Palabra Clave, (La Plata), octubre 2019-marzo 2020, vol. 9, n° 1, e075. ISSN 1853-9912
Universidad Nacional de La Plata
Facultad de Humanidades y Ciencias de la Educación
Departamento de Bibliotecología

LuXMeL: hacia la interoperabilidad Redalyc/AmeliCA-Scielo

LuXMeL: towards a RedALyC/AmeliCA-SciELO interoperability

Lucía Correa
Universidad Nacional de La Plata, Argentina
lucorrea1@hotmail.com

Francisco Chiarullo
Investigador independiente, Argentina
franciscochiarullo@hotmail.com

Resumen:

En este artículo se presenta una herramienta de trabajo que permite la automatización de las tareas necesarias para la conversión de un XML-JATS, producto de la herramienta de marcado de RedALyC y AmeliCA, al formato de XML-JATS utilizado por SciELO Argentina para la ingesta de revistas a su plataforma.

Palabras clave: Acceso Abierto, AmeliCA, RedALyC, SciELO, Planilla de cálculo, XML-JATS.

Abstract:

In this article we present a new work tool that allows the user the automatization of the work needed to convert an XML-Jats archive generated by the marking tool of RedALyC and AmeliCA, to the XML format that SciELO Argentina uses to upload journals to its platform.

Keywords: Open Access, AmeliCA, RedALyC, SciELO, Spreadsheet, XML-JATS.

Introducción

Dentro de lo que es la comunicación científica en América Latina, no podemos dejar de hablar de acceso abierto al conocimiento. Este modelo es una forma de trabajo para los países latinoamericanos desde el inicio de la edición científica a finales del siglo XIX. Es habitual, para la región, que los investigadores publiquen sus resultados en revistas editadas por las universidades, las cuales son gratuitas tanto para el que publica como para el que lee. En palabras de Dominique Babini (2019): "Nuestra fortaleza en América Latina es que estamos donde ellos quieren llegar: tenemos comunicaciones científicas gestionadas por la misma comunidad científica, en formas colaborativas, sin tercerización comercial."

Una de las maneras de determinar la calidad editorial de estas publicaciones es su inclusión en plataformas de indización y evaluación, donde pueden ingresar para ser reconocidas a nivel internacional y tener un sello de calidad. Para este trabajo nos enfocaremos en tres plataformas que promueven el acceso abierto: RedALyC (http://www.redalyc.org/), AmeliCA (http://www.amelica.org/) y SciELO (https://scielo.org/). Todas estas plataformas le otorgan a las revistas una enorme visibilidad y difusión.

RedALyC y SciELO son bases que requieren altos criterios de calidad que las revistas deben cumplir para poder ingresar. AmeliCA, por otro lado, es una iniciativa basada en la tecnología RedALyC y pretende ser una infraestructura de comunicación para la publicación académica y la ciencia abierta. Sus requisitos de ingreso son mínimos, y las herramientas de edición que ofrece son variadas. AmeliCA utiliza para su plataforma de marcado en XML-JATS tecnología provista por RedALyC (MarcALyC). Esta herramienta, además del

Recepción: 24 de septiembre de 2019 | Aceptación: 20 de octubre de 2019 | Publicación: 28 de octubre de 2019

Cita sugerida: Correa, L. y Chiarullo, F. (2019). LuXMeL: hacia la interoperabilidad Redalyc/AmeliCA-Scielo. *Palabra Clave (La Plata), 9*(1), e075. https://doi.org/10.24215/18539912e075

XML, provee al editor distintos formatos de publicación (html, pdf, epub), y le permite la descarga para la publicación en su página web.

SciELO, por otro lado, ofrece su tecnología para poder incluir las revistas que fueron invitadas a ese portal, pero sin beneficios para la propia edición de las revistas (Banzato y Rozemblum, 2019), ya que el proceso de marcado, si bien arroja también un XML, es únicamente para utilización de SciELO, que publica los números en su portal.

Impulsados por los mayores beneficios obtenidos, le hemos dado prioridad en el flujo de trabajo a Marcalyc/AmeliCA, y, con el apoyo del equipo mexicano y del personal del CAICYT (SciELO Argentina) (http://www.caicyt-conicet.gov.ar/sitio/), diseñamos un proceso de validación para poder continuar trabajando con Marcalyc/AmeliCA y reutilizar su XML para SciELO Argentina (http://www.scielo.org.ar/scielo.php) sin duplicar las tareas.

