



SOFTWARE JSEEDITOR PARA LA MARCACIÓN AUTOMÁTICA DE ARTÍCULOS DE LA REVISTA BIOTECNOLOGÍA VEGETAL PARA LA BASE DE DATOS SCIELO

JSEEDITOR SOFTWARE FOR THE ARTICLES' AUTOMATIC MARKING OF VEGETABLE BIOTECHNOLOGY JOURNAL

*Lisvandy Amador Penichet, Hansel Castellanos León², Yelenys Alvarado Capó³, Lester R. Santos
Martínez⁴*

1 Instituto de Biotecnología de las Plantas. Universidad Central "Marta Abreu" de Las Villas, Cuba, lisvandy@ibp.co.cu, 7ma del Oeste #27 / 1ra del Sur y 2da del Sur, Placetas, Villa Clara

2 Facultad de Matemática, Física y Computación. Universidad Central "Marta Abreu" de Las Villas, Cuba.

3 Instituto de Biotecnología de las Plantas. Universidad Central "Marta Abreu" de Las Villas, Cuba, yelenys@ibp.co.cu

4 Instituto de Biotecnología de las Plantas. Universidad Central "Marta Abreu" de Las Villas, Cuba, lester@ibp.co.cu

RESUMEN: SciELO es un modelo para publicación electrónica de revistas creado para responder a las necesidades de la comunicación científica en los países en desarrollo y particularmente de América Latina y el Caribe. La marcación de textos de los artículos científicos en HTML debe responder a las exigencias de la metodología SciELO para la publicación de documentos electrónicos. Este trabajo persiguió como objetivo desarrollar un software para la marcación automática de artículos de la revista Biotecnología Vegetal. Se emplearon archivos electrónicos en formato HTML de artículos de la revista. Se definieron los metadatos requeridos para la marcación, entre ellos: título, autores, palabras claves, referencias bibliográficas. Los resultados se compararon con el programa Markup versión 4.0 de SciELO. Con el software desarrollado fue posible la marcación automática de los artículos científicos y con ello se redujo el tiempo de marcado y el número de errores por artículo. Este programa permite continuar el trabajo de la revista Biotecnología Vegetal para su indexación en SciELO.

Palabras Clave: marcación de metadatos; revistas científicas; indexación; XML.

ABSTRACT: SciELO is a model for electronic journal publication created to respond to the needs of scientific communication in developing countries and particularly in Latin America and the Caribbean. The marking of texts of scientific articles in HTML must meet the requirements of the SciELO methodology for the publication of electronic documents. This work aimed to develop a software for the automatic marking of articles belonging to the Biotecnología Vegetal journal. Electronic files in HTML format of the journal's articles were used. The metadata required for the marking were defined, among them: title, authors, keywords, bibliographical references. The results were compared with the Markup version 4.0 SciELO program. With the developed software, it was possible to mark automatically the scientific articles, reducing marking time and number of errors per article. This program allows continuing the work of the journal Biotecnología Vegetal for its indexing in SciELO.

Key Words: metadata marking; scientific journals; indexing; XML

1. INTRODUCCIÓN

La gestión de la información y el conocimiento a partir de la bibliografía científica disponible en la web se ha convertido en una tarea cotidiana y necesaria para los investigadores de las diferentes ramas. Es por ello que se hace común encontrar cada día en internet más revistas de carácter científico que se dediquen a la publicación de los resultados obtenidos por los investigadores de todo el mundo, con el objetivo de lograr un acceso multipolar al conocimiento.

Dentro de estas revistas las de acceso abierto han ganado en popularidad ya que garantizan que los contenidos que en ellas se publican puedan ser vistos, descargados, impresos y distribuidos por los usuarios sin la necesidad de tener que pagar por ello [2]. Es así que surgen bases de datos internacionales que agrupan este tipo de revistas con el objetivo de facilitar el acceso al conocimiento por parte de los científicos. Dentro de estas bases de datos se destacan Redalyc, Latindex y SciELO.

La base de datos SciELO es de las más reconocidas cuando se habla de acceso abierto. SciELO tiene como objetivo aumentar de forma sostenible la accesibilidad, visibilidad, uso, calidad e impacto de las revistas que publica [4]. Esta base de datos agrupa un gran número de revistas de diferentes áreas del conocimiento [3], teniendo todas ellas en común, el libre acceso a sus contenidos por parte de los usuarios.

El proceso de indexación de una revista en la base de datos SciELO requiere el cumplimiento de una serie de parámetros que garantizan que la revista cuente con el rigor determinado para ser incluida en esta base de datos. Entre estos se pueden mencionar la periodicidad, puntualidad, revisión por pares, acceso abierto, estar indexada en los catálogos de DOAJ y Latindex y su inclusión en otros índices internacionales.

Además es requisito necesario tener marcados todos los artículos que van a ser indexados en esta base de datos para generar las versiones de los mismos en formato XML. Esto garantiza que aumente la visibilidad de la revista dado que los motores de búsquedas trabajan más eficientemente con los archivos en este formato.

El proceso de marcación es bastante engorroso ya que el software proporcionado para realizar el mismo requiere de una total interacción con el usuario, por lo tanto, es necesario usar personal especializado que conozca cómo funciona el proceso para que

este sea satisfactorio. Esta tarea consume bastante tiempo dada las dificultades mencionadas. En el mejor de los casos el software permite la marcación semiautomática de los artículos. Para ello es necesario configurar un autómata que sea capaz de reconocer la norma en que se escribe la revista [5]. Este proceso llega a ser tan complicado y los beneficios que proporciona tan escasos, que en la mayoría de los casos las revistas optan por la opción de marcación manual.

La Revista Biotecnología Vegetal se prepara para ser indexada por la Base de Datos SciELO y se hace necesario economizar el tiempo de trabajo del colectivo editorial de esta revista, dado que sus miembros son investigadores del Instituto de Biotecnología de las Plantas y por ende disponen de un tiempo limitado. Es por ello que el presente trabajo tiene como objetivo general: desarrollar un software para la marcación automática de artículos de la revista Biotecnología Vegetal de acuerdo a las normas de la base de datos SciELO.

2. CONTENIDO

Actualmente el proceso editorial de la Revista Biotecnología Vegetal perteneciente al Instituto de Biotecnología Vegetal (IBP) se realiza de forma manual. La publicación de la revista es de carácter trimestral por lo que se procesan 4 números al año y cada número posee un cierto número de artículos que varía (normalmente entre 7 y 12). Los pasos principales de este proceso comprenden:

- (1) recepción de artículo por parte del editor.
- (2) asignación a los árbitros.
- (3) decisión de publicación por parte de los árbitros.
- (4) si el artículo se considera publicable, el autor debe realizar los arreglos sugeridos por los árbitros.
- (5) revisión final por parte del editor.
- (6) diagramación del artículo.
- (7) obtención del artículo en formato PDF y HTML.
- (8) marcación del artículo para su entrega a la base de datos SciELO.
- (9) publicación del artículo.

La marcación de los artículos científicos para luego generar los XML es una de las partes que más tiempo requieren en el proceso editorial. Esta demora está dada debido a que el proceso se realiza de forma manual a través del conjunto de programas del paquete *scielo-pc-programs-4.0.091* de la base de datos SciELO y su herramienta de marcación

convencional Markup. Esta herramienta también cuenta con un módulo que permite la marcación semiautomática. Esto se hace mediante la configuración de un autómata que sea capaz de reconocer la norma de la revista cuyos artículos deseen ser marcados. A pesar de que la configuración del autómata reduce considerablemente el tiempo de marcación de los artículos, este puede ser reducido aún más si se logra la marcación totalmente automática. También es válido aclarar que la mayoría de las revistas marcan sus artículos manualmente debido a que la configuración del autómata es un proceso complejo. Esta configuración requiere de conocimientos que normalmente las personas integrantes de los cuerpos editoriales no tienen.

Por lo descrito anteriormente se decide crear un módulo en la herramienta JSEditor encargado de realizar la marcación de los artículos científicos de forma automática. Este módulo se complementa con los programas de SciELO como son el SGML Parser encargado de realizar la comprobación del proceso de marcación y la herramienta Converter la cual permite la conversión al formato XML de los artículos marcados.

2.1 Proceso de marcación del artículo. El caso particular de las referencias bibliográficas.

El proceso de marcación se le realiza al artículo en formato HTML. La marcación consiste en la inyección de código HTML al artículo (inclusión de la etiquetas de marcación) que permite estructurar cada una de las partes del mismo. Luego el artículo es revisado con la herramienta SGML Parser para comprobar que las etiquetas insertadas están acorde al DTD usado por la metodología SciELO para la obtención de los artículos en formato XML. Con el posterior uso de la herramienta Converter se obtiene una versión del artículo en formato XML. Esto garantiza una mayor visibilidad del mismo en la web, dado

que los motores de búsquedas encuentran más fácilmente la información en los documentos XML. Es por ello que este lenguaje se ha hecho popular para almacenar la información científica.

Las partes a marcar en una publicación son: título, título en inglés, afiliación, resumen, palabras clave, abstract, keywords, cuerpo del documento, referencias, fecha de recibido y aceptado. Además a esto se agrega la marcación del tipo de artículo, que para el caso de la revista Biotecnología Vegetal puede ser: Artículo Original, Comunicación Corta, Reseña Bibliográfica y Nota Técnica.

Dentro de las secciones a marcar resaltan las referencias bibliográficas debido a su gran diversidad (9 tipos de referencias) y a la gran importancia que estas tienen. La correcta marcación de las referencias bibliográficas posibilita que en los procesos de recuperación de la información el artículo pueda ser encontrado más fácilmente, ya que estas constituyen uno de los metadatos fundamentales que usan los motores de búsqueda para encontrar la información relevante a partir de las consultas generadas por los usuarios. También las referencias contienen elementos que permiten realizar el proceso de referencia cruzada. Dentro de estos elementos está el DOI y el ISBN. Mediante este proceso se pueden encontrar otros artículos que aborden temas similares al artículo que se trata en cuestión.

Las referencias encontradas en los artículos de la revista Biotecnología Vegetal pueden ser clasificadas en los siguientes tipos: Book, Book Section, Edited Book, Conference Proceeding, Journal Article, Web Page, Patent, Thesis, Online Database.

Cada uno de los tipos de referencias mencionadas anteriormente cuenta con una estructura que se adecúa a la norma de publicación de la revista (ISO 690 con modificaciones) lo que facilita el proceso de clasificación y revisión de las mismas. Dicha estructura se muestra en la Figura 1.

Book:	Autor (Año) Título. Editorial, Lugar de publicación [; doi: ISBN:]
Book section:	Autor (Año) Título. <In: En:> Editor <(ed) (eds)>. Título del Libro, pp. -. Editorial, Lugar de publicación [; doi: ISBN:]
Edited book:	Editor <(ed) (eds)> (Año) Título. Editorial, Lugar de publicación [; doi: ISBN:]
Conference proceeding:	Autor (Año) Título. In: Editor <(ed) (eds)> Nombre de la Conferencia, Lugar de la Conferencia, día inicial-día final/mes/año, pp. - . Editorial, Lugar de publicación [; doi: ISBN:]
Journal article:	Autor (Año) Título. Revista Volumen [(Número)]:Páginas– [; doi:]
Online database:	Título (Año) Editorial. URL. . <Accessed Accedido> Fecha
Patent:	Autor (Año) Título. País Número de Patente, Fecha
Thesis:	Autor (Año) Título. <Thesis Tesis> Grado de la Tesis, Universidad, Lugar de Publicación, País
Web page:	Autor (Año) Título. <Available in: Disponible en:> URL. <Accessed Accedido> Fecha

Figura 1 Estructura de cada uno de los tipos de referencias permitidos para la norma usada por la revista Biotecnología Vegetal.

Las palabras y los símbolos resaltados en negrita son de carácter obligatorio. Las palabras pueden aparecer en idioma inglés o en español en dependencia del idioma original de artículo que se referencia. Los parámetros encerrados entre corchetes "[]"

son opcionales, por tanto pueden aparecer en las referencias o no. Los parámetros encerrados entre paréntesis angulares "< >" son excluyentes.

A partir de esta estructura fue posible conformar un

árbol de decisión Figura 2 que permitió dada una referencia clasificarla. Esta clasificación puede tener dos salidas: (I) en caso que la referencia esté correctamente estructurada devuelve el tipo de la referencia, (II) en otro caso informa al usuario que la referencia no se adecúa a ninguno de los tipos predefinidos como válidos.

Una vez clasificadas cada una de las referencias este módulo se encarga de revisarlas estructuralmente. En caso de error el usuario será informado

mediante un mensaje. Además el programa brinda una pista de dónde pudiera estar el error y sugiere el tipo al cual pudiera pertenecer la referencia para facilitar al editor su corrección. Esta corrección puede ser realizada directamente desde la interfaz del programa, lo que constituye un ahorro de tiempo significativo al no tener que recurrir al artículo en formato Word para hacer los arreglos pertinentes. Si los arreglos se hacen directamente sobre el artículo en formato Word es necesario convertir nuevamente el artículo a HTML y volver a revisar la estructura.

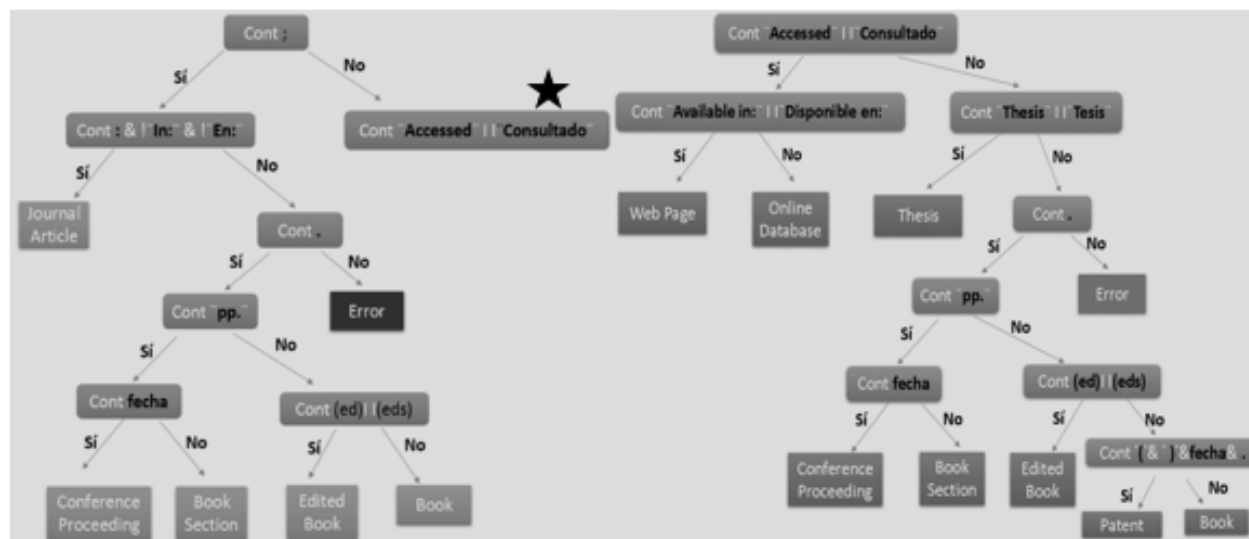


Figura 2 Árbol de decisión para la clasificación de las referencias bibliográficas. El árbol de la derecha es la continuación del nodo marcado con una estrella.

Desde el instante que el artículo fue revisado y no contiene ningún tipo de error este está listo para ser

2.2 Acerca de la herramienta

La implementación del software JSEditor versión 1.1 se llevó a cabo usando el lenguaje de programación Java en el entorno de desarrollo NetBeans 8.0.2 para Windows con JDK versión 1.7.

Para llevar a cabo el procesamiento del artículo en formato HTML se utilizó la biblioteca jsoup versión 1.8.3, la misma contiene una serie de clases que facilitan la manipulación de documentos HTML permitiendo moverse por los nodos del árbol DOM de cada uno de ellos.

2.3 Validación de la herramienta

Con el objetivo de realizar la validación de la herramienta desarrollada se decidió someterla a pruebas de criterio de expertos y de esta manera comprobar los resultados.

Muestra

Para validar este módulo se tomaron como muestra 15 artículos pertenecientes al volumen 16 de la

sometido al proceso de marcación.

revista Biotecnología Vegetal. Es importante resaltar que este procedimiento pudo ser realizado por cualquier usuario ya que la herramienta desarrollada no exige el conocimiento del proceso de marcación ni de las etiquetas que el mismo incluye pues la marcación se realiza de forma automática.

Procedimiento y Resultados

Se seleccionó una muestra de 15 artículos y se sometieron al proceso de marcación con la herramienta desarrollada. Aquí se manifiesta una de las ventajas principales de este software y es que permite la marcación de una colección completa. Esta facilidad no está disponible en el software convencional usado para la marcación.

Packer en [5] refiriéndose al tiempo que ocupa la marcación de un artículo científico con la herramienta convencional Markup, plantea que la marcación manual solamente de las referencias bibliográficas tarda 84.36 minutos y la marcación semiautomática a través de la configuración del autómata solamente de las referencias bibliográficas tarda 47.81

minutos. Para una muestra de 15 artículos los tiempos de marcación ascienden a:

- para la marcación manual un total de 21.9 horas,
- para la marcación semiautomática asciende a 11.95 horas

En ambos casos solo se mide el tiempo de la marcación de las referencias bibliográficas.

Con la aplicación implementada (JSEditor) el tiempo de marcación promedio de un artículo es 0.2 segundos. La marcación de los 15 artículos sometidos a prueba demoró 2.68 segundos.

En estudios más recientes, referidos al tiempo de marcación con la herramienta convencional Markup, Guzmán-Useche y Rodríguez-Contreras en [1] plantean que:

1. Crear la revista, lo cual se hace una sola vez, toma entre 4 y 10 minutos.
2. Crear las secciones usadas en el número que se está procesando toma entre 1 y 2 minutos.
3. Crear el número de la revista a ser marcado toma entre 1 y 2 minutos.
4. La preparación del documento para ser marcado toma entre 20 y 95 minutos.
5. La marcación es el proceso que más tiempo ocupa, entre 20 y 110 minutos.
6. Generar el archivo XML al finalizar el marcado toma entre 1 y 6 minutos (si no hay errores en la marcación es más rápido, pero si hay errores de marcación puede tardar hasta 10 minutos).
7. Depurar los errores toma entre 0 (si no hay ningún error) y 15 minutos (siempre y cuando quien procesa tenga conocimientos de las etiquetas de SciELO y de XML).

Es válido aclarar que los tiempos anteriormente mencionados fueron obtenidos por un grupo de personas con más de 10 años de experiencia con el uso de la herramienta Markup.

Para la comparación que se realiza solamente son relevantes los tres últimos tiempos. Al interpretar la información detalladamente se aprecia que:

- Para el mejor caso con el uso de la herramienta Markup el tiempo de marcación es de $20+1+0=21$ minutos (suma de los pasos 5, 6 y 7). Esto constituye el escenario óptimo para este procedimiento, caso cuya probabilidad es bastante baja.
- Para el peor de los casos el tiempo de marcación asciende a $110+10+15=135$ minutos, cifra está bastante considerable.

En cualquier circunstancia con el uso de la herramienta desarrollada el tiempo de marcación es de solamente 1 minuto, que es el tiempo empleado para la generación del artículo en formato XML. Los tiempos de marcación como se apreció anteriormente son despreciables y la depuración de errores de marcación no es necesaria ya que la automatización de este proceso asegura la no incursión en errores de esta índole.

3. CONCLUSIONES

- La marcación automática de los artículos científicos de la revista Biotecnología Vegetal, mediante el uso de la herramienta JSEditor, disminuye considerablemente el tiempo requerido para esta tarea.
- Los resultados obtenidos al someter el módulo de marcación a pruebas de desempeño demuestran el software desarrollado constituye una solución factible y viable para la marcación de los artículos científicos de la revista Biotecnología Vegetal.

4. REFERENCIAS BIBLIOGRÁFICAS

1. Guzmán-Useche, E., y F. 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 No. 64, 2016.
2. Mendivelso, J. C.: Retos para las revistas electrónicas de libre acceso en Iberoamérica. A propósito de la I jornada sobre revistas científicas electrónicas españolas en acceso abierto. Ar@cne: revista electrónica de recursos en internet sobre geografía y ciencias sociales, 2006.
3. Miguel, S.: "Revistas y producción científica de América Latina y el Caribe: su visibilidad en SciELO, RedALyC y SCOPUS", Revista Interamericana de Bibliotecología, Vol. 34, No. 2, pp. 187-199, 2011
4. Packer, A. L.: "El SciELO para nuestras revistas científicas", Interciencia, Vol. 27, No. 6, pp. 274, 2002.
5. Packer, A. L.; M. Rocha Biojone; I. Antonio; R. Mayumi Takemaka; A. Pedrosa García; A. Costa da Silva y R. F. Delbucio: "SciELO: una metodología para la publicación electrónica", Acimed, 9(Supl.), 2001.

5. SÍNTESIS CURRICULARES DEL AUTOR PRINCIPAL

Nacido en Placetas, Villa Clara el 5 de mayo de 1991. Estudió Ciencia de la Computación en la Universidad Central "Marta Abreu"

de Las Villas en el periodo 2010-2015. Línea de investigación principal la minería de texto.

Eventos.

1-**Lisvandy Amador Penichet**, Idalmis Bermúdez Caraballos, Neysi Parada Gamboa. Evaluación de la percepción pública de los Organismos Genéticamente Modificados a través del software MultiTransG 1.0. XII Simposio Internacional de Biotecnología vegetal (2016).

2- **Lisvandy Amador Penichet**, Hansel Castellanos León, Lester R. Santos Martínez, Yelenis Alvarado Capó. Desarrollo del software BioVegMarker para el procesamiento de metadatos en artículos de la revista Biotecnología Vegetal para la base de datos SciELO. Fórum Provincial de Ciencia y Técnica (2017) Premio Relevante, Impacto a la vitalidad del centro en la producción y los servicios.

3- **Lisvandy Amador Penichet**, María M. García Lorenzo, Damny Magdaleno Guevara. Gestión del conocimiento en artículos científicos a partir de la información brindada por las referencias bibliográficas. XI Jornada Nacional Bibliotecaria. (2017)

Publicaciones

1- **Lisvandy Amador Penichet**, María M. García Lorenzo, Daniel Gálvez Lío, Damny Magdaleno Guevara (2017) SemClustDML: algoritmo para agrupar artículos científicos basado en la información brindada por las referencias bibliográficas. Revista Cubana de Ciencias Informáticas. 11 (2): 46-60.