LA INVESTIGACIÓN BIBLIOTECOLÓGICA Y DE LA INFORMACIÓN HACIA EL 2030: DESARROLLO SOSTENIBLE

Catalina Naumis Peña
Ariel Alejandro Rodríguez García
Coordinadores

Z669.7
I58

La investigación bibliotecológica y de la información hacia el 2030 : desarrollo sostenible / Coordinadores Catalina Naumis Peña, Ariel Alejandro Rodríguez García. - México : UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2022.
xiv, 322 p. - (Sistemas bibliotecarios de información y sociedad)
ISBN: 978-607-30-6258-9

1. Investigación bibliotecológica. 2. Objetivos de Desarrollo Sostenible. 3. Bibliotecas - Desarrollo sustentable. 4. Desarrollo sustentable - Aspectos sociales. I. Naumis Peña, Catalina, coordinadora. II. Rodríguez García, Ariel Alejandro, coordinador. III. ser.

Diseño de la portada: Wendy Chávez
Primera edición: julio de 2022

Instituto de Investigaciones Bibliotecológicas y de la Información
Circuito Interior s/n, Torre II de Humanidades,
pisos 11, 12 y 13, Ciudad Universitaria, C. P. 04510,
Alcaldía Coyoacán, Ciudad de México

ISBN: 978-607-30-6258-9

Publicación dictaminada

Impreso y hecho en México

Contenido

Información e investigación en desarrollo sostenible

El covid-19, análisis de un dominio emergente

Posibilidades del XML JATS para el tratamiento y la recuperación de información: El caso del sistema de indización automática SISA

Isidoro Gil Leyva
Facultad de Comunicación y Documentación, Universidad de Murcia

Introducción

De la comunicación científica tradicional a la publicación semántica

Hasta mediados del s. XVII, la principal vía de comunicación de los logros y descubrimientos científicos era la correspondencia privada que propiciaba un avance del conocimiento muy lento. En 1665, aparecieron las dos primeras revistas científicas: la francesa ***Journal des Sçavans*** y la británica ***Philosphical Transactions,*** y desde entonces hasta la actualidad, las revistas científicas se han ido consolidando como el principal canal de comunicación de la ciencia por encima de libros, informes, encuentros académicos o cualquier otra forma de difusión.

Desde la creación de estas revistas, los resultados de la investigación científica se difunden por medio del artículo científico como eje central de la comunicación científica. Desde 1665 hasta 1950 aproximadamente, no se produjeron cambios significativos en el flujo de trabajo de la edición y difusión de la ciencia, más allá de las innovaciones progresivas de impresión o el uso de máquinas de escribir a partir de 1830. A mitad del siglo XX registramos la invención de los ordenadores y poco después su uso generalizado, a lo que habría que sumar después la aparición de

software de edición de textos como Wordstar (1978), la Web (1989) y nuevos formatos para presentar la información como el HTML (1991) o el PDF (1993). A continuación, a finales de la década de 1990 y comienzos de 2000 se crearon y extendieron herramientas y lenguajes (eXtensible Markup Language, RDF, Sparql, Ontology Web Language) encaminados a crear una Web dotada de significado con datos legibles por aplicaciones informáticas. Esta combinación tecnológica hará que se vaya abandonando una web basada en documentos aislados en favor de una nueva Web (semántica) conformada por documentos legibles por humanos, por máquinas, procesables e interconectados, permitiendo, como señalaron los inventores de la Web semántica Berners-Lee, Hendler y Ora,[1] que las computadoras y las personas trabajen en cooperación.

A esta irrupción tecnológica, hay que sumar a partir de 2002 el movimiento de Ciencia abierta que persigue un cambio en el modelo de comunicación científica. Durante la década de 1990, comenzaron a publicarse las primeras revistas electrónicas que, en mayor o menor medida, se han ido beneficiando de este florecimiento tecnológico. En un primer estadio, a comienzos de la década de 1990 las revistas científicas aprovecharon un entorno Web con formatos como el HTML y PDF, empleados extensivamente por los editores para el proceso último de comunicación que facilitaba la difusión y el intercambio de los artículos científicos. De hecho, hoy en día, según recientes datos del informe Scholastica,[2] plataforma web de pago desde la que más de novecientos editores llevan a cabo la gestión integral de sus revistas académicas, los formatos más empleados siguen siendo todavía el PDF (98%), HTML (48%), papel (43%), EPUB (14%), XML (3%). Si bien, la publicación en XML ya comienza a estar presente y muchos de esos artículos publicados en formato PDF, HTML O EPUB se generan a partir de un documento base en XML JATS.

En un segundo estadio, las revistas científicas al inicio de la década de 2000, tomaron un nuevo impulso a raíz del movimiento

1 Berners-Lee, Hendler y Ora. "The Semantic Web", 3.

2 The State of Journal Production and Access 2020.

de Ciencia abierta y la Web semántica con la tecnología XML, RDF, OWL a lo que hay que sumar desde 2012 el estándar Journal Article Tag Suite (JATS) para describir el contenido textual y gráfico de los artículos. Según datos del ya mencionado informe Scholastica de 2020, el 35 por ciento de los editores ya están usando el formato XML principalmente JATS. Así pues, esta combinación de Ciencia abierta y Web semántica ha dado lugar, en la última década, a la denominada publicación semántica que termina coadyuvando el procesamiento automático, recuperación, difusión, intercambio y reutilización de información.

Publicación semántica

Durante trescientos años, se ha venido usando un formato lineal para el artículo científico; sin embargo, con la entrada de la publicación electrónica se abrieron nuevas posibilidades. En este sentido, Joost G. Kircz[3] propuso la ruptura estructural lineal de los artículos científicos mediante la separación y el almacenamiento de distintos tipos de información en módulos textuales cognitivos bien definidos, vinculados y adaptados a las necesidades de los lectores, lo que propició además la reutilización de la información.

La publicación semántica se sustenta en documentos electrónicos que además de ser legibles por humanos, contienen metadatos sobre su estructura y contenido legibles por máquinas. En 2009, Shotton[4] definió la publicación semántica "[...] como cualquier cosa que mejore el significado de un artículo de revista publicado, facilite su descubrimiento automatizado, permita su vinculación a artículos relacionados semánticamente, permita el procesamiento de los datos del artículo o facilite la integración de datos entre artículos". Así pues, la publicación semántica que aquí nos interesa se fundamenta en el uso de formatos legibles por máquina que permitan su procesamiento para diferentes fines: extracción, anotación y relación de información. La extracción de información

3 Kircz, "Modularity".

4 Shotton, "Semantic", 86.

para la condensación, indización o clasificación; la anotación para facilitar la lectura y búsqueda de información, y la conexión entre entidades presentes en el contenido de cada pieza permite la relación entre diferentes documentos generando redes de conocimiento.

Desde finales de la década de 2000 hasta la actualidad, se han realizado numerosas propuestas que están conformando lo que se entiende por publicación semántica, algunas de las cuales ya han llegado a los editores que son los que en última instancia las materializan. A continuación, se presentan varias propuestas de mejora semántica agrupadas en tres grados de complejidad para su puesta en práctica:

Complejidad baja. La estructura lineal del artículo se rompe en favor de unidades bien reconocibles y navegables apreciables normalmente en la parte superior o lateral, como por ejemplo, introducción, secciones, conclusiones, referencias o cualquier otra; incorporación de resúmenes gráficos y en video a los tradicionales

Figura 1. Elementos configuradores de la publicación semántica

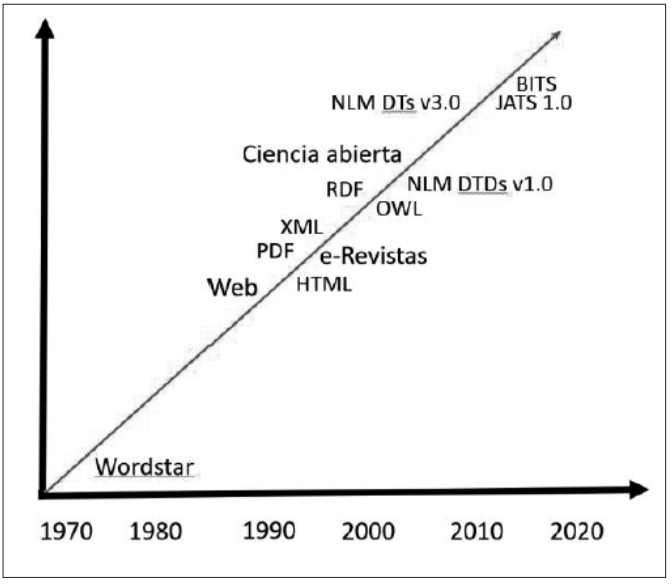
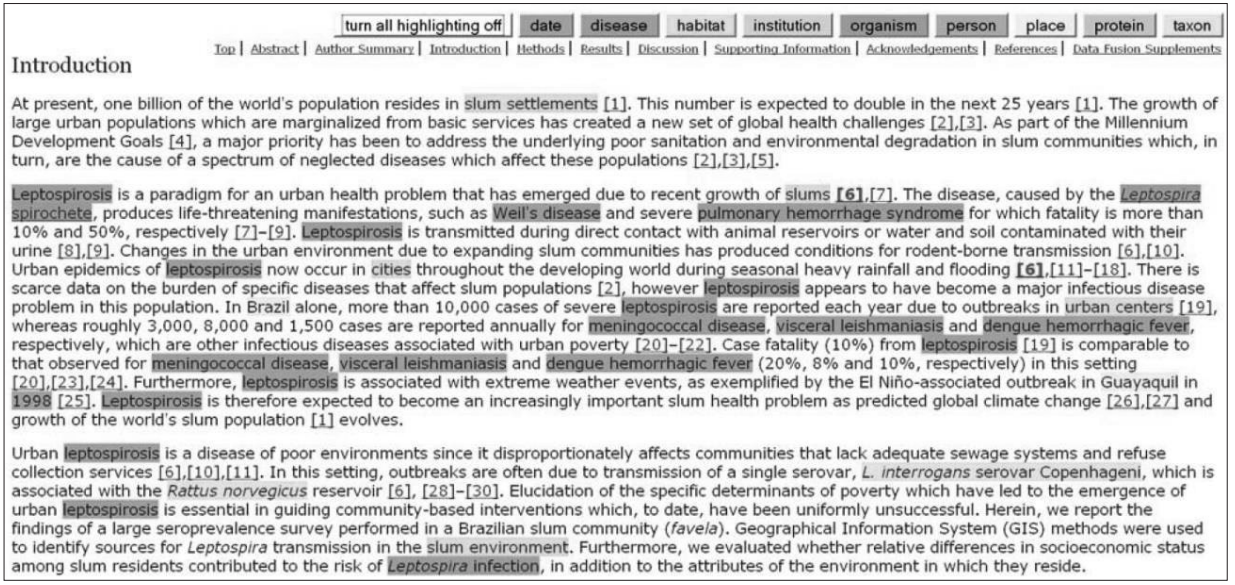