Desarrollo (o ¿por qué una planilla de cálculos?)

En la Facultad de Humanidades y Ciencias de la Educación (FaHCE) de la Universidad Nacional de La Plata (UNLP) –Argentina– nos encontramos, luego de la incorporación de todas nuestras revistas a AmeliCA, con el problema del doble trabajo para poder mantener las indizaciones en SciELO, ya que, como explican Banzato y Rozemblum:

> Hacia finales de 2012, con la instalación de varias revistas en el portal OJS de la FaHCE el flujo de edición no incluía el marcado que requería SciELO para la publicación, sino que se hacía luego de la publicación, sólo con fines de visibilidad, ya que es un sistema muy engorroso y lento para trabajar. Además, en 2014, SciELO comienza a tomar decisiones que no cumplían con nuestras expectativas del Acceso Abierto, por lo cual hemos cambiado las prioridades de nuestro flujo de trabajo, redireccionando nuestro esfuerzo al sistema de marcado XML-JATS de RedALyC (Banzato y Rozemblum, 2019).

Es a partir de estos cambios en los flujos de trabajo que surge la necesidad de modificar los XML que arroja el marcador de AmeliCA (y el de RedALyC), para enviarlos a SciELO y evitar duplicar esfuerzos y recursos, manteniendo nuestro modelo de trabajo sin perder la visibilidad que otorga SciELO.

Las primeras pruebas fueron modificaciones realizadas a mano con base en las correcciones que íbamos recibiendo desde el Equipo SciELO Argentina en el CAICYT, el cual se encarga de realizar en nuestro país la indización y carga de las revistas en SciELO. Sin embargo, las pruebas no otorgaron los resultados esperados: muchas veces, tras lo que parecía ser una validación exitosa, el sistema arrojaba errores.

En julio de 2019, logramos, gracias al aporte de Fernando Javier Rodríguez Contreras, miembro de RedALyC y AmeliCA, y de Carina Gordillo, de CAICYT, establecer una lista de modificaciones necesarias y obligatorias para los XMLs, y así pasamos las validaciones y que el sistema de SciELO las aceptara.

Sin embargo, pasada la euforia inicial del éxito, pero envalentonados por el mismo, nos surgió una nueva necesidad: automatizar el proceso para acortar tiempos. El motivo fue que modificar manualmente un XML proveniente de RedALyC o AmeliCA, nos llevaba casi tanto tiempo como marcar el artículo en el sistema de SciELO, y resultaba muy propenso a errores humanos, ya que implicaba la edición y escritura de un archivo en un lenguaje de programación que no todos los editores de la UNLP (sin mencionar Argentina) conocen, ni tienen por qué conocer.

Fue en ese momento que confirmamos el modelo de trabajo desarrollado en nuestra facultad, que tiene en cuenta la necesidad de incorporar a los equipos de gestión de la información una persona con capacidades en sistemas informáticos, ya que la idea del procesador estaba, pero no la podíamos llevar a cabo. Asimismo, Rozemblum y Banzato (2012, p. 94) hablan de la importancia del bibliotecario como eslabón para la edición y difusión de las revistas digitales y del conocimiento científico en general. También dicen que "la manipulación de la información para lograr interoperabilidad entre los sistemas propios y ajenos es una capacidad adquirida por los bibliotecarios en su formación profesional". Teniendo en cuenta esto, creemos de suma importancia

para esta labor la necesidad de un equipo de trabajo interdisciplinar que incluya un informático capaz de llevar a la práctica las ideas.

En una primera instancia decidimos realizar un procesador en hojas de cálculo. Estos programas proveen una interfaz conocida para la mayoría de los usuarios de computadoras, y ofrecen innumerables listados de funciones que permiten modificar contenido, independientemente de si es texto o número, por lo que de inmediato decidimos que era la herramienta adecuada para esta primera etapa del proyecto.

Funcionamiento del Procesador LuXMeL

El procesador denominado LuXMeL es un libro que posee tres hojas y que puede emplearse con la mayoría de los programas de hoja de cálculo. Se recomienda además contar con un editor de XML, preferentemente el NOTEPAD++, en el cual realizar los cambios y correcciones previas y posteriores al uso del procesador.

Hoja "XML a procesar"

Para procesar el archivo hay que corroborar previamente:

- Que no posea tabulaciones. [1]
- Que no posea filas en blanco
- Que no haya filas que no comiencen con "<" (en ciertos casos los XML insertan saltos de línea que hay que eliminar, en particular en las afiliaciones)
- Que no haya autores que no declaren su mail o que posean más de una afiliación.
- Que el artículo tenga las fechas de recibido y aceptado.

En ese orden, estos errores pueden invalidar el funcionamiento del procesador LuXMeL. Las tabulaciones impiden que se copie correctamente el XML original y deben eliminarse en tu totalidad reemplazándolas, a lo sumo, por espacios. Cada una de las otras dificultades tiene menos probabilidad de entorpecer el proceso que la anterior. De todas maneras, la hoja de "PROCESADOR" informa si se ha producido algún desfase de filas durante el procesamiento de los datos.

Una vez limpio y revisado el XML se debe copiar completamente y pegar en la celda A1 de la hoja "XML a procesar". Es importante nunca usar las opciones de copiar o cortar texto desde esta hoja, ya que puede dañar el procesador.

Hoja PROCESADOR

Esta hoja posee varias columnas ocultas con fórmulas; estas no deben alterarse de modo alguno. Luego, a partir de la columna R, se encuentran dos paneles descriptivos del proceso. Las columnas A a la P realizan las siguientes funciones:

A) Copia el contenido de la hoja "XML a procesar".
B) Chequea si en la fila A se encuentra un DOI.
C) Chequea las filas en las que se encuentra un "<contrib>" e indica el N° del autor.
D) Chequea si hay un mail y toma el N° de autor al que pertenece de la columna C; luego chequea si hay filas que indiquen el "<alt-text>" de las imágenes y marca esas líneas para eliminarlas.
E) Enumera la cantidad de líneas que deben ignorarse en total a medida que avanza el documento.
F) Copia las celdas de la columna A, pero salteando las de los mails de los "<contrib>" y de los "<alt-text>" de las imágenes.

G) Marca las líneas entre las que se encuentran las "<aff>" para poder incorporarles los mails si hace falta.
H) Marca las líneas en las que se deben incorporar los mails quitados de dentro de "<aff>".
I) Marca la cantidad de líneas que deberán saltearse una vez incorporados los mails en "<aff>".
J) Incorpora dentro de las líneas de "<aff>" la línea "Línea_para_MAIL_creada_por_LuXMeL" en los casos en los que corresponda agregar allí un mail.
K) Agrega los mails quitados de "<contrib>" dentro de "<aff>" según se indica en la columna J.
L) Controla si las dos líneas subsiguientes a "<history>" son de paginación, y, si no, las marca.
M) Corrige las dos primeras líneas del XML, y, si están marcadas en la columna K, agrega las líneas de paginación ficticia.
N) Busca la línea de "</counts>", y, si se incorporó paginación ficticia, marca que allí debe ir la línea de conteo de páginas.
O) Agrega la línea de conteo de páginas "<page-count count="02"/>" si corresponde paginación ficticia.
P) Para control, lista las líneas de la hoja "XML a procesar" que forman parte del XML, no están en blanco, y tampoco empiezan con "<".

Los paneles descriptivos informan, el primero, datos adicionales útiles para la edición del XML y el segundo sobre los errores o posibles dificultades durante el procesamiento. En cada caso, listan a su derecha las primeras 20 líneas en las que se encuentre cada caso, para poder ser consultadas y, si correspondiere, corregirlas en la hoja "XML a procesar". Los paneles son:

RESUMEN DE DATOS

- FILAS MÁXIMAS CORREGIDAS: La fila hasta la cual existen las fórmulas del procesador LuXMeL; si el total de filas del XML a corregir supera este total, el casillero se pondrá en rojo, para informar que no es posible procesar ese documento sin ampliar las fórmulas a más filas.
- FILAS TOTALES DEL XML: Controla la fila de la hoja "XML a procesar" en la que se encuentra el "</article> ".
- TOTAL DE AUTORES: Obtiene de la columna C el total de autores del XML.
- IMÁGENES: Obtiene de la columna D el total de imágenes cuyos deban eliminarse.
- SECCIONES SIN TIPIFICAR: Lista el total de secciones sin tipificar.
- TOTAL DE DOIs: Lista el total de DOIs mencionados en el XML, por si hiciera falta consultarlos.
- CORRESPONDE PAG. FICTICIA: Informa si el procesador LuXMeL ha detectado que debe incorporarse o no la paginación ficticia.
- PÁG. INICIAL FICTICIA: [2] El usuario debe completar manualmente la celda que indica el N° de página inicial que se desee crear para la paginación ficticia.

CONTROL DE ERRORES

- TABULACIONES: El procesador LuXMeL solo puede trabajar con la columna A de la hoja "XML a procesar". Cualquier parte del XML que, por tabulaciones, hubiera sido copiada en otra columna, no se verá reflejada en el resultado final. Cualquier XML que posea tabulaciones debe considerarse como erróneamente procesado.
- FILAS EN BLANCO: Varias fórmulas del procesador LuXMeL dependen del contenido de otras celdas. Al ser todas las celdas en blanco idénticas, es imposible distinguirlas entre sí y esto puede producir errores imprevisibles. No es recomendable que se carguen XML que contengan filas en blanco, ya que, de todas maneras, estas deben eliminarse antes de poder subir el archivo a SciELO.
- FILAS DUDOSAS: Simplemente enumera las celdas que poseen contenido y no empiezan con "<" en caso de que se deseare corroborar que sean correctas.

- AUTORES SIN EMAIL en "<contrib>": Lista los autores para los cuales no se encontró un email en el apartado de "<contrib>" ya que se ha detectado en versiones anteriores que, en algunos casos excepcionales, esto puede producir comportamientos inesperados de parte del procesador LuXMeL.
- ERRORES EN EL PROCESAMIENTO: Indica si se han producido errores inesperados durante el procesamiento y las columnas en las que se los detectó. Esto facilita determinar cuál puede llegar a ser el inconveniente con el XML cargado.
- PROCESAMIENTO EXITOSO: Busca en la hoja de Resultado la fila que contenga el "</article> " y calcula si es la que le corresponde de acuerdo a la cantidad de filas que el procesador haya eliminado y/o incorporado. Si es así, informa el total de líneas del XML corregido, y, si no, informa la cantidad de filas de desfase. Si el número de filas de desfase es positivo, es posible que se hubieran duplicado líneas por error. Si es negativo, es posible que se hubieran eliminado líneas que no correspondía eliminar.

Hoja Resultado

En la columna A de esta hoja se encuentra copiado el XML de la hoja "XML a procesar", con las correcciones realizadas mediante la hoja PROCESADOR. Si se copian todas hasta la fila de "</article>" en un archivo nuevo en el Notepad++, y se guarda dicho archivo con la extensión .xml (Extensible Markup Language), se obtendrá una copia corregida del XML requerido por SciELO Argentina.

Revisiones posteriores

Existen otros errores que pueden impedir la validación del XML, que no provienen de la incompatibilidad de los formatos utilizados por RedALyC/AmeliCA y SciELO, sino que se originan durante el marcado del artículo, es decir, que son errores humanos.

Por lo general estas faltas no influyen en el marcado ni en la publicación de los artículos en ninguno de sus formatos, ni en la plataforma de RedALyC/AmeliCA ni en un OJS de uso particular, ya que simplemente se trata de la ausencia de determinadas etiquetas en las referencias (producto de la omisión durante el marcado, o un marcado equivocado). Las etiquetas que tienen que figurar obligatoriamente son:

- Las referencias tipo *book* deben tener año de publicación "**<year>**". En el caso de la etiqueta "<year> " es muy importante, además, verificar que la fecha de publicación del documento referenciado no sea mayor al año de publicación del artículo (algo lógico, pero que a veces no tomamos en cuenta cuando publicamos números con retraso).
- Las referencias tipo *webpage* deben tener un URL etiquetado como ""**<ext-link>**".
- Las referencias tipo *journal* no pueden omitir el título del artículo "**<article-title>**".
- as referencias tipo *confproc* (Congreso/Conferencias) están obligadas a tener la etiqueta "**<conf-name>**" o nombre de la conferencia.
- Cuando la referencia completa tiene información sobre el doi, hay dos formas de procesarlas:

  Si el doi no tiene URL, sino solamente el código doi:

  a) En Referencia Completa se deja igual, y en Referencia detallada se usa la etiqueta "**<pub-id pub-id-type="doi">10.....</pub-id>**"

  Si el doi aparece con un enlace URL en la referencia:

  a) En Referencia Completa se coloca un enlace externo al URL "**<ext-link> ... </ext-link>**"

En Referencia detallada se utiliza:
"**<pub-id pub-id-type="doi">10.....</pub-id>**", pero también se agrega "**<ext-link> ... </ext-link>**".

Es indispensable también, dentro del grupo de referencias, revisar que la gramática del lenguaje no tenga errores, particularmente casos de doble entrecomillado, como en el ejemplo que compartimos a continuación:

```
<xref ref-type=""bibr"" rid=""redalyc_350554796006
_ref25"">Kirchof, 2016,  p. 213</xref>
```

En este caso se ven resaltadas con color las dobles comillas.

Por otro lado, en secciones con tipología definida (como puede ser Introducción, Metodologías, o Resultados), el validador de SciELO exige que estén marcadas dentro de la etiqueta que precede a cada sección "<sec **sec-type="..."**>". En la tabla 1 pueden observarse las tipologías que SciELO provee en su página web (SciELO, 2017), en la columna izquierda se puede ver la denominación para cada tipo de sección, y en la derecha, la explicación de dicho tipo.

TABLA 1
Sec-Type de SciELO

cases	Cases/Case Reports
conclusions	Conclusions/Comment
discussion	Discussion/Interpretation
intro	Introduction/Synopsis
materials	Materials
materials|methods	• Materials and Methodology
methods	Methods/Methodology/Procedures
nd	undefined
results	Results/Statement of Findings
results|conclusions	• Results and Conclusions
results|discussion	• Results and Discussion
results|discussion|conclusions	• Results, Discussion, Conclusions
subjects	Subjects/Participants/Patients
supplementary-material	Supplementary materials

Fuente: SciELO, 2017

A modo de cierre

Junto con este artículo, dejamos este procesador, para aquel a quien le interesare utilizarlo, en dos formatos diferentes: un archivo .ods (para LibreOffice Calc), [3] y un archivo .xlsx (para Excel 2016 en adelante).

Luego de varias pruebas, pudimos definir el tiempo promedio de procesamiento de un XML –desde el momento en que se baja de la plataforma de RedALyC o AmeliCA, hasta que está listo para enviar a SciELO Argentina– en aproximadamente veinte minutos, si se tiene en cuenta que un XML que está perfectamente marcado y no requiere ningún tipo de intervención posterior se procesa en menos de cinco minutos, y uno que necesita varias modificaciones (particularmente en la sección de referencias) puede llevar hasta cuarenta minutos.

Como ejemplo, podemos decir que el mismo proceso realizado a mano en un número de una revista, con 10 artículos con un marcado básico de referencias, nos llevó aproximadamente 28 horas, y con la herramienta lo redujimos a 8 horas de trabajo.

Este procesador no es, de ninguna manera, la forma definitiva que buscamos aplicar para sortear este problema. Mientras esperamos por una unificación desde los sistemas para la utilización de los XML, estamos trabajando en una aplicación o programa que genere los XML automáticamente y que acorte los tiempos aún más, sin necesidad de un programa externo.

Sin embargo –dada la necesidad de una solución que sea rápida–, nos otorga la posibilidad de cumplir con las continuidades de las revistas en SciELO, mientras se trabaja en una herramienta más apropiada, o, al menos, más elegante.

Referencias

Babini, D. (2019). La comunicación científica en América Latina es abierta, colaborativa y no comercial. Desafíos para las revistas. Palabra Clave, 8(2), e065. https://doi.org/10.24215/18539912e065

Banzato, G., y Rozemblum, C. (2019). Modelo sustentable de gestión editorial en Acceso Abierto en instituciones académicas. Principios y procedimientos. *Palabra Clave, 8*(2), e069. https://doi.org/10.24215/18539912e069

SciELO (2017). *SciELO PC Programs' Documentation. List of Codes*. Recuperado de https://scielo.readthedocs.io/projects/scielo-pc-programs/en/latest/code_database.html#sec-type

Rozemblum, C., y Banzato, G. (2012). La cooperación entre editores y bibliotecarios como estrategia institucional para la gestión de revistas científicas. *Información, cultura y sociedad*, 27, 91-106. Recuperado de http://www.m emoria.fahce.unlp.edu.ar/art_revistas/pr.5536/pr.5536.pdf

Notas

1 Para eliminar las tabulaciones de manera eficaz, sugerimos usar la opción de reemplazar tabulaciones por espacios del Notepad++: se selecciona todo el texto se elige *Edit → Blank Operation → TAB to Space*.

2 Debido a las publicaciones digitales y los varios formatos, la paginación clásica que se arrastra desde las versiones impresas ha ido quedando en desuso; actualmente se aplican nuevas paginaciones digitales que denominan al artículo en sí. Dado que SciELO utiliza aún la paginación tradicional en su código, es necesario para aquellas revistas que ