia describiendo la HNDM y elaborando nuestra propuesta.

El objetivo de este artículo es reflexionar en torno a los sistemas de organización documental, a partir de un ejemplo concreto, el de la HNDM, a través del ejercicio de crear una ontología basada en el esquema de estructuración de metadatos del sistema.

Desarrollo

Información: tipos y organización

La información estructurada es información que ha sido analizada o procesada de alguna forma con el fin de dividirla en sus componentes estructurales. Para poder realizar este

procesado es necesario que previamente se haya determinado qué tipo de estructura se quiere reconocer en la información, ya que depende en gran medida de cuál es la utilidad que se le vaya a dar a la misma. Sin embargo, y frente a este estado ideal de la información, es habitual encontrar la información como información no estructurada, que constituye la mayoría de las veces la fuente de información primaria, y que es aquella que no ha sido tratada, que no tiene definido un modelo de datos asociado o a la que todavía no se le ha asignado un esquema de estructura interna.

El ejemplo más paradigmático de información no estructurada es el del texto, sobre el que se pueden realizar diversos tipos de análisis, según se atienda a su configuración física (caracteres, palabras, líneas, párrafos, secciones, etc.), a su configuración sintáctica (partículas elementales, sintagmas, oraciones, etc.), a su contenido semántico (qué tópicos aborda y cómo los relaciona entre sí), etc.

En 1998, Shilakes y Tylman (1998) hicieron una estimación acerca de la situación en que se encontraba la información potencialmente usable dentro del mundo de los negocios y llegaron a la conclusión de que aproximadamente un 80% de la misma estaba en forma no estructurada. Para 2010 (Grimes, 2008), se estimó que la cantidad de datos no estructurados aumentará alrededor de un 800% en 5 años, y que su ritmo de crecimiento es un orden de 10 a 50 veces más rápido que el de la información estructurada.

Muchas veces una ventaja de estructurar la información es que para poder lograrlo es preciso dar un nivel explícito de representación que clasifique los componentes que la conforman. Este proceso de nombrar las cosas y clasificarlas es un paso hacia la creación de una ontología que permita exportar el procedimiento realizado en el modelo a otros modelos similares, dando como resultado una metodología general de investigación. Al planificar un proyecto de estructuración para analizar información hay dos cuestiones principales que deben abordarse: qué estructuras son importantes y cuáles han de codificarse. El cómo han de codificarse es, sin duda, importante, pero su importancia pertenece a un nivel cualitativamente inferior al hecho de qué se codifica y qué no, por lo que suele ser decidido con posterioridad, cuando se han aclarado las cuestiones anteriores. Cualquiera de estas decisiones se enfrenta, en cualquier proyecto, a decisiones de tipo económico (tiempo, recursos necesarios, etc.) e intelectuales.

A continuación, se presentan algunas cuestiones específicas que pueden ayudar a reconocer qué componentes y relaciones han de ser tenidas en cuenta:

1. ¿Seguiría considerándose el componente si se efectúa una reordenación de la información?
2. ¿Puede ser útil para otros propósitos además del específico para el que se plantea?
3. ¿Existe algún otro caso en el que ya se esté utilizando?
4. ¿Será un componente que sirva para realizar búsquedas sobre él y caracterice (o ayude a caracterizar) el texto posteriormente?
5. ¿Está relacionado de alguna forma con algún otro componente que pueda ser interesante?

El punto final para decidir qué estructura codificar es plantearse si existe ya un estándar que permite realizarlo de manera más sencilla, lo que podría facilitar muchas de las decisiones de diseño (por las respuestas y éxitos/fracasos experimentados en situaciones similares) y evitar sorpresas posteriores que obliguen a rehacer parte del trabajo.

Una de las áreas importantes de investigación en la bibliotecología es el problema de la catalogación, ordenamiento, organización documental, recuperación, etc. Los intentos de organizar grandes masas de información han dado lugar a métodos organizativos que han cumplido, con mayor o menor acierto, su función. Tradicionalmente, estas grandes masas de información están concentradas en las bibliotecas, en las que la diversidad y cantidad de información han requerido de una invención continua de metodologías que permitieran la recuperación de los documentos bajo distintos criterios. Es por ello que la mayoría de los métodos que encontramos desde un punto de vista histórico tienen una fuerte relación con la ordenación en este tipo de material. Sin embargo, encontramos ejemplos de formas de organización, principalmente de aplicación en épocas más recientes, que no están directamente relacionadas con la ordenación y catalogación de publicaciones.

Las bibliotecas gestionan grandes cantidades de datos a los cuales implementa una estructura, a pesar de que -en su mayoría- lo que resguarda sean contenidos textuales. A partir de implementar búsquedas en texto completo usando archivos con reconocimiento óptico de caracteres (OCR por sus siglas en inglés) dentro de un repositorio, la recuperación de texto se hace posible; pero hay que notar que la estructuración de la colección se hace a partir de los metadatos, no de los contenidos textuales (Peña Pimentel, 2011).

Los problemas a los que se enfrentan los métodos sistemáticos de organización se agrupan en cuatro tipos principales:

1. Fuentes de información dispersas. Un problema que se presenta más frecuentemente con la aparición de tecnologías que permiten la distribución y almacenamiento masivo de una forma descentralizada.
2. Una cantidad ingente de información, que supera con creces la capacidad humana y hace imposible su catalogación manual.
3. Una alta variedad de temáticas, que dificulta una catalogación clara.
4. Alto número de relaciones entre los elementos a organizar, lo que dificulta el uso de metodologías que hacen uso de una jerarquización fuerte.

Existen diferentes formas para organizar la información, algunos de estos son comunes a la catalogación y la descripción de metadatos de las bibliotecas/hemerotecas. Sin embargo, antes de concretar un modelo es necesario conocer las características de los materiales contenidos, las necesidades del usuario y las recomendaciones de interoperabilidad entre sistemas de recuperación de información; teniendo en cuenta que la mayoría de bibliotecas/hemerotecas digitales cuentan también con un acervo físico de los materiales digitalizados, por lo cual la elección de los estándares de metadatos debe contemplar la materialidad de los documentos originales y buscar una relación entre los dos formatos (físico y digital) de los “mismos” materiales.

Revisaremos algunos de los modelos comúnmente utilizados en bibliotecas digitales:

Tesauro: Es un intento por clasificar ordenadamente la mayor cantidad posible de información derivada de un dominio específico. Según algunas definiciones, un tesauro es una lista que contiene los "términos" empleados para representar los conceptos, temas o contenidos de los documentos, con miras a efectuar una normalización terminológica que permita mejorar el canal de acceso y comunicación entre los usuarios y las unidades de información.

La construcción básica de un tesauro es semejante a la de un diccionario, la diferencia principal entre ambos es que el tesauro se construye a partir de sinónimos y campos relacionados que pueden explicarse por sí solos ya que no sólo contiene definiciones, sino información adicional, situando la explicación del término y su utilidad bajo un marco de acción específico (este caso se da especialmente cuando el término que se define es ambiguo y hace falta este contexto para aclarar su significado). Por lo general la estructura de un tesauro es jerárquica y parte del sustantivo y sus variaciones semánticas, aunque

hay tesauros en los que sus relaciones son funcionales, de sinonimia, antonimia, etc. (Gilchrist, 2003).

Bases de datos: Normalmente, se suelen considerar también las bases de datos como un método de organizar la información, aunque desde un punto de vista computacional representa más bien un marco general en el que se pueden proyectar diversos métodos de organización. Una base de datos es una colección de datos organizados y estructurados según un determinado modelo de información que refleja no sólo los datos en sí mismos, sino también las relaciones que existen entre ellos. Una base de datos se diseña con un propósito específico y debe ser organizada con una lógica coherente. Los datos podrán ser compartidos por distintos usuarios y aplicaciones, pero deben conservar su integridad y seguridad al margen de las interacciones de ambos. La definición y descripción de los datos han de ser únicas para minimizar la redundancia y maximizar la independencia en su utilización (Lapuente & Lapuente, s/f).

En este sentido, una biblioteca puede considerarse una base de datos compuesta en su mayoría por documentos y textos impresos en papel e indexados para su consulta. En la actualidad, y debido al desarrollo tecnológico de campos como la informática y la electrónica, la mayoría de las bases de datos están en formato digital, hasta el punto de que ya va ligado el concepto de base de datos con el de base de datos digital.

Ontología: Otro método de organización y clasificación de información que se aleja del mundo de las publicaciones, y que ha tenido una gran influencia y uso extendido en el mundo científico (y últimamente con un uso extensivo en su nueva adaptación al mundo informático) son las ontologías. Una ontología es una formulación de un exhaustivo y riguroso esquema conceptual dentro de un dominio dado, con la finalidad de facilitar la comunicación y la compartición de la información entre diferentes sistemas. Típicamente, las ontologías en este campo se relacionan estrechamente con vocabularios fijos (lo que se conoce como ontología fundacional) con cuyos términos debe ser descrito todo lo demás.

La intención de un sistema estandarizado de metadatos es la de imponer una descripción (clasificación, catalogación y demás) a una serie de recursos y que éstos sean recuperables y, sobre todo que no sean restrictivos a un único sistema; es decir, que sean interoperables y potenciables.

Las ontologías hacen posible construir una semántica a partir de la estructuración de metadatos: “Una ontología es una especificación de una conceptualización, esto es, un

marco común o una estructura conceptual sistematizada y de consenso no sólo para almacenar la información, sino también para poder buscarla y recuperarla. Una ontología define los términos y las relaciones básicas para la compresión de un área del conocimiento, así como las reglas para poder combinar los términos para definir las extensiones de este tipo de vocabulario controlado" (Otero-Cerdeira, Rodríguez-Martínez y Gómez-Rodríguez, 2015).

La implementación de una ontología trata de convertir la información en conocimiento mediante una estructura de conocimiento formalizada que referencie los datos, por medio de estándares de metadatos, sobre algún dominio de conocimiento.

Los beneficios de utilizar ontologías se pueden resumir de la siguiente forma:

- Proporcionan una forma de representar y compartir el conocimiento utilizando un vocabulario común.
- Permiten usar un formato de intercambio de conocimiento.
- Proporcionan un protocolo específico de comunicación.
- Permiten una reutilización del conocimiento.

En resumen, una ontología es un sistema de representación del conocimiento que resulta de seleccionar un dominio o ámbito del conocimiento, y aplicar sobre él un método con el fin de obtener una representación formal de los conceptos que contiene y de las relaciones que existen entre dichos conceptos. Además, una ontología se construye en relación a un contexto de utilización. Esto quiere decir que una ontología especifica una conceptualización o una forma de ver el mundo, por lo que cada ontología incorpora un punto de vista. Además, una ontología contiene definiciones que nos proveen del vocabulario para referirse a un dominio. Estas definiciones dependen del lenguaje que usemos para describirlas. Todas las conceptualizaciones; definiciones, categorizaciones, jerarquías, propiedades, herencia, etc. de una ontología pueden ser procesables por máquina.

Según Gruber (1995) las ontologías se componen de:

- Conceptos: son las ideas básicas que se intentan formalizar. Los conceptos pueden ser clases de objetos, métodos, planes, estrategias, procesos de razonamiento, etc.
- Relaciones: representan la interacción y enlace entre los conceptos de un dominio.

- Suelen formar la taxonomía del dominio. Por ejemplo: subclase-de, parte-de, parte-exhaustiva-de, conectado-a, etc.
- Funciones: son un tipo concreto de relación donde se identifica un elemento mediante el cálculo de una función que considera varios elementos de la ontología. Por ejemplo, pueden aparecer funciones como: asignar-fecha, categorizar-clase, etc.
- Instancias: se utilizan para representar objetos determinados de un concepto.
- Reglas de restricción o axiomas: son teoremas que se declaran sobre relaciones que deben cumplir los elementos de la ontología. Por ejemplo: "Si A y B son de la clase C, entonces A no es subclase de B", "Para todo A que cumpla la condición B1, A es C", etc. Los axiomas, junto con la herencia de conceptos, permiten inferir conocimiento que no esté indicado explícitamente en la taxonomía de conceptos.

Las posibles aplicaciones y usos de las ontologías son:

- Repositorios para la organización del conocimiento.
- Servir de herramienta para la adquisición de información.
- Servir de herramientas de referencia en la construcción de sistemas de bases de conocimiento que aporten consistencia, fiabilidad y falta de ambigüedad.
- Normalizar los atributos de los metadatos aplicables a los documentos.
- Posibilitar el trabajo cooperativo al funcionar como soporte común de conocimiento.
- Permitir el tratamiento ponderado del conocimiento para recuperar información de forma automatizada.
- Permitir la construcción automatizada de mapas conceptuales y mapas temáticos.
- Permitir la interoperatividad entre sistemas distintos.
- Establecer modelos normativos que permitan la creación de la semántica de un sistema y un modelo para poder extenderlo y transformarlo entre diferentes contextos.
- Servir de base para la construcción de lenguajes de representación del conocimiento.

Tesauros *vs* ontologías

La Documentación y la Bibliotecología han tenido que asumir el impacto de Internet y sus tecnologías asociadas. La progresiva digitalización de las bibliotecas afecta los recursos de información, las herramientas de representación y recuperación, y los requerimientos de los usuarios. Son varias las disciplinas que estudian la representación de información: la Lingüística, la Inteligencia artificial, la Documentación, la Ingeniería lingüística; las cuales han producido herramientas de recuperación de la información como taxonomías, sistemas de clasificación, lexicones computacionales, bases de datos léxicas, tesauros, listas de encabezamiento, bases de conocimiento, mapas conceptuales, ontologías, anillos de sinónimos y redes semánticas, etc.

Dentro de este tipo de herramientas, especialmente en la Bibliotecología destacan dos, los tesauros y las ontologías. Revisaremos las diferencias y similitudes entre ambas. De acuerdo con Arano (2004), "Un tesauro es una herramienta documental utilizada en el ámbito de la representación y recuperación de información, que representa un ámbito del conocimiento determinado mediante su estructuración conceptual". Los tesauros proporcionan estructuras conceptuales y de significado de los términos representados; son estructuras jerarquizadas con asociaciones y equivalencias. Mientras que "[u]na ontología es una representación formal y explícita de la estructura conceptual de un campo del conocimiento. Una ontología es el soporte semántico de las palabras que son descritas como objetos lingüísticos en una base de datos léxica o terminológica" (Arano, 2004), a diferencia del tesauro, las ontologías tienen relaciones conceptuales menos jerarquizadas puesto que cada organización corresponde al campo de conocimiento que estructura. Algunas de las ventajas de una ontología, por encima de un tesauro son:

- Una ontología puede estar elaborada de acuerdo con diferentes requerimientos y, al mismo tiempo, puede funcionar como un esquema de base de datos, como una auténtica base de conocimiento, para definir varias tareas o aplicaciones.
- Una ontología potencia la comunicación entre humanos y ordenadores.
- Una ontología promueve la normalización y reutilización de la representación de la información mediante la identificación del conocimiento común y compartido.
- Las ontologías añaden valor a los tesauros a través de una semántica más profunda, así como desde un prisma conceptual, relacional e informático. Una mayor profundidad semántica puede implicar niveles más profundos de jerarquía,

relaciones enriquecidas relaciones entre clases y conceptos, así como la capacidad de formular reglas de inferencia, etc.

