

Marco procedimental para facilitar la interoperabilidad en el contexto de la Biblioteca Virtual de Salud en Cuba: el modelo Ontomed

Procedural framework to facilitate interoperability in the context of the Virtual Health Library of Cuba: the Ontomed model

Keilyn Rodríguez Perojo,^I Amed Abel Leyva Mederos,^{II} José Antonio Senso Ruíz^{III}

^I Centro Nacional de Información de Ciencias Médicas. La Habana, Cuba.

^{II} Universidad Central de Las Villas "Martha Abreu". Villa Clara, Cuba.

^{III} Universidad de Granada, España.

RESUMEN

La inexistencia de un modelo de datos enlazados en el contexto de la Biblioteca Virtual en Salud de Cuba (BVS) provoca deficiencias en la interoperabilidad semántica de los contenidos y la recuperación de información. Además, y como problema añadido, las entidades documentales que lo componen presentan distintos niveles de descripción bibliográfica y de alcance temático. El presente artículo propone el diseño de un marco procedimental para facilitar la interoperabilidad semántica entre conjuntos de datos enlazados de la BVS, y estandarizar sus principales clases y propiedades. Este procedimiento, que se implementa en el modelo ontológico denominado Ontomed, constituye una aproximación teórico-conceptual que permitirá modelar las principales entidades, atributos y relaciones en el contexto de la BVS. Se espera un impacto directo en tres dimensiones fundamentales: 1) caracterización del contexto de aplicación del marco procedimental 2) formalización de los componentes del modelo ontológico y 3) selección de una muestra experimental en el escenario de la BVS.

Palabras clave: modelo de datos enlazados; datos enlazados; Ontomed; web semántica; interoperabilidad; estructura de datos; datos bibliográficos; bases de datos; recuperación de información; Biblioteca Virtual en Salud.

ABSTRACT

The absence of a linked data model in the context of the Virtual Health Library of Cuba (VHL) causes deficiencies in the semantic interoperability of content and information retrieval. In addition, as added problem, the documentary component entities have different levels of bibliographic description and subject scope. This article proposes the design of a procedural framework to facilitate semantic interoperability between linked datasets of VHL, standardizing their core classes and properties. Such a procedure, implemented in the ontological model called Ontomed, is a theoretical and conceptual approach that allows modeling the main entities, attributes and relationships in the context of the VHL. A direct impact will expect on three key dimensions: 1) characterization of context implementing the procedural framework 2) formalization of the components of the ontological model and 3) selecting an experimental sample on the stage of the VHL.

Key words: linked data model; linked data; Ontomed; semantic web; interoperability; data structure; bibliographic data; databases; information retrieval; virtual health library.

INTRODUCCIÓN

El crecimiento acelerado de los volúmenes de datos en el contexto virtual plantea la necesidad urgente de reformular los modelos de organización y representación de información bibliográfica desde una perspectiva integradora, creativa y centrada en dimensiones sociales, culturales, lingüísticas e históricas. Tal y como aconteció con el formato inicial de Elementos Semánticos de Europeana (*Europeana Semantics Elements (ESE)*), el principal problema de la mayoría de los esquemas de datos del universo bibliográfico para evolucionar a estructuras de datos más abiertas y flexibles es su enfoque generalizado hacia un modelo de descripción "plano", que no permitía incluir referencias a recursos externos, ni la extensibilidad a otros sistemas especializados de mayor granularidad.¹ De la misma manera, la Biblioteca del Congreso de los Estados Unidos, junto a un grupo de instituciones bibliotecarias, museos y archivos de reconocido prestigio internacional, han propuesto un nuevo marco conceptual para la descripción bibliográfica, conocido como *Library of Congress New Bibliographic Framework Initiative* o BIBFRAME,² con el objetivo de realinear el esquema bibliográfico vigente a partir de la adaptación de los Formatos MARC21 al modelo de la Web Semántica y Datos Abiertos y Enlazados (*Linked Open Data*). El propio reconocimiento internacional de las limitaciones de esquemas de datos bibliográficos tradicionales como MARC21, es un medio eficaz de tratar la información bibliográfica desde una perspectiva diferente, tomando como referencia la necesidad de representar objetos no documentales dentro de una Capa de Datos Semántica (*Semantic Data Layer*).

En este escenario, las nuevas normas de Recursos, Descripción y Acceso (*Resource Description and Access (RDA)*) juegan un rol importante en la manera en que se crean y se utilizan datos bibliográficos, en tanto enfatizan en las fortalezas de las Reglas de Catalogación Anglo Americanas (RCAA2),³ pero tienen características novedosas que lo hacen más aplicable al ambiente digital actual, incluyendo un alineamiento más cercano con modelos de datos internacionales como los

Requerimientos Funcionales de los Registros Bibliográficos (FRBR) y los Requerimientos Funcionales para Datos de Autoridad (FRAD).

En el contexto de la red Infomed conviven esquemas de datos heterogéneos que, en su mayoría, carecen de enfoque relacional explícito entre sus campos de datos homólogos. Esta heterogeneidad provoca la pérdida de datos de contexto en espacios de integración bibliográficos como la Biblioteca Virtual en Salud de Cuba (BVS), definida como una colección descentralizada y dinámica de fuentes y servicios de información especializados en Ciencias de la Salud, producidos nacional e internacionalmente y disponibles a través de internet.⁴ La herramienta LILACS, Descripción Bibliográfica e Indización para Web (LILDBI Web),⁵ emplea la Metodología LILACS en la generación de registros bibliográficos dentro del ecosistema de la Biblioteca Virtual, donde poco se puede captar el principio de procedencia y orden natural de los documentos en formato electrónico, así como las relaciones explícitas entre entidades, expresiones y manifestaciones en las interfaces de recuperación de información de la BVS.

LOS MODELOS DE DATOS ENLAZADOS EN CONTEXTOS BIBLIOGRÁFICOS

En el contexto bibliográfico, existen cuatro ideas fundamentales que han estado presentes, con mayor o menor fuerza, en los últimos años: las necesidades del usuario, el concepto de "obra", la estandarización y la internacionalización.⁶ Antes de ser desarrollados e introducidos los FRBR, la percepción que se tenía sobre los objetos bibliográficos era que éstos estaban constituidos por obras, textos, traducciones y ediciones.⁷ Posteriormente, *Le Bouef*⁸ ejemplifica la existencia de los objetos bibliográficos por medio de la analogía que hace del libro como objeto físico y abstracto, para llegar a definir qué se entiende por la entidad "obra" en el contexto de los FRBR. De ahí que las normas que regulan la catalogación se han nutrido de los aportes de distintas personas e instituciones que, en el afán de lograr una normalización para facilitar la identificación y el acceso a los documentos, han contribuido a la construcción de un corpus que sustenta el desarrollo alcanzado. Estas contribuciones han reconfigurado la noción de "Marcos de Referencia Bibliográficos" para el intercambio de datos estructurados, enfocado en el uso de la Web Semántica, los principios y mecanismos de Datos Enlazados (*Linked Data*) y el Marco de Descripción de Recursos (Resource Description Framework o RDF⁹) para el intercambio de datos estructurados. Entre los más relevantes, se encuentran:

BIBFRAME:² En noviembre del año 2012 la Biblioteca del Congreso anunció la publicación de un informe titulado *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services (BIBFRAME)*. Este se enfoca en el uso de la Web Semántica, los principios y mecanismos de *Linked Data* y *Resource Description Framework (RDF)* como marco de referencia básico para el intercambio de datos estructurados. La iniciativa tiene como objetivo la evolución hacia un nuevo esquema que deberá adaptarse mejor a las futuras necesidades para representar la información bibliográfica en la web, centrándose en cuatro clases principales: las obras de creación, instancias, las autoridades y anotaciones.¹⁰

Modelo Conceptual de Europea:¹ El Modelo de Datos de Europea (Europeana Data Model o EDM) es un vocabulario centrado en la representación de metadatos de objetos culturales, que proporciona acceso a las representaciones digitales de estos. EDM se sitúa en un contexto de agregación de datos, donde los objetos pueden ser complejos y en el que diferentes proveedores de datos pueden alojar diferentes visiones de estos. Ha sido incorporado por otros conjuntos de elementos, principalmente OAI-ORE (*Open Archives Initiative Object Reuse and Exchange*), Dublin Core, SKOS y CIDOC CRM.¹

*Modelo de Referencia Conceptual CIDOC CRM:*¹¹ Modelo semántico de referencia elaborado en el año 1994, primero por el Grupo de Normalización Documental (Documentation Standards Group) del Comité internacional para la documentación del Consejo internacional de los museos (ICOM-CIDOC), después por un grupo de trabajo especialmente constituido a dicho efecto, el CRM-SIG.¹¹ En el año 2006, ISO publicó el CIDOC CRM como norma internacional (ISO 21127:2006). Se trata de un modelo semántico que constituye una "ontología" de la información relativa al patrimonio cultural, y formaliza relaciones que unen los conceptos fundamentales de este tipo de información. Su presentación se basa en el enfoque "orientado al objeto". El modelo proporciona un lenguaje común a yacimientos de informaciones heterogéneas y permite la integración de estas, más allá de sus eventuales incompatibilidades tanto semánticas como estructurales.

Es importante tener en cuenta que estos modelos de datos enlazados se concretan en casos de uso representativos de colecciones de bibliotecas, archivos y museos. Precisamente el valor de estos casos radica en que el análisis de sus similitudes y diferencias permiten inferir buenas prácticas para la aplicación en otros proyectos del sector. Una de las diferencias entre los modelos anteriores y Ontomed es que no solo se concibe para contextos de datos patrimoniales, sino como punto de contacto para correlacionar datos administrativos de valor contextual como es caso del código REEUP (Registro Estatal de Empresas y Unidades Presupuestadas) con datos de afiliación institucional y geográfica. Otra de las diferencias fundamentales es la particular atención que se presta al enfoque multinivel de las clasificaciones temáticas en las Ciencias Biomédicas y de la Salud, y su articulación con vocabularios controlados como los Descriptores en Ciencias de la Salud (DeCS). Queda además el reto de transformar la estructura de éste último a datos enlazados para reconciliar términos y definiciones con vocabularios de alto nivel como los que integra el Unified Medical Language System (UMLS).¹²