Figura 2. Anotación automática de artículo



turn all highlighting off | date | disease | habitat | institution | organism | person | place | protein | taxon

Top | Abstract | Author Summary | Introduction | Methods | Results | Discussion | Supporting Information | Acknowledgements | References | Data Fusion Supplements

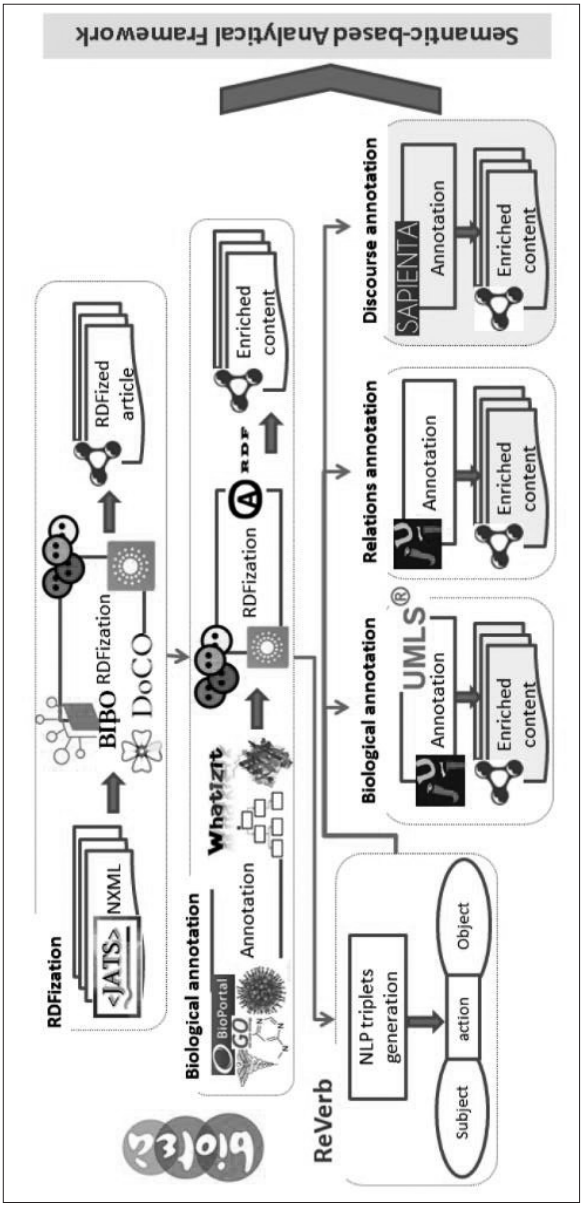
Introduction

At present, one billion of the world's population resides in slum settlements [1]. This number is expected to double in the next 25 years [1]. The growth of large urban populations which are marginalized from basic services has created a new set of global health challenges [2],[3]. As part of the Millennium Development Goals [4], a major priority has been to address the underlying poor sanitation and environmental degradation in slum communities which, in turn, are the cause of a spectrum of neglected diseases which affect these populations [2],[3],[5].

Leptospirosis is a paradigm for an urban health problem that has emerged due to recent growth of slums **[6]**,[7]. The disease, caused by the *Leptospira* spirochete, produces life-threatening manifestations, such as Weil's disease and severe pulmonary hemorrhage syndrome for which fatality is more than 10% and 50%, respectively [7]–[9]. Leptospirosis is transmitted during direct contact with animal reservoirs or water and soil contaminated with their urine [8],[9]. Changes in the urban environment due to expanding slum communities has produced conditions for rodent-borne transmission [6],[10]. Urban epidemics of leptospirosis now occur in cities throughout the developing world during seasonal heavy rainfall and flooding **[6]**,[11]–[18]. There is scarce data on the burden of specific diseases that affect slum populations [2], however leptospirosis appears to have become a major infectious disease problem in this population. In Brazil alone, more than 10,000 cases of severe leptospirosis are reported each year due to outbreaks in urban centers [19], whereas roughly 3,000, 8,000 and 1,500 cases are reported annually for meningococcal disease, visceral leishmaniasis and dengue hemorrhagic fever, respectively, which are other infectious diseases associated with urban poverty [20]–[22]. Case fatality (10%) from leptospirosis [19] is comparable to that observed for meningococcal disease, visceral leishmaniasis and dengue hemorrhagic fever (20%, 8% and 10%, respectively) in this setting [20],[23],[24]. Furthermore, leptospirosis is associated with extreme weather events, as exemplified by the El Niño-associated outbreak in Guayaquil in 1998 [25]. Leptospirosis is therefore expected to become an increasingly important slum health problem as predicted global climate change [26],[27] and growth of the world's slum population [1] evolves.

Urban leptospirosis is a disease of poor environments since it disproportionately affects communities that lack adequate sewage systems and refuse collection services [6],[10],[11]. In this setting, outbreaks are often due to transmission of a single serovar, *L. interrogans* serovar Copenhageni, which is associated with the *Rattus norvegicus* reservoir [6], [28]–[30]. Elucidation of the specific determinants of poverty which have led to the emergence of urban leptospirosis is essential in guiding community-based interventions which, to date, have been uniformly unsuccessful. Herein, we report the findings of a large seroprevalence survey performed in a Brazilian slum community (*favela*). Geographical Information System (GIS) methods were used to identify sources for *Leptospira* transmission in the slum environment. Furthermore, we evaluated whether relative differences in socioeconomic status among slum residents contributed to the risk of *Leptospira* infection, in addition to the attributes of the environment in which they reside.

Fuente: Shotton *et al.*, 7.



Fuente: Castro *et al.* 2013, 53.

Figura 3. Representación de Biotea

abstracts textuales; acceso a datos procesables comprendidos en tablas, figuras, etc. permitiendo su manipulación y descarga; fusión de datos del artículos como dibujos, planos o mapas con recursos externos como Google Maps, lo que posibilita su superposición; presentación de las citas en su contexto original, de tal manera que el lector puede ver el texto original que está siendo citado; introducción de animación en las figuras, incluso en 3D; la reorganización personalizada de las referencias bibliográficas de distintas maneras a criterio del lector como orden alfabético, año de publicación, frecuencia de citación en el texto u orden original.

Complejidad media/alta. Identificación semántica de la estructura mediante la creación de artículos en formato XML o JATS, además de los habituales formatos PDF, HTML o EPUB procedentes de matrices JATS, e identificación semántica del contenido al proporcionar valor añadido a los textos resaltando entidades como conceptos temáticos, fechas, nombres propios, lugares, etcétera, tras una anotación manual o automática del texto, con el fin de facilitar su lectura y comprensión. Estas entidades también pueden ser enlazadas a recursos externos como bases de datos, glosarios, vocabularios u ontologías.

Complejidad alta. Los documentos XML JATS se RDFizan empleando técnicas de procesamiento del lenguaje natural para posteriormente efectuar una anotación semántica automática a partir de vocabularios, bases de datos especializadas u ontologías. Por último, una vez identificadas entidades y sus relaciones, este contenido enriquecido se interconecta con Linked Data, lo que posibilita la creación de redes de conocimiento.

Journal Article Tag Suite

El *Journal Article Tag Suite* (JATS) es un estándar para el marcado XML de artículos científicos que fue publicado en 2012 como ANSI/NISO Z39.96. Los antecedentes de JATS son las Definiciones de Tipo de Documento (DTDs) que la Biblioteca Nacional de Medicina de los Estados Unidos creó de 2003 a 2008.

En JATS se define un conjunto de elementos y atributos que describen los datos fuente, el contenido textual y gráfico de los artículos en sentido amplio porque también sirve para la descripción de artículos de investigación, de revisión, cartas al editor, las editoriales, instrucciones para autores o reseñas de libros.

Un artículo JATS está compuesto por etiquetas con una estructura jerárquica en árbol cuya etiqueta principal es <article> ,de la que cuelgan <front> con los metadatos de la revista y el artículo; <body> metadatos del contenido textual y gráfico del artículo; <back> lista de referencias, agradecimientos, apéndices o glosarios; <floats, group> para figura o tablas, y por último, <response> o <sub-article>, la primera para identificar un comentario del artículo y la segunda para referir la traducción del artículo a otro idioma.

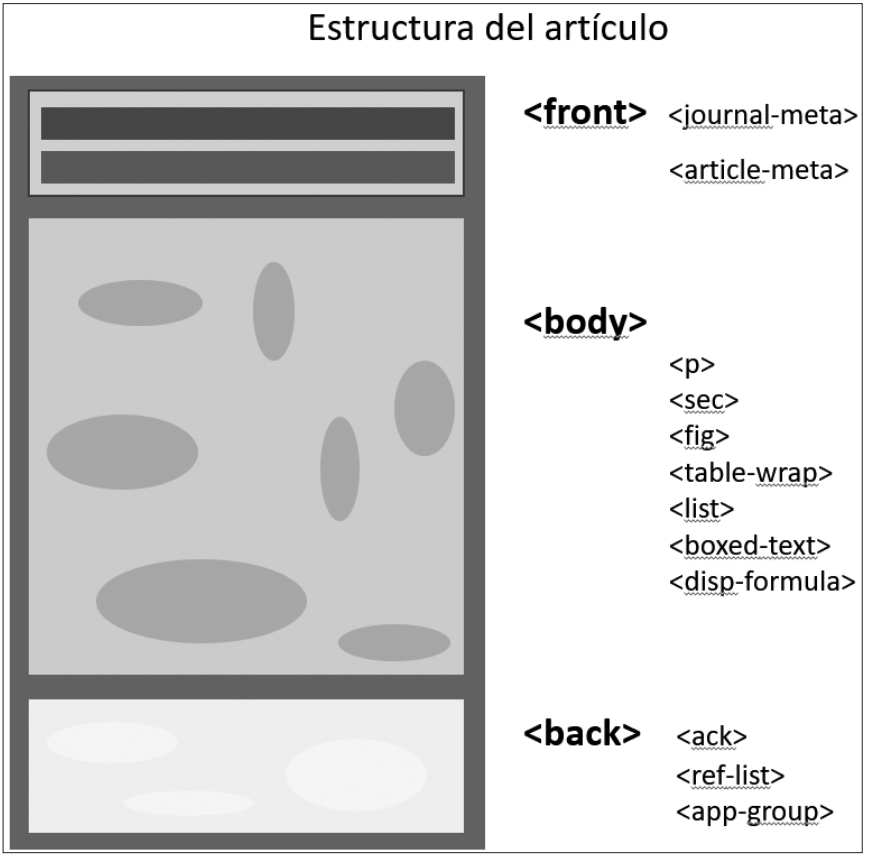
JATS incluye tres conjuntos de etiquetas o modelos de artículos con sus respectivos esquemas en DTD W3C XML Schema: para archivo e intercambio de artículos, para publicación de artículos y para la creación de artículos. En la actualidad, el modelo más extendido es el de publicación de artículos.

Asimismo, basado en JATS se han creado dos extensiones: Book Interchange Tag Suite (BITS), modelo XML para libros académicos y Standards Tag Suite para normas. BITS supone un formato común para que los editores de libros intercambien contenido de libros, incluidas partes o capítulos de libros. Y por otro lado, la STS: Standards Tag Suite, ANSI/NISO Z39.102-2017 y la ISOSTS (ISO Standards Tag Set) como estándares para la codificación XML de documentos de normas para propiciar el uso e intercambio de esta tipología documental.

Diversas y numerosas plataformas y editoriales de revistas electrónicas están requiriendo o admitiendo JATS. Algunas veces, los editores utilizan JATS más como un instrumento de seguridad y preservación digital, mientras que en otras ocasiones usan JATS para a partir de éste generar y publicar los artículos en formatos HTML, EPUB y PDF (PubMed Central, Redalyc, ScienceCentral) o incluso publican directamente en formato XML JATS (PLOS o PeerJ).

El coste de tiempo de conversión de artículos científicos de formato Word a JATS mediante editores específicos varía de una

Figura 4. Las tres principales etiquetas de JATS

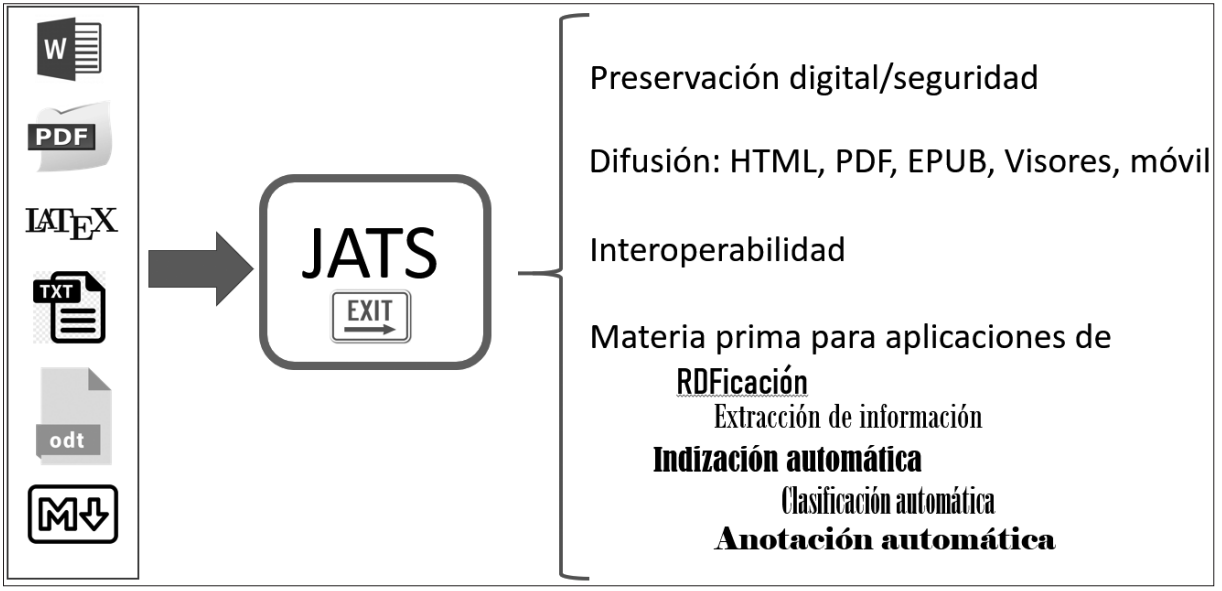


Fuente: Alves, Tony *et al.* (2017).

hora y media a cuatro horas aproximadamente, en función de la experiencia y pericia de los usuarios.[5] Si bien para reducir costes se han desarrollado *softwares* para la conversión automática de numerosos formatos de origen al JATS, e incluso se han implemen-

5 Guzmán-Useche, "Sustentabilidad,"; Redalyc; Eikebrokk, "EPUB as publication".

Figura 5. JATS visto como materia prima



Fuente: elaboración propia.

tado editores de texto para ser usados por los autores en el proceso de la creación de los textos, generando archivos XML JATS que son remitidos e incorporados directamente al flujo de trabajo de revisión por pares de las revistas y editoras. Parece solo cuestión de tiempo que todos terminemos tecleando en un editor que genere documentos JATS.

Así pues, una vez disponibles los documentos en formato XML JATS, ya están preparados para su preservación digital, para su difusión y comprensión por humanos tras la generación de formatos como HTML, PDF o EPUB, para el inicio de procesos de interoperabilidad o para convertirse en materia prima para procesos de procesamiento automática de información, puesto que XML JATS es un formato legible por ordenador. Así pues, los documentos XML JATS se convierten en el punto de partida de procesos para la generación de tripletas RDF, clasificación automática de contenidos, la indización automática de documentos o la anotación automática de información con el fin de hacer los contenidos más comprensibles.

Estudio de caso: indización automática y recuperación de información a partir de artículos científicos en formato XML JATS con SISA

SISA es un sistema de indización automática en entorno web que procesa el texto completo de artículos científicos, leyes/decretos y sentencias judiciales. Sus primeros pasos se encuentran en la tesis doctoral de este autor defendida en 1997, y desde entonces las sucesivas implementaciones han sido usadas para labores de docencia e investigación. El espíritu basal de SISA se centra en dos ideas: la metaestructura de los documentos y el conocimiento se transmite a través de la terminología; es decir, el valor de un término para la indización viene determinado por el lugar en el que se ubica y por su propio significado.[6] La idea de usar el lugar don-

6 Gil-Leiva, *Manual de indización,* 368-384 ; Gil-Leiva "SISA-Automatic", 139-162.

de aparecen los términos en los documentos no debe considerarse original ni nueva; de hecho, la propia norma ISO sobre indización de documentos ofrece indicaciones de a qué partes de los documentos prestar mayor atención.

Así pues, cada una de las partes de las tipologías documentales es reconocida por una marca inicial y marca final. De esta manera, el título de este artículo sería: #ITI# Posibilidades del XML JATS para el tratamiento y recuperación de información: El caso del sistema de indización automática SISA#FTI#. En la siguiente tabla se muestran las etiquetas para marcar la estructura y contenido de documentos procesados en SISA. Para la identificación de la terminología, se emplean vocabularios controlados con relaciones de sinonimia o tesauros.

Tabla 1. **Marcas empleadas en SISA**

Artículos	Leyes-Decretos	Sentencias
Título TI	Título TL / TD	Resumen RE
Resumen RE	Preámbulo PRE	Encabezamiento EN
Palabras clave PC	Índice IND	Antecedentes AN
Epígrafe EP	Título TI	Hechos HE
Primer párrafo PP	Artículo ART	Fundamentos FU
Título de tabla TT	Primer Párrafo Artículo PPA	Fallo FA
Título de Figura TF	Título Capítulo TC	
Conclusiones CO	Disposición Adicional DA	
Referencias RF	Disposición Transitoria DT	
	Disposición Derogatoria DD	
	Disposición Final DF	

Actualmente, SISA reconoce directamente la estructura y contenido de artículos publicados en formato XML por la *Revista Española de Documentación Científica* y en formato HTML por la revista *Information Research: An International Electronic Journal*, así como en formato HTML de los artículos publicados en Redalyc procedentes de documentos matrices JATS. De tal manera que solamente hay que descargarlos de sus respectivos sitios y cargarlos para su procesamiento en SISA.

Por tanto, desde la concepción a mitad de la década de 1990 y las sucesivas versiones implementaciones hasta hoy, este sistema de indización automática ha estado alineado espiritualmente con JATS incluso antes de su gestación, puesto que el funcionamiento de SISA y el estándar JATS se focalizan en la estructura y contenido de los documentos (ver figura 4 y tabla 1).

INDIZACIÓN AUTOMÁTICA DE ARTÍCULOS XML JATS

Disponer de artículos en formato XML procedentes de la *Revista Española de Documentación Científica* del Consejo Superior de Investigación Científica español o bien en HTML producidos desde JATS por el sistema de información Redalyc propicia que su procesamiento sea muy fácil y rápido por herramientas que acepten estos formatos. Una prueba ejecutada para esta ocasión con cuarenta artículos en formato XML que comprende la descarga de los documentos de la web de la *Revista Española de Documentación Científica*, la carga en SISA, la indización automática y la exportación de los metadatos descriptivos y la indización de cada artículo (Figura 6) ha tomado un tiempo de siete minutos en total, lo que significa unos 10 segundos para disponer de metadatos descriptivos y de indización, mientras que la obtención manual de estos metadatos por un profesional consumiría entre diez y trece horas (Tabla 2).

Figura 6. 'Collage' con los procesos para el procesamiento de artículos XML JATS

Fuente: elaboración propia.

Tabla 2. Tiempo estimado para el procesamiento automático de artículos científicos en XML o HTML procedentes de JATS

	Tiempo estimado		
	Artículos	**Indizador humano**	**SISA**
Metadatos descriptivos + Metadatos de indización	1	15-20 minutos	10 segundos
Metadatos descriptivos + Metadatos de indización	40	10-13 horas	7 minutos

Ordenación por relevancia de los resultados de búsquedas de información

Las bases de datos bibliográficas y repositorios con registros únicamente descriptivos conformados por los metadatos habituales de Título, Autores, Filiación, Revista, Resumen, Palabras clave o Descriptores, poco a poco, están siendo complementadas por los textos completos. Por tanto, las estrategias de búsqueda de información pueden abarcar el texto completo, ampliando los límites de los metadatos descriptivos a los confines de todo el texto como los títulos de epígrafes, títulos de tablas o figuras u otras partes del documento.

Como ya se ha mencionado anteriormente, uno de los fundamentos para la obtención de la indización automática por parte de SISA es la posición que ocupan los términos en los textos. Por tanto, tras el procesamiento automático de un artículo en formato XML JATS, SISA recopila datos sobre los términos (o sinónimos) que conforman un documento como en qué lugares aparecen y con qué frecuencia; de tal manera que esta valiosa información además de usarse para generar términos de indización se reutiliza durante el proceso de búsqueda de información, concretamente, para la ordenación de los documentos resultantes tras una búsqueda.

En una base de datos o repositorios con centenares de miles de documentos almacenados, es muy probable que una ecuación de

búsqueda recupere varios cientos o miles de documentos. De tal manera que los sistemas de información bibliográfica ofrecen opciones para ordenar y mostrar los resultados de la búsqueda por año, nombre de la revista, Primer autor; por el número de citas recibidas, y otra opción común es la de ordenar los resultados por 'relevancia'.

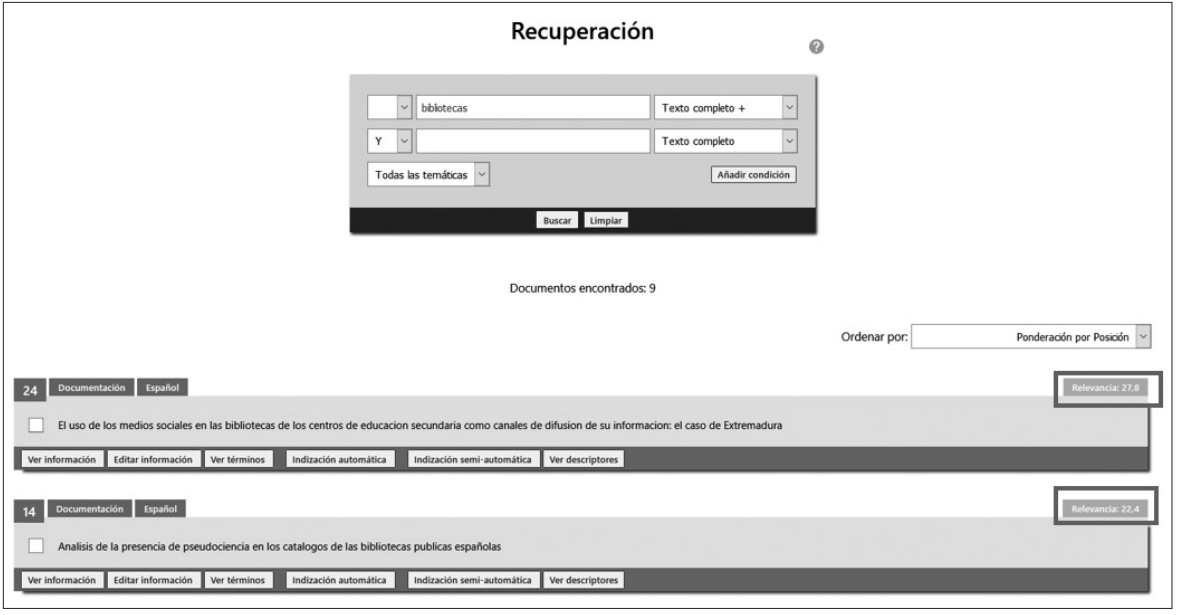
Bases de datos comerciales como Econlit, ERIC, WoS, SCOPUS, Arts & Humanities, FSTA o LISTA y muchas otras, ofrecen la ordenación por 'Relevancia', que es la relación existente entre el término o frase usada en la búsqueda con el contenido de los registros bibliográficos de la base de datos, si bien en estas bases de datos no es fácil identificar a primera vista en qué consiste exactamente esa relación que termina presentando una lista de documentos al usuario.

El módulo de recuperación de SISA brinda la posibilidad de ordenar los resultados de una búsqueda por la relación entre el término o frase de búsqueda con el contenido íntegro de los documentos. En concreto, permite ordenar los resultados de una búsqueda por cuatro maneras diferentes de relevancia: Posición, Frecuencia, TFIDF y de forma conjunta por Posición Frecuencia y TFIDF.

Los datos disponibles en SISA tras el procesamiento de los documentos permiten asignar una ponderación numérica a cada palabra o frase que posteriormente podría ser reutilizada para establecer la ordenación y presentación de los documentos por 'Relevancia' dada a los usuarios como se muestra en la Figura 7.

Veamos esto con un ejemplo. El Doc24 de la Figura 7, cuyo título es 'El uso de los medios sociales en las bibliotecas de los centros de educación secundaria como canales de difusión de su información: el caso de Extremadura', cuenta con una Ponderación por posición de 27.8 porque la palabra 'bibliotecas' que fue la usada para efectuar la búsqueda en el módulo de recuperación de SISA aparece en el Título una vez, tres veces en el Resumen, una vez en los Epígrafes, cinco veces en los Primeros párrafos de un epígrafe, veintiocho veces en Otros párrafos, una vez en el Título de una figura, diez veces en las conclusiones y siete veces en

Figura 7. Búsqueda en el módulo de recuperación de SISA por el término 'bibliotecas'



las referencias bibliográficas del artículos; mientras que la palabra 'bibliotecas' en otros documentos cuenta con una Ponderación por posición 0.2 porque aparece solamente en Otros párrafos, una sola vez y, además, en un lugar no significativo del texto; de ahí que ocupe esa posición.

Tabla 3. Ordenación de resultados por fecha y ponderación por posición

Ordenados por		
Orden	**Fecha de incorporación al sistema**	**Ponderación por posición**
1	Doc1	Doc5 Ponderación: 27.8
2	Doc2	Doc3 Ponderación: 17.2
3	Doc3	Doc7 Ponderación: 8.9
4	Doc4	Doc1 Ponderación: 2.4
5	Doc5	Doc4 Ponderación: 0.7
6	Doc6	Doc2 Ponderación: 0.2
7	Doc7	Doc6 Ponderación: 0.2

Por tanto, parece razonable pensar que a la mayoría de los usuarios que interrogaran con una base de datos similar buscando documentos sobre 'bibliotecas' les sería de mayor utilidad el Doc5 que los documentos Doc4, Doc2 y Doc6; de ahí que en la lista de resultados se debería mostrar el Doc5 antes que los otros. Extrapolando esto a un sistema de información en el que los documentos recuperados para una determinada búsqueda sean varios cientos o miles resultaría de gran ayuda ahorrando tiempo y esfuerzo a los usuarios.

CONCLUSIONES

Desde el siglo XVII hasta bien avanzada la segunda mitad del siglo XX no se produjeron cambios significativos en el flujo de trabajo de la edición y difusión de la ciencia. Con la aparición y universalización de los ordenadores y la llegada de la tecnología web

en primer lugar, y posteriormente la web semántica, desde 2010 aproximadamente los editores vienen empleando cada vez más la denominada publicación semántica que permite un fácil y más rápido procesamiento de la información al tiempo que amplía las posibilidades de difusión, intercambio y reutilización de información. Han sido mostrados unos pocos ejemplos de lo que supone la publicación semántica. Se trata de prácticas relativamente recientes pero que con seguridad tendrán un largo recorrido que se revela espoleada o delimitada únicamente por los confines de la creatividad de los editores.

Por otro lado, han sido ofrecidos ejemplos de cómo artículos en formato XML de la *Revista Española de Documentación Científica* del CISIC y en formato HTML generado a partir de JATS en el sistema de información Redalyc son materia prima para un procesamiento automático óptimo y fácil. SISA logra en apenas unos segundos, que con toda seguridad podrán reducirse a la mitad en mejores entornos de trabajo, un conjunto útil de metadatos descriptivos y de indización para cada artículo procesado. De igual modo, también se ha explicado cómo SISA a partir de los mencionados artículos XML JATS puede reducir el coste- tiempo de los usuarios a la hora de revisar los resultados mostrados tras una búsqueda, ordenando los mismos por una relevancia por posición.

Por tanto, se anima a los editores de publicaciones científicas a incorporar en sus flujos de trabajo el formato XML JATS. Este formato garantiza la preservación de sus materiales originales y además facilita el intercambio de información científica; favorece que los propios editores o terceros generen redes de conocimiento enlazado; asimismo, incentiva el desarrollo de aplicaciones y herramientas que podrían llegar a las unidades documentales en forma de *software* libre. Permite, en definitiva, que las computadoras y las personas trabajen en cooperación.

AGRADECIMIENTOS

Agradecemos a la Unidad de Análisis Documental y Producción de Bases de datos del Centro de Ciencias Humanas y Sociales del

CSIC; en concreto a Teresa Abejón Peña por proporcionarnos el vocabulario controlado que usan para la indización de documentos en el área de Biblioteconomía y Documentación que nos permite llevar a cabo tareas de evaluación con SISA.

REFERENCIAS

Berners-Lee, Tim, J. Hendler y O. Lassila. "The Semantic Web". *Scientific American (*mayo 2001): https://doi.org/10.1038/35074206.

Castro, Leyla Jael Garcia, Rafael Berlanga, Dietrich Rebholz-Schuhmann y Alexander Garcia. "Connections across scientific publications based on semantic annotations". En 3rd Workshop on Semantic Publishing (2013), 10th Extended Semantic Web Conference, Montpellier, France, 26 mayo 2013, 51-62.

Eikebrokk, Trude, Tor Ame Dahl y Siri Kessel. "EPUB as Publication Format in Open Access Journals: Tools and Workflow". *Code4Lib Journal* (2014).

Gil-Leiva, Isidoro. *Manual de indización. Teoría y práctica*. Gijón: Trea, 2008.

___. "SISA-Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules Versus TF-IDF Rules". *Knowledge Organization* 44, (2017): 139-162. https://doi.org/10.5771/0943-7444-2017-3-139.

Guzmán-Useche, Eliana y Fernando Rodríguez-Contreras. "Sustentabilidad de las iniciativas latinoamericanas de publicación de revistas científicas en acceso abierto utilizando el estándar XML JATS: el caso de SciELO". *Biblios: Revista electrónica de bibliotecología, archivología y museología*, 64 (2016). https://doi.org/10.5195/biblios.2016.290.

Kircz, Joost. "Modularity: The next form of scientific information presentation?". *Journal of Documentation*, 54, (1998): 210-235. https://doi.org/10.1108/EUM0000000007185.

Scholastica survey: The State of Journal Production and Access 2020, https://lp.scholasticahq.com/journal-production-access-survey/.

Shotton, David. "Semantic publishing: The coming revolution in scientific journal publishing". *Learned Publishing* 22, núm. 2 (2009): 85-94. https://doi.org/10.1087/2009202.

Shotton, Dadid, Katie Portwin, Graham Klyne y A. Alistair Miles (2009). "Adventures in semantic publishing: exemplar semantic enhancements of a research article". *PLoS Computational Biology*, 5, núm. 4 (2009). https://doi.org/10.1371/journal.pcbi.1000361.

La investigación bibliotecológica y de la información hacia el 2030: desarrollo sostenible. Instituto de Investigaciones Bibliotecológicas y de la Información/UNAM. La edición consta de 100 ejemplares. Coordinación editorial, Anabel Olivares Chávez; revisión especializada, Valeria Guzmán González; corrección de pruebas, Carlos Ceballos Sosa; revisión de pruebas, Valeria Guzmán González y Carlos Ceballos Sosa; formación editorial, Sonia Wendy Chávez Nolasco. Fue impreso en papel cultural de 90 gr en los talleres de Litográfica Ingramex, Centeno 162-1,Col. Granjas Esmeralda, Iztapalapa, C.P. 09810, Ciudad de México. Se terminó de imprimir en septiembre de 2022.