Ontologías y bases de datos

La principal diferencia entre las ontologías y las bases de datos se encuentra entre OWA (Open World Assumption) y CWA (Close World Assumption). Si bien las ontologías utilizan el sistema OWA de representación del conocimiento, las bases de datos utilizan el CWA. Una base de datos explota la UNA (Asunción de nombre única) para nombrar entidades. Cualquier información que falte en un sistema de base de datos tiene el valor de "0". Cualquier elemento de información que falte en un sistema de ontología se considera desconocido (Sir, Bradac y Fiedler, 2015).

Otra diferencia importante entre las ontologías y las bases de datos es el propósito para el que se crean. Mientras que las ontologías se centran en agregar significado y comprensión, las bases de datos se concentran en el almacenamiento de datos. En otras palabras, las bases de datos se crean como almacenes de datos efectivos, mientras que las ontologías se forman para una mejor comunicación, interoperabilidad y como el puente de comunicación entre un ser humano y una máquina.

Ambos sistemas utilizan diferentes métodos de creación. Una base de datos se crea desde cero, lo que significa que todas las tablas y sus contenidos están prediseñados. Cuando diseñamos un sistema de ontología, intentamos aprovechar las ontologías existentes o la estructuración de sistemas sobre una ontología existente (extendiendo una ontología existente).

Al crear un sistema de base de datos, aplicamos la normalización de tablas; dicha normalización se utiliza para eliminar datos redundantes de las tablas y reducir la complejidad y creamos formularios. Los formularios son un conjunto de reglas que ayudan a corregir la transformación de las entidades y las relaciones con la estructura de la disposición física de las tablas.

La metodología para la creación de ontología no incluye formas normales. Un importante método de creación de ontologías consiste en patrones de diseño. Estos patrones, sin embargo, no son tan estrictos como las formas normales: en lugar de eso, crean reglas generales. En estos patrones se reconocen seis áreas: estructural, sintáctico, contenido, presentación, consideración y correspondiente. (Ontology Design Patterns, s/f)

¿Cómo se construye una ontología?

Para la construcción de una ontología es necesario tomar en cuenta la existencia de ontologías similares, conocer las posibilidades de integración entre ontologías y seguir una serie de características que se describen a continuación:

- Posibilidad de integración: el marco que se aplique para construir la ontología debe permitir algún tipo de reutilización del conocimiento (reconstrucción de la ontología en un marco diferente).
- Identificación de los supuestos y los compromisos ontológicos: los aspectos presentados se describen en el modelo conceptual y en las especificaciones del documento de requisitos de la ontología.
- Reconocer el conocimiento que se representará en cada módulo: es necesario determinar qué conocimiento debe representarse en cada componente básico.
- Conocimiento de las ontologías candidatas: este imperativo se subdivide en encontrar ontologías disponibles y elegir entre ellas las que podrían constituir posibles candidatos para la integración.
- Elección de las ontologías de origen: debemos elegir la ontología de origen que mejor se adapte a nuestras necesidades y propósitos. El mejor candidato es el que se puede adaptar mejor o más fácilmente. Se debe considerar la compatibilidad entre ontologías y su relación con la ontología que estamos creando.
- Aplicar las operaciones de integración: las operaciones de integración relacionadas especifican cómo se incluirá el conocimiento de una ontología integrada y se combinará con el conocimiento de la ontología resultante o se modificará antes de su inclusión.
- Analizar la ontología resultante: después de la integración del conocimiento, uno debe evaluar y analizar la ontología resultante. Además de exhibir un diseño adecuado y el cumplimiento de los criterios de evaluación, la ontología debe tener un nivel uniforme general. La ontología resultante debe ser consistente y coherente en todo (Sir et al., 2015).

Obtención de una ontología a partir de una base de datos

Hay muchos métodos y soluciones disponibles para facilitar la conversión de datos desde una base de datos relacional. López et al. (2011) llamaron a este enfoque Ontología de Ingeniería inversa. En general, estas técnicas permiten el acceso a los datos almacenados en la base de datos. En un primer momento se deben identificar las similitudes entre ambas herramientas y crear las correspondencias entre clases y tablas; atributos y características y restricciones y axiomas. Esto se logra realizando una serie de pruebas con los diferentes recursos de la base de datos:

- Usando los datos de la base de datos
- Utilizando un análisis de consulta
- Utilizando un análisis del esquema relacional de la base de datos (López et al., 2011)

La clasificación muestra el tipo de información conocida sobre la base de datos, la cual se utilizará para la construcción de la ontología; sin embargo, se debe tomar en cuenta la pérdida de información.

Es cierto que las bases de datos proporcionan comodidad, seguridad y facilitan el intercambio de datos; garantizan accesibilidad a los datos y su almacenamiento es confiable y "a prueba" de errores. Sin embargo, esta estructura no presenta información semántica que sí se presenta en la ontología. La elección entre una herramienta u otra depende directamente del uso que se le dará a la información estructurada. Por ejemplo, para un estudio humanístico, el contenido semántico y la contextualización de los datos aportarán información importante para el estudio de los mismos.

Estructura de los metadatos de la HNDM: JSON

JSON ("JSON", s/f) es un metaformato que es útil para expresar estructuras de datos basadas en árboles. El beneficio clave de JSON es que expresa de forma relativamente clara los tipos de estructuras de datos que se utilizan comúnmente en los lenguajes de programación, en particular en Javascript, lo que hace que sea muy fácil de crear y consumir.

JSON rara vez se utiliza con un esquema declarativo y, por lo general, se procesa analizándolo en las estructuras de datos nativos utilizadas por un lenguaje de

programación particular. La notación de objetos de JavaScript (JSON) es un formato de texto para la serialización de datos estructurados. Se deriva del objeto *literals* de JavaScript, tal como se definen en la Programación ECMAScript Idioma estándar, en su tercera edición [ECMA-262].

JSON puede representar cuatro tipos primitivos (cadenas, números, booleanos, y nulo) y dos tipos estructurados (objetos y matrices). Una cadena es una secuencia de cero o más caracteres Unicode [UNICODE].

Un objeto es una colección desordenada de cero o más nombre / valor pares, donde un nombre es una cadena y un valor es una cadena, número, booleano, nulo, objeto o matriz.

Una matriz es una secuencia ordenada de cero o más valores. Los términos "objeto" y "matriz" provienen de las convenciones de JavaScript.

Los objetivos de diseño de JSON eran que fuera mínimo, portátil, textual y un subconjunto de JavaScript.

La referencia a ECMA-404 es normativa, no con el significado habitual de que los implementadores deben consultarla para entender este documento, pero para enfatizar que no hay inconsistencias en la definición del término "texto JSON" en cualquiera de sus especificaciones. Sin embargo, ECMA-404 permite varias prácticas que esta especificación recomienda evitar en el Intereses de máxima interoperabilidad.

La intención es que la gramática sea la misma entre los dos documentos, aunque se utilizan diferentes descripciones. Si hay una diferencia entre ellos, ECMA y el IETF trabajarán juntos para actualizar ambos documentos. Si se encuentra un error con cualquiera de los documentos, el otro debe ser examinado para ver si tiene un error similar. Si lo hace, debería ser fijo, si es posible. Si alguno de los documentos se modifica en el futuro, el ECMA y el IETF trabajarán juntos para garantizar que los dos documentos permanezcan alineados a través del cambio.

En la HNDM la anotación JSON está concentrada en un sistema de base de datos NoSQL orientado a documentos de código abierto llamado MongoDB. Esta base de datos no guarda los datos en tablas (como se hace en una base de datos relacional); en su lugar guarda estructuras de datos con un esquema dinámico, esto hace que la integración de datos sea más fácil y que la información de los documentos tenga mayor complejidad que la que podría tener en una tabla.

La HNDM utiliza esta notación desde 2015, momento en que se liberó la versión actual; sin embargo su construcción comenzó en 2014 tras dos años de reconocimiento y

evaluación de los materiales digitales, sus metadatos y la revisión de otros posibles formatos de estructuración para crear la base de datos.

La Hemeroteca Digital cuenta con el mismo formato de metadatos (BEJ) con distintos tipos de contenido; los cuales, a su vez se relacionan con dos tipos principales: $.publicación y con _id, como se muestra en la figura 1. La HNDM cuenta con nueve tipos de contenidos que agrupan las 45 etiquetas de los metadatos y que pueden observarse en el anexo.

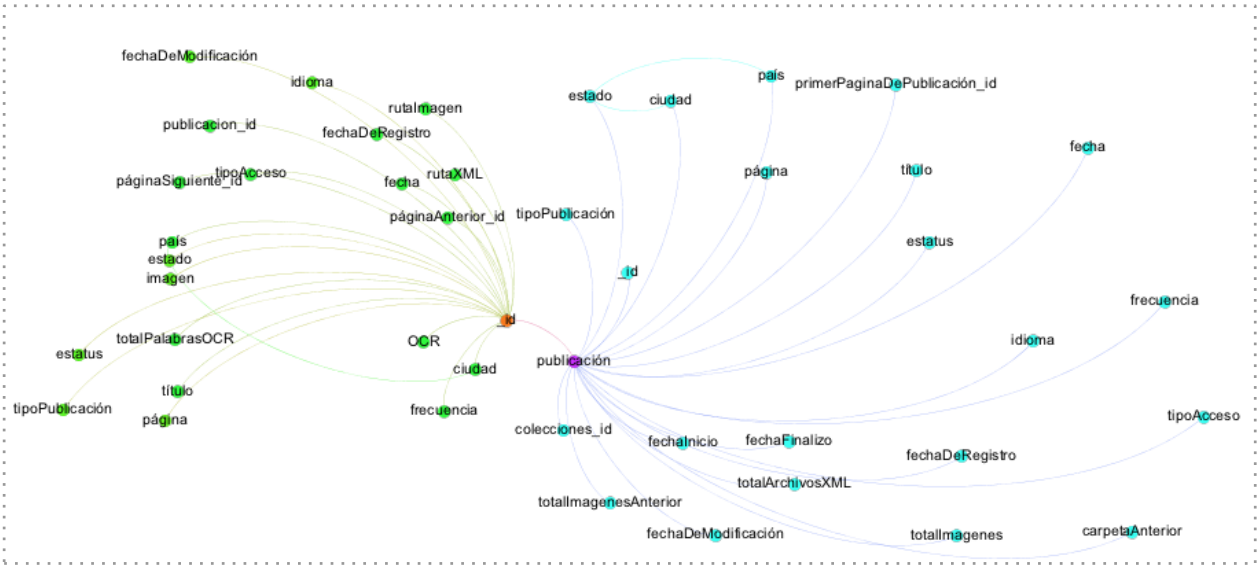

Fig. 1- Metadatos HNDM (JSON).

La base de datos de la HNDM cuenta con una aparente neutralidad en la organización de la base de datos que aloja y clasifica los metadatos para su despliegue. Al implementar una ontología debe considerarse que esa aparente neutralidad se pierde (Hajmoosaei & Skoric, 2016); dada la intrínseca jerarquización que implica ontologar contenidos, la estructura en la cual se van a reorganizar los metadatos dependerá directamente de la ontología creada. Así pues, la implementación de una ontología en un sistema ligeramente jerarquizado (como lo es la base de datos que gestiona esta colección) cambia el propósito del mismo y, aunque gana independencia en la recuperación de recursos específicos, la funcionalidad con respecto al patrón establecido y contextualización en la recuperación de datos, se hace más complejo.

El ejercicio propuesto para esta ontología responde directamente a las necesidades de investigación del proyecto Intercambios Oceánicos y, principalmente, a la necesidad de generar una ontología interoperable con otros sistemas de estructuración ya que la HNDM está estructurada en una base de datos en JSON. Para la creación de la ontología usamos

como fichero base el esquema JSON (Fig. 2) en la cual está construida la base de datos de los metadatos, no se implementó ninguna ontología adicional (externa), sino que se generó a partir de sí misma (Munir & Sheraz Anjum, 2018).

A primera vista es evidente que este proceso hizo que la nueva organización pasara de 140 líneas de código a 3611; esta modificación en la extensión del código fuente se debe, principalmente, a que cada elemento ha obtenido "independencia", por ejemplo, aquellos elementos que se encontraban dispuestos a manera de cadena (*string*), ahora son elementos independientes; pero mayor jerarquizados. A pesar de que la estructura se vuelve más compleja y desanuda las cadenas, permitiendo una búsqueda por elemento mínimo; la recuperación de información bajo este formato se hace más compleja, se hace más lenta.

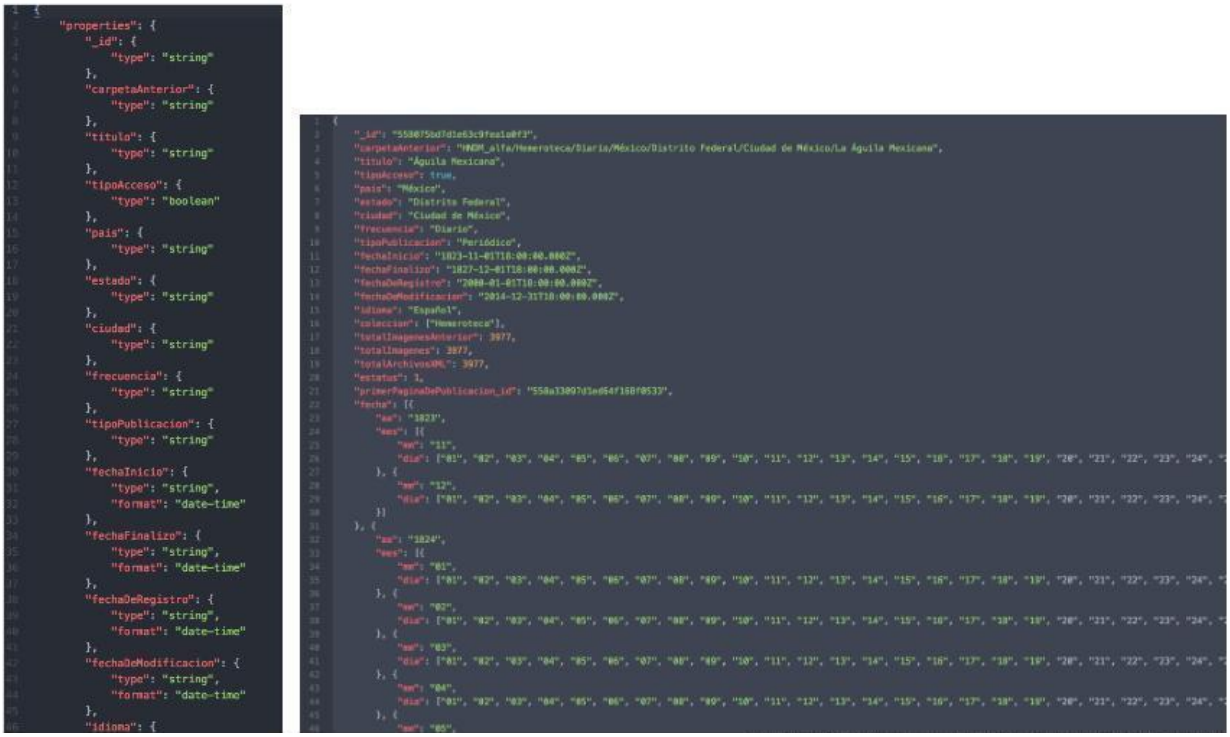

Fig. 2- Ontología (derecha) y Esquema original JSON (izquierda).

Al no imponer una ontología externa (alguna estructurada por idioma, como era el objetivo de este experimento), la interoperabilidad de nuestra propuesta se cierra en sí misma. Es decir, de implementarse la ontología propuesta, la hipótesis es que lograremos que la recuperación de datos sea mucho más específica y que los resultados complementarios tengan mayor relación contextual con la búsqueda realizada; pero no necesariamente será posible entablar un diálogo con el resto de recursos disponibles a

este proyecto (el resto de las colecciones de hemerotecas digitales), a menos que éstos implementen nuestra ontología como sistema de estructuración.

Dada la magnitud de los recursos disponibles, utilizar una ontología común convertiría esta masa de información en un sistema complejo, el cual restringe su interoperabilidad a los elementos comunes -independientemente de la jerarquización-, como la paginación, fechas y demás elementos no textuales que están condicionados al idioma y su grafía. Sin embargo, si se lograse la implementación de una ontología multilingüe, la recuperación de datos se potencializa dada la interoperabilidad generada entre los sistemas de estructuración y, principalmente, dada la falta de restricción dentro de la búsqueda en los OCR de los periódicos digitalizados.

Agradecimientos

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). Para más información ver: http://iib.unam.mx/intercambiosoceanicos.

Referencias bibliográficas

Arano, S. (2004). La ontología: Una zona de interacción entre la Lingüística y la Documentación. *Hipertext.net*, (2). Recuperado de http://www.hipertext.net

Castells, P. (2005). La web semántica. En C. Bravo Santos & M. Á. Redondo Duque (Eds.), *Sistemas interactivos y colaborativos en la web*. Cuenca: Ediciones de la Universidad de Castilla-La Mancha.

Galina, I, & Priani Saisó, E. (En prensa). Políticas de digitalización para la investigación. El caso de la HNDM y el proyecto Oceanic Exchanges. En *Humanidades Digitales. Una visión desde México*. Querétaro, México: Fondo Editorial de la Universidad Autónoma de Querétaro.

Gilchrist, A. (2003). Thesauri, taxonomies and ontologies – an etymological note. *Journal of Documentation*, *59*(1), 7–18. https://doi.org/10.1108/00220410310457984

Grimes, S. (2008). *Unstructured Data and the 80 Percent Rule*. Recuperado de Forrester Research website: http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-

the-80-percent-rule/

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, *43*(5), 907–928. https://doi.org/10.1006/ijhc.1995.1081

Hajmoosaei, A., y Skoric, P. (2016). Museum Ontology-Based Metadata. *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, 100–103. https://doi.org/10.1109/ICSC.2016.74

JSON. (s/f). Recuperado el 24 de septiembre de 2019, de https://www.json.org/json-es.html

Latham, S. y Scholes, R. (2006). The Rise of Periodical Studies. *PMLA* 121 (2): 517-31. https://doi.org/10.1632/003081206X129693

Lapuente, C. L., y Lapuente, M. J. L. (s/f). Bases de datos [Tesis]. Recuperado el 24 de septiembre de 2019, de http://www.hipertexto.info/documentos/b_datos.htm

Lopez, G., Servetto, A. C., Echeverría, A., Jeder, I., Grossi, M. D., & Rey, E. J. (2011). *Ontologías en arquitecturas dirigidas por modelos*.

Munir, K., y Sheraz Anjum, M. (2018). The use of ontologies for effective knowledge modelling and information retrieval. *Applied Computing and Informatics*, *14*(2), 116–126. https://doi.org/10.1016/j.aci.2017.07.003

Noel, Yuhanna. (2010, noviembre). Government Summit. Recuperado el 17 de julio de 2011, de MarkLogic Newsletter website: http://newsletter.marklogic.com/2010/11/government-summit-noel-yuhanna-forrester-keynote/

Ontology Design Patterns. Org (ODP)—Odp. (s/f). Recuperado el 24 de septiembre de 2019, de http://ontologydesignpatterns.org/wiki/Main_Page

Otero-Cerdeira, L., Rodríguez-Martínez, F. J. y Gómez-Rodríguez, A. (2015). Ontology matching: A literature review. *Expert Systems with Applications*, *42*(2), 949–971. https://doi.org/10.1016/j.eswa.2014.08.032

Peña Pimentel, Miriam. (2011). *El Gracioso en el Teatro de Calderón: Un análisis desde las Humanidades Digitales* (Doctoral, UWO). Recuperado de http://ir.lib.uwo.ca/etd/307

Shilakes, C. C. y Tylman, J. (1998). *Enterprise Information Portals*. Recuperado de Merryl Linch website: http://ikt.hia.no/perep/eip_ind.pdf

Sir, M., Bradac, Z. y Fiedler, P. (2015). Ontology versus Database. *IFAC-PapersOnLine*, *48*(4), 220–225. https://doi.org/10.1016/j.ifacol.2015.07.036

Anexo

TIPO DE CONTENIDO: Nulo (Null)	TIPO DE CONTENIDO: Coordenadas	TIPO DE CONTENIDO: Fecha (Date)
ETIQUETA: $, página	ETIQUETA: OCR	ETIQUETA: fechadeModificación, fechaDeRegistro, fechaFinalizo, fechaInicio
DESCRIPCIÓN: contenido nulo	DESCRIPCIÓN: 4 coordenadas numéricas que delinean un segmento en una imagen	DESCRIPCIÓN: fecha (ISODate)
VALOR: ninguno	VALOR: ninguno	VALOR: ninguno
RELACIÓN: $ y $.publicación (respectivamente)	RELACIÓN: $.publicación	RELACIÓN: $.publicación
TIPO DE CONTENIDO: Fecha (Date)	TIPO DE CONTENIDO: Fecha (Date)	TIPO DE CONTENIDO: Nombre del archivo (File name)
ETIQUETA: fecha, fechadeModificación, fechaDeRegistro	ETIQUETA: fecha	ETIQUETA: carpetaAnterior
DESCRIPCIÓN: fecha (ISODate)	DESCRIPCIÓN: rango de fecha (aa,mes,día)	DESCRIPCIÓN: colección (HNDM_NAS)
VALOR: ninguno	VALOR: ninguno	VALOR: ninguno
RELACIÓN: $.publicación.página	RELACIÓN: $.publicación	RELACIÓN: $.publicación
RELACIÓN SUPERIOR: $.publicación		
TIPO DE CONTENIDO: Nombre del archivo (File name)	TIPO DE CONTENIDO: Cadena de contenido (String content)	TIPO DE CONTENIDO: Cadena de contenido (String content)
ETIQUETA: imagen, rutaImagen, rutaXML	ETIQUETA: ciudad, estado, país, título	ETIQUETA: ciudad, título
DESCRIPCIÓN: tif, url.tif, url.xml (respectivamente)	DESCRIPCIÓN: cadena de caracteres	DESCRIPCIÓN: cadena de caracteres
VALOR: ninguno	VALOR: [país] nombre de país	VALOR: ninguno
RELACIÓN: $.publicación.página	RELACIÓN: $.publicación	RELACIÓN: $.publicación
RELACIÓN SUPERIOR: $.publicación		RELACIÓN SUPERIOR: $.publicación

TIPO DE CONTENIDO: Opciones múltiples predefinidas (multiple pre-defined choices)	TIPO DE CONTENIDO: Opciones múltiples predefinidas (multiple pre-defined choices)	TIPO DE CONTENIDO: VALORes numéricos (Numeric value)
ETIQUETA: estatus, frecuencia, idioma, tipoAcceso, tipoPublicación	ETIQUETA: estado, estatus, frecuencia, idioma, país, tipoAcceso, tipoPublicación	ETIQUETA: totalArchivosXML, totalImágenes, totalImágenesAnterior
DESCRIPCIÓN: valores predeterminados (sobre idiomas, verdadero/falso, periódico/monografía etc.)	DESCRIPCIÓN: valores predeterminados (sobre idiomas, verdadero/falso, periódico/monografía etc.)	DESCRIPCIÓN: valores numéricos
VALOR: ninguno	VALOR: ninguno	VALOR: ninguno
RELACIÓN: $.publicación	RELACIÓN: $.publicación	RELACIÓN: $.publicación
RELACIÓN SUPERIOR: $.publicación	RELACIÓN SUPERIOR: $.publicación.página	RELACIÓN SUPERIOR: $.publicación.página
TIPO DE CONTENIDO: Acrónimo/identificador único (unique ID/acronym)	TIPO DE CONTENIDO: VALORes numéricos (Numeric value)	TIPO DE CONTENIDO: Acrónimo/identificador único (unique ID/acronym)
ETIQUETA: _id, colecciones_id, primerPáginaDePublicación_id	ETIQUETA: página, totalPalabrasOCR	ETIQUETA: _id, páginaAnterior_id, páginaSiguiente_id, publicación_id
DESCRIPCIÓN: ID de objeto (Object_id)	DESCRIPCIÓN: valores numéricos	DESCRIPCIÓN: ID de objeto (Object_id)
VALOR: ninguno	VALOR: ninguno	VALOR: ninguno
RELACIÓN: $.publicación	RELACIÓN: $.publicación	RELACIÓN: $.publicación
	RELACIÓN SUPERIOR: $.publicación.página	RELACIÓN SUPERIOR: $.publicación.página

Conflicto de intereses

Los autores declaran que no existe conflicto de intereses.

Contribuciones de los autores

Las autoras participaron por igual en la investigación: conceptualización, diseño metodológico, escritura, análisis de encuestas, interpretación de resultados, conclusiones, recomendaciones, etcétera.