PREMISAS PARA LA APERTURA DEL MODELO DE DATOS ENLAZADOS DE LA BIBLIOTECA VIRTUAL DE SALUD

*Tim Berners Lee*¹³ introdujo en su declaración sobre Linked Data los principios básicos de la vinculación de datos. Estos principios están orientados a la publicación e interconexión con el objetivo de facilitar la interoperabilidad, aprovechando tanto la arquitectura distribuida como los estándares Web:

- Utilizar URIs (*Uniform Resource Identifier*) para identificar los recursos en cada uno de los componentes de las declaraciones RDF: sujeto, predicado y objeto.
- Emplear URIs creados de acuerdo con el protocolo HTTP, de forma que puedan ser consultados y desreferenciados en la web por las personas y, sobre todo, por los sistemas automáticos.
- Debe proporcionarse información útil sobre los recursos identificados con los URI desreferenciables, para cuando alguien los consulte. Con esta finalidad, es necesario utilizar estándares de la web semántica, como RDF y SPARQL.
- Establecer enlaces con otros recursos (utilizando sus URI) en el momento de publicar datos en la web, de forma que se puedan descubrir más datos.

A estas premisas generales se añaden restricciones que se consideran importantes para homogeneizar la estructura de datos relativos a la BVS, tales como:

- Las clases responden a una visión conceptual del dominio Salud de Cuba como un sistema de relaciones, simples y complejas.

- El comportamiento de cada clase principal está dado por el entorno bibliográfico de la BVS, sus atributos y relaciones en el contexto de las Ciencias Biomédicas y de la Salud cubana.
- Todas las clases están provistas de objetos primitivos o abstractos que permiten modelar el contexto como un todo y parte a la vez.
- El sistema de relaciones de clases que mapean el dominio es susceptible de reutilizar ontologías que refuercen y armonicen el modelo ontológico.

El objetivo principal del artículo es exponer los elementos conceptuales y metodológicos relativos al modelo ontológico Ontomed, a partir de la aplicación de un procedimiento experimental para modelación de datos enlazados en el contexto bibliográfico de la BVS. Se pretende, además, mejorar la interoperabilidad semántica de sus recursos de información con otros conjuntos de datos estructurados.

MODELO DE DATOS ENLAZADOS DE LA BIBLIOTECA VIRTUAL DE SALUD

Por la complejidad y el alcance que supone el diseño de un ecosistema de datos enlazados en el contexto virtual de Infomed, no será objetivo esencial del presente artículo la profundización en la filosofía de datos abiertos,¹⁴ si bien los conjuntos de datos seleccionados para la transformación a este modelo, en el contexto de la Biblioteca Virtual de Salud en Cuba, pueden ser reutilizados por cualquier persona o aplicación de software, sujetos, cuando más, a requerimientos de atribución legal y de intercambio de la misma manera en que aparecen. Teniendo en cuenta estos principios, se propone, como primer paso, la transformación de una muestra representativa de fuentes de información de la BVS a datos enlazados, bajo las siguientes condiciones:¹⁵

- Definición, contexto y alcance de intervención del universo de datos de la BVS de Cuba.
- Definición del Modelo de Referencia de Datos Ontomed. El propósito de estandarizar la identificación, descripción, uso e intercambio de información entre diferentes niveles bibliográficos de la BVS de Cuba.
- Selección de una muestra representativa de recursos de información de la BVS de Cuba, susceptible de transformarse en conjuntos de Datos Enlazados (*Linked Data*).
- Definición de los metadatos primarios para normalizar la descripción de los conjuntos de datos.

METODOLOGÍA

En este apartado se desarrolla un caso de estudio a partir de la selección de una muestra representativa de fuentes de información, siguiendo el enfoque procedimental propuesto por *Hidalgo Delgado*.¹⁶ Este tiene como objetivo la validación del marco procedimental propuesto para la construcción del modelo de datos enlazados de la BVS, utilizando *Open Refine*¹⁷ (también conocido como *Google Refine*) en las tareas de depuración, organización, limpieza, homogeneización, transformación y enlazado de los datos provenientes de las fuentes seleccionadas (Fig. 1).



Fig. 1. Representación gráfica de las etapas para la aplicación del arco procedimental.

Los resultados obtenidos tendrán valor práctico en la implementación de aplicaciones de software en otras fases de la investigación en cuestión, y demuestran que la propuesta de solución incrementa la interoperabilidad semántica de los metadatos bibliográficos en el contexto de la BVS, atendiendo al número de consultas que pueden formularse utilizando un enfoque de búsqueda facetada.

Extracción de datos

El objetivo de esta actividad es extraer y almacenar los metadatos bibliográficos provenientes de de la BVS. Desde el punto de vista estructural, los metadatos utilizados por cada fuente siguen el estándar Dublin Core, con elementos de cualificación a partir de la representación de datos propios como el código REEUP en el caso del Directorio de Instituciones de Salud. La extracción de metadatos de las fuentes de información seleccionadas se realizó a partir de la exportación de sus bases de datos en MySQL a ficheros de texto separados por coma (Comma Separated Values o CSV), de forma semi automática. Una vez obtenidas las fuentes de datos, se procede a la importación en la herramienta Open Refine¹⁷ con el objetivo de normalizar las entidades que serán empleadas en el proceso de homogeneización con las 5 clases propuestas en el modelo de datos enlazados de la BVS:

- Personas (autores, editores, colaboradores, audiencias, profesores).
- Temáticas (clasificaciones, categorías).
- Documentos (libros, revistas, artículos, tesis, directorios, sitio web).
- Eventos (marco temporal por años, acontecimiento, tiempo de duración de una obra).
- Instituciones (nombres de instituciones del Sistema Nacional de Salud).

Procesamiento previo de las fuentes de datos

El objetivo de esta actividad es normalizar algunos campos de metadatos a partir de la transformación de datos para su normalización, tales como: fechas, afiliaciones institucionales, nombre de autores, clasificaciones, palabras clave, entre otros. El procesamiento previo de los datos se realiza con la finalidad de limpiar los metadatos obtenidos en la actividad anterior, a partir de la base de datos intermedia obtenida en la extracción de datos. La salida de esta actividad es la misma base de datos bajo las siguientes condiciones:

- Desambiguación de los nombres de los autores, eliminando espacios en blanco entre caracteres, así como cadenas de caracteres no relacionadas por adición o sustracción (Fig. 2).
- Normalización de las entradas de autores bajo una forma autorizada, según las normas Vancouver.

Cluster & Edit column "Author"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method: Keying Function: 4 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none"> • Figeroa, Rodrigo A. (1 rows) • Figueroa, Rodrigo A (1 rows) 	<input checked="" type="checkbox"/>	Figeroa RA
2	6	<ul style="list-style-type: none"> • Díaz Mastellari, Marcos (5 rows) • Díaz Mastellari, Marcos, Comp. (1 rows) 	<input checked="" type="checkbox"/>	Díaz M
2	3	<ul style="list-style-type: none"> • Retamales Saavedra Rodrigo (2 rows) • Retamales Saavedra, Rodrigo (1 rows) 	<input checked="" type="checkbox"/>	Retamales R
2	2	<ul style="list-style-type: none"> • García Artiles, Manuel (1 rows) • García Artiles, Manuel B. (1 rows) 	<input checked="" type="checkbox"/>	García M

Rows in Cluster

Average Length of Choices

Length Variance of Choices

0 — 6

19 — 26.5

0 — 3.5

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Fig. 2. Desambiguación de los nombres de autores procedentes del Localizador de Información en Salud, utilizando el algoritmo *key collision-metaphone*.

Para solucionar el problema de la ambigüedad relativa a los registros de los autores, se utilizaron los algoritmos integrados a *Open Refine* conocidos como *key collision-fingerprint* y *key collision-metaphone* con el objetivo de normalizar las cadenas de textos a minúsculas, las transformaciones a caracteres ASCII, la eliminación espacios y símbolos de puntuación, así como el empleo de funciones de comparación fonética.

Modelación

Desde el punto de vista técnico, lo más importante son las especificaciones que conducen al diseño de un nuevo Modelo de Datos Enlazados para la BVS, a partir de la reutilización de diferentes ontologías, más las clases definidas para Ontomed. Entre ellas, la definición de la estructura de objetos digitales y agregaciones de estos proviene del estándar *The Open Archive Initiative-Object Reuse and Exchange*.¹⁸ Los conceptos se definen a partir de *The Simple Knowledge Organization System (SKOS)*,¹⁹ mientras que Dublin Core complementa la definición de clases y propiedades (cuadro 1).²⁰⁻²⁴

Cuadro 1. Ontologías escogidas para el modelado de Ontomed

BIBO	La ontología bibliográfica (BIBO por sus siglas en inglés) es una ontología para la Web Semántica utilizada para describir entidades bibliográficas como libros, revistas, tesis, documentos legales, entre otros.
FOAF (Friend of a Friend)	Es una ontología ampliamente utilizada en la descripción de personas, de las relaciones con otras personas y con los recursos Web.
VIVO Core	VIVO articula un grupo de ontologías con información acerca de investigadores, técnicos u otros individuos que les permite destacar sus áreas de experiencia, desplegar credenciales académicas, visualizar sus redes de trabajo y mostrar información sobre publicaciones, proyectos, servicios.
Dublin Core	La Iniciativa de Metadatos Dublin Core (DCMI) es un esfuerzo internacional e interdisciplinar dirigido a definir un conjunto básico de elementos para la recuperación de información en Internet. Está constituido por quince elementos de datos que pueden usarse para describir una amplia variedad de objetos de información en Internet.
SKOS	Simple Knowledge Organization System (Sistema Simple de Organización del Conocimiento) proporciona un modelo para la representación de la estructura básica y el contenido de esquemas de conceptos como tesauros, esquemas de clasificación, listas de encabezamientos de materia, taxonomías, folksonomías y otros vocabularios controlados similares.

La intención fundamental es centrar la recuperación de información en los siguientes elementos del contexto: personas, eventos, materias, instituciones, recursos de información o las combinaciones de objetos digitales más complejos en forma de colecciones o agrupaciones funcionales, siguiendo el método analítico de responder a las siguientes preguntas: Quién, Qué, Cómo, Cuándo, Dónde y Porqué (Fig. 3).



Fig. 3. Vista del modelo conceptual de datos enlazados de la Biblioteca Virtual de Salud

Estructura de clases e instancias del modelo

La mayor parte de los esfuerzos durante las últimas décadas se han dedicado al desarrollo de algoritmos de razonamiento sobre la estructura de las ontologías, lo que se denomina en lógica de descripciones la TBox.²⁵ Sin embargo, las aplicaciones en la Web Semántica necesitan un método de recuperación de instancias y consultas que, además de permitir recuperar el conocimiento, sea capaz de inferir información no sólo sobre la estructura de la ontología, sino también sobre las instancias, lo que en lógica de descripciones se denomina ABox.²⁶ En las Lógicas Descriptivas (DL) existe una distinción entre la llamada TBox (caja terminológica) y la ABox (caja de aserciones). De forma general, la TBox contiene sentencias que describen conceptos jerárquicos o relaciones entre conceptos, mientras la ABox contiene sentencias específicas indicando a donde pertenecen los individuos en la jerarquía, es decir, relaciones entre individuos y conceptos.²⁷ La unión de TBox y ABox deriva en la formación de una Base de Conocimientos, equivalente a un conjunto de axiomas relativos a la Lógica de Primer Orden,²⁸ donde es posible definir cálculos de inferencia que permiten derivar conocimiento implícito a partir del explícito.

El objetivo principal es la definición de la estructura de datos y patrones de razonamiento, tanto de la TBox como la ABox de Ontomed, que se integrará al Modelo de Datos Enlazados de la BVS, resultante del proceso de homogeneización de las fuentes de datos seleccionadas con las ontologías externas.

ESTRUCTURA DE DATOS Y PATRONES DE RAZONAMIENTO DE ONTOMED

TBox de Ontomed

El TBox contempla las clases principales de Ontomed:

- Clases principales de "Ontomed".
- Autor:= Authority.
- Profesor:= Person.
- Legislación:= Resource.
- Artículo:= Document.
- Hospital:= Organization.
- Colección de Enfermería:= Collection.
- Especialidades Médicas:= Classification.
- Documental:= Work.

Donde:

":=" simboliza que el primer elemento "está definido en la clase" correspondiente.

Para Ontomed, la representación de la clase "Persona" y sus posibles relaciones con otras clases a través de sus propiedades, sería la siguiente forma:

foaf:Person

- owl:Class.
- rdfs:subClassOf owl:Thing.
- rdfs:label"Person"@en.

Property:preferredName

- owl:DatatypeProperty.
- rdfs:subPropertyOf skos:prefLabel, foaf:name, rdfs:label ; vivo:Authorship.
- rdfs:domain foaf:Person.
- rdfs:label "preferred name"@en.

ABox de Ontomed

El ABox contiene afirmaciones acerca de individuos nombrados en términos de vocabulario:

- Dr. Gustavo Kourí: Authority.
- Pedro Kourí: Person

- Resolución Ministerial No 286/2014: Resource.
- Atlas de embriología humana: Document.
- Hospital Clínico Quirúrgico Docente Dr. "Salvador Allende": Organization.
- Colección de Enfermería: Collection.
- Hematología: Classification.
- Analgesia acupuntural quirúrgica ¿Falacia o realidad?: Work.

Donde:

":" simboliza el rol que juega la instancia con respecto a la clase correspondiente.

Para Ontomed, la instanciación de la clase "Persona" y sus posibles relaciones con otras clases a través de sus propiedades, sería la siguiente forma:

- <#pedrokouri>.
- foaf:Person.
- preferredName "Pedro Kourí".

Para la construcción de la ontología de dominio derivada de la TBox y ABox de Ontomed, se utilizó la metodología *Methontology*²⁹ como parte del ciclo de vida basado en la reutilización de otras existentes.

Enlazado

La generación de enlaces entre el grafo RDF generado en la actividad anterior y grafos similares publicados en la web es una de las tareas fundamentales para el enriquecimiento de recursos publicados como datos enlazados. Para generar los enlaces se utiliza la extensión para Open Refine nombrada "Name-Entity-Recognition"³⁰ para facilitar tareas de reconciliación y normalización de enlaces con fuentes de datos externas. En el caso de estudio se generaron enlaces entre los autores, instituciones y temáticas asociadas. Para establecer la comparación entre los nombres de los autores entre ambos grafos, origen y destino, se utilizó el algoritmo *key collision-fingerprint* con un umbral de distancia de 2, lo que significa que solo se enlazaran aquellos autores que contengan exactamente el mismo nombre en ambos grafos RDF.

Publicación

La publicación de datos enlazados significa la creación de conjuntos de datos (data sets) en el contexto de una o más entidades, donde se agregan y mantienen las declaraciones RDF, formando un grafo global, sin límites de registros (bibliográficos y de autoridades) y catálogos locales donde se generaron inicialmente. De esta forma, estarán listos para:

- Enlazarse, en origen, con datos RDF de otros depósitos digitales e incluso de otros conjuntos de datos ajenos a las colecciones documentales.
 - Ser enlazados por conjuntos de datos externos que quieran ser enriquecidos.
-

- Ser consumidos directamente por aplicaciones que los necesiten para la generación de nuevos recursos y servicios, mediante descargas masivas (dump) o mediante consultas selectivas de datos a través de puntos de acceso SPARQL (SPARQL Endpoint).

RESULTADOS

CASO DE ESTUDIO

El caso de estudio tiene como objetivo medir la aplicabilidad del marco de trabajo propuesto. Se resumen las actividades relativas a la selección, modelado, transformación y enlazado de datos enlazados a partir de cuatro fuentes de datos de la BVS.

Selección de las fuentes de datos

Para la selección, se tuvieron en cuenta dos criterios esenciales (tabla):

- La intencionalidad de la BVS en cuanto a la cobertura temática de contenidos relacionados a los diez principales problemas de salud de Cuba.
- Los niveles de uso de estas fuentes de información a partir del análisis de series de datos temporales provenientes de herramientas de analítica web.

Modelado, transformación y enlazado de datos

En esta actividad se modelan los metadatos bibliográficos extraídos y procesados con clases y propiedades propias de las ontologías especificadas en la tabla. Se analizaron las ontologías existentes que presentan componentes susceptibles a modelar en el contexto de los bibliográficos de la BVS. Una vez ejecutado procesamiento previo de los metadatos, se definen las clases y propiedades del modelo ontológico para compartir y anotar semánticamente los metadatos procesables, tanto por los humanos como por las computadoras (cuadro 2). El objetivo final de la actividad es crear enlaces semánticos procesables por máquina para conocer las relaciones y enlazarlas con entidades provenientes de fuentes de datos externas (Fig. 4).

Tabla. Fuentes de datos empleadas en el caso de estudio, que se encuentran entre las diez más importantes de la Biblioteca Virtual en Salud de Cuba

Fuente de datos	URI	Número de registros
Localizador de Información de Salud	http://liscuba.sld.cu/	8 832
Directorio de Instituciones de Salud	http://dirinstituciones.sld.cu/	4 837
Legislación en Salud	http://legislacion.sld.cu/	513
Catálogo de Obras Audiovisuales	http://obrasaudiovisuales.sld.cu/	288
Total		14 217

Cuadro 2. Clases y propiedades de las ontologías reutilizadas para la construcción del modelo ontológico

Entidad	Clase	Metadatos asociados	Propiedades
Libro	bibo: Book vivo: Book vivo: Document	Título	dc: title
		Descriptores	bibo: subject dc: subject
		Editorial	bibo: publisher vivo: publisherName
		Correo electrónico	foaf: mbox
Autor	foaf: Person vivo: authorship dc: creator	Autor Creador Fuente	bibo: authorList vivo: isCorrespondingAuthor
Temática	dc:subject vivo:SubjectArea	Temática Categoría Clasificación	vivo:hasSubjectArea
Afiliación	bibo: Organization vivo: Organization	Institución	vivo: hasAffiliate

Como no existe un catálogo integrado de la BVS, en el siguiente caso de uso se modela un grafo conceptual con un patrón de secuencia de interrelaciones a partir de la propiedad "sameAs", perteneciente al destacado médico cubano "Pedro Kourí, 1900-1964", disponible como institución científica dedicada a la investigación en el actual Localizador de Información en Salud de Cuba ([Fig. 5](#)):

CONCLUSIONES

La concepción de un Modelo de Datos Enlazados para el dominio de las Ciencias Biomédicas y de la Salud cubano, especialmente en el contexto bibliográfico de la BVS, es un proceso complejo, ya que uno de los enfoques más adecuados es contar con expertos que proporcionen herramientas que se adapten, al menos, a los requerimientos básicos de expresión del vocabulario, no solo respecto a la limitada expresividad semántica, sino también en relación con la gestión de vínculos, descripción de metadatos o posibilidades de ofrecer una estrategia válida de preservación. Además, las mínimas condiciones de fiabilidad y calidad obligan a emplear medios semiautomáticos para efectuar los modelados, pues la alternativa manual no es eficiente e introduce la posibilidad del error en porcentajes más altos que los deseables.

RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: <http://localhost:3333/> [edit](#)

RDF Skeleton

RDF Preview

This is a sample turtle representation of (up-to) the *first 10* rows

```
@prefix gn: <http://www.geonames.org/ontology#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix vivo: <http://vivoweb.org/ontology/core#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix mesh: <http://org.snu.bike/MeSH#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

<http://localhost:3333/275> a bibo:Webpage ;
  dc:title "Sexología y Sociedad" ;
  dcterms:created "2005-04-05 00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  dc:subject "http://dbpedia.org/ontology/genre" ;
  foaf:mbox "cenesex@infomed.sld.cu" ;
  vivo:hasSubjectArea "Especialidades -- Sexología" ;
  dc:description "
  Revista especializada del Centro Nacional de Educación Sexual (CENESEX), bajo el auspicio del Gobierno Cubano y del Fondo de Población de Na
  " .

<http://localhost:3333/7387> a bibo:Webpage ;
  dc:title "Sexual development" ;
  dcterms:created "2010-10-21 11:16:00"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
  dc:subject "http://dbpedia.org/ontology/genre" ;
  vivo:hasSubjectArea "Especialidades -- Sexología" ;
  dc:description "Revista que publica artículos originales y especializados en las diferentes temáticas de sexualidad que permite un acer
```

Fig. 4. Reconciliación y enlazado entre el grafo RDF obtenido y DBpedia, a partir de la propiedad "subject".

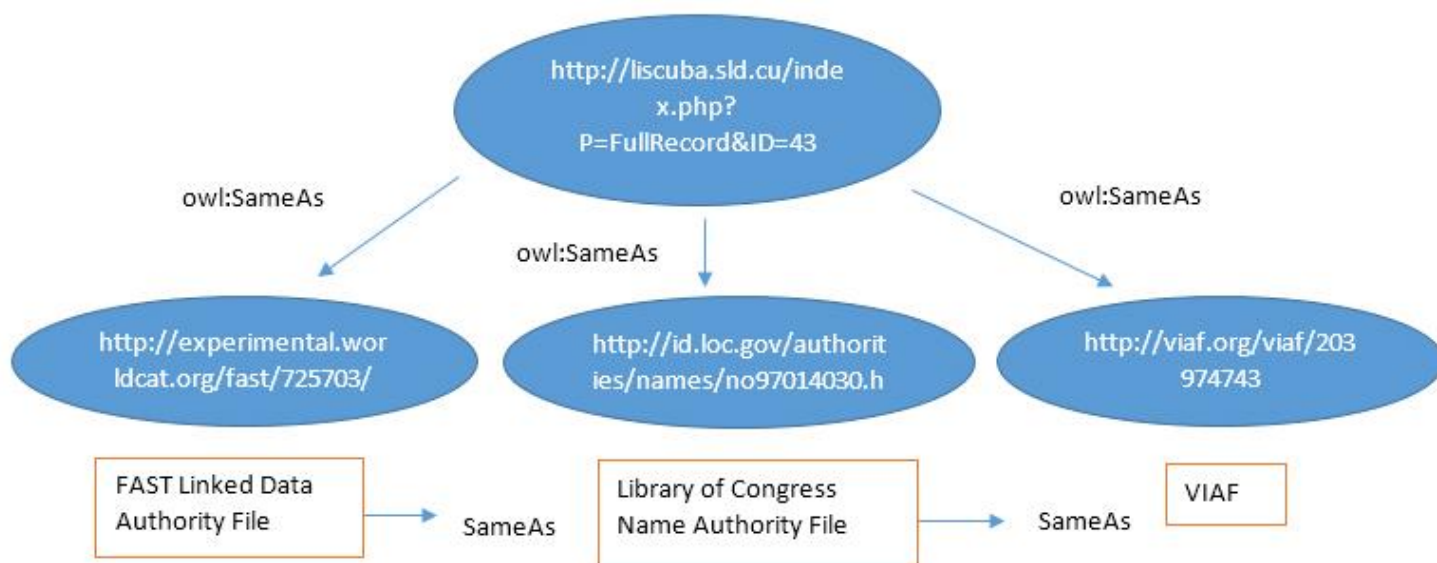


Fig. 5. Representación del proceso de enlazado de clase genérica "Autor", a partir de las posibles relaciones entre conjuntos de datos internos y externos.

A pesar de los obstáculos y los resultados obtenidos, se concluye que la aplicación del marco procedimental posibilitó constatar el poco desarrollo de la aplicación de tecnologías de la web semántica en el contexto bibliográfico de la BVS. Las colecciones digitales disponibles necesitan de un esfuerzo institucional intencionado para enriquecer sus datos, usando formatos de serialización en RDF con enlaces externos y creando puntos de consulta SPARQL.

El caso de estudio demuestra la falta de normalización en los principales metadatos que comparten las fuentes de datos seleccionadas. Este aspecto influye sobremanera en los niveles de agregación de la propia colección o del conjunto de las colecciones de una institución que aporta contenidos a la BVS. Sería conveniente considerar niveles superiores, como los agregadores nacionales (Biblioteca Nacional de Cuba) o internacionales como Europea³¹ y la Biblioteca Pública Digital de las Américas.³² De esta forma, aumentaría la proyección y la difusión de la información contenida en estas colecciones.

La propuesta de diseño de Ontomed ha permitido capturar una porción del dominio de la Salud cubano desde una perspectiva relacional e intencional cualitativamente superior, en tanto todas y cada una de las fuentes de datos seleccionadas en la muestra poblacional no se relacionan más allá de su propia estructura de Base de Datos. Por tanto, la BVS de Cuba es una representación virtual que integra parte de la producción documental y científica de la Salud cubana, pero no toda.

En el diseño de las clases principales del modelo se presta especial atención a la distinción de puntos de acceso vitales como las Autoridades, las Clasificaciones y la noción "Obra" como entidades que proporcionan mayor expresividad al modelo, dado que la simplicidad de los componentes de la web ha sido el substrato de su éxito. A juicio personal del autor de la presente investigación, simplificar la gestión semántica de vocabularios supondría la generalización de la publicación de vocabularios bibliotecarios de calidad como datos enlazados.

CONTRIBUCIÓN DE LOS AUTORES

Keilyn Rodríguez Perojo, José Antonio Senso Ruíz y Amed Abel Leiva Mederos contribuyeron por igual al diseño del estudio y a la redacción del artículo. Todos los autores revisaron y aprobaron la versión final.

CONFLICTO DE INTERESES

Los autores declaran que no existe conflicto de intereses.

REFERENCIAS BIBLIOGRÁFICAS

1. Doerr M, Gradmann S, Hennicke S, Isaac A, Meghini C, van de Sompel H. The european data model. In: World Library and Information Congress: 76th IFLA General Conference and Assembly; 2010. pp. 10-5.
2. Miller E, Ogbuji U, Mueller V, MacDougall K. BIBFRAME Primer: Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services. Library of Congress; 2012:[42 p.].
3. Rovira Jarque A, Bonsón À. Les noves normes per a un catàleg. RDA: Documento; 2015.
4. Packer AL, de Castro E. Biblioteca Virtual en Salud: BIREME. Sao Paulo: El Centro; 1998.
5. Santana-Arroyo S. Comparación de campos de LILACS Descripción Bibliográfica e Indización (LILDBI) con entidades y atributos de los requerimientos funcionales para registros bibliográficos. Rev Esp Docum Cient. 2013;36(2):6.
6. García R, Alejandro A. Los objetos bibliográficos confirmados en la integración compleja de la descripción y acceso a recursos. Investig Bibliotecol. 2009;23(48):33-59.
7. Wilson P, Svenonius E. The Intellectual Foundations of Information Organization. Cambridge: Digital Libraries and Electronic Publishing; 2001;62(2):203-4.
8. Le Boeuf P. Principios de catalogación IFLA: Hacia un código internacional de catalogación; 2005:2.
9. Andreu S, Miquel C, Jorge F, Jesús G. Aplicación del modelo de datos RDF en las colecciones digitales de bibliotecas, archivos y museos de España. Rev Esp Docum Cient. 2016;39(1):5.
10. Miller E, Ogbuji U, Mueller V, MacDougall K. Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services. Washington, DC: Library of Congress; 2012 [citado 20 de mayo de 2013]. Disponible en: <http://www.loc.gov/bibframe/pdf/marclid-report-11-21-2012.pdf>
11. Doerr M, LeBoeuf P. Modelling intellectual processes: the FRBR-CRM harmonization. Digital libraries: Research and development: Springer; 2007. p. 114-23.
12. Bodenreider O. The unified medical language system: integrating biomedical terminology. Nucleic acids research. 2004;32:267-70.

13. Bizer C, Heath T, Berners-Lee T. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*. 2009:205-27.
14. Monino JL, Sedkaoui S. *Big Data, Open Data and Data Development*: John Wiley & Sons; 2016.
15. Solar M, Meijueiro L, Daniels F. *A Guide to Implement Open Data in Public Agencies*. Springer; 2013. p. 75-86.
16. Delgado YH, Álvarez LR, Mederos AL, García MdMR, Montes JFA. BM2LOD: platform for publishing bibliographic data as linked open data. *7th IADIS International Conference Information Systems*; 2014.
17. Verborgh R, De Wilde M, Sawant A. *Using OpenRefine: the essential OpenRefine guide that takes you from data analysis and error fixing to linking your dataset to the Web*; 2013.
18. Lagoze C, Van de Sompel H, Johnston P, Nelson M, Sanderson R, Warner S. *Open archives initiative object reuse and exchange*. Presentation at JCDL; 2007.
19. Miles A, Matthews B, Wilson M, Brickley D, editors. *SKOS core: simple knowledge organisation for the web*. *International Conference on Dublin Core and Metadata Applications*. 2005. pp. 3.
20. Dimic Surla B, Segedinac M, Ivanovic D, editors. *A BIBO ontology extension for evaluation of scientific research results*. In: *Proceedings of the Fifth Balkan Conference in Informatics*. Association for Computing Machinery (ACM); 2012. pp. 275-8.
21. Brickley D, Miller L. *FOAF vocabulary specification 0.98*. Namespace document. 2012:9.
22. Krafft DB, Cappadona NA, Caruso B, Corson-Rikert J, Devare M, Lowe BJ, et al. *Vivo: Enabling national networking of scientists*. 2010.
23. Weibel SL, Koch T. *The Dublin core metadata initiative*. *D-lib magazine*. 2000;6(12):1082-9873.
24. Cox SJD, Yu J, Rankine T. *SISSVoc: A Linked Data API for access to SKOS vocabularies*. *Semantic Web*. 2016;7(1):9-24.
25. Schaerf A. *Reasoning with individuals in concept languages*. *Data & Knowledge Engineering Data & Knowledge Engineering*. 1994;13(2):141-76.
26. Xu J, Shironoshita P, Visser U, John N, Kabuka M. *Module Extraction for Efficient Object Queries over Ontologies with Large ABoxes*. *AIA Artificial Intelligence and Applications*. 2015;2015(1):8-31.
27. Villazón-Terrazas BM. *A method for reusing and re-engineering non-ontological resources for building ontologies*: IOS Press; 2012.
28. Baader F. *The description logic handbook theory, implementation and applications*. Cambridge, UK; New York: Cambridge University Press; 2003.
29. Fernández-López M, Gómez-Pérez A, Juristo N. *Methontology: from ontological art towards ontological engineering*. Madrid: Universidad Politécnica de Madrid; 1997.
30. Van Hooland S, De Wilde M, Verborgh R, Steiner T, Van de Walle R. *Exploring entity*

recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities*. 2015;30(2):262-79.

31. Portal Europeana. Modelo de datos Europeana; 2015 [citado 20 de mayo de 2016]. Disponible en: <http://www.europeana.eu/portal/>

32. Library of Congress. Digital public library of America. 2015 [citado 20 de mayo de 2016]. Disponible en: <https://www.dp.la/>

Recibido: 2 de junio de 2016.

Aprobado: 2 de julio de 2016.

Keilyn Rodríguez Perojo. Centro Nacional de Información de Ciencias Médicas. Calle 27 No. entre N y M. El Vedado.
Correo electrónico: keilyn@infomed.sld.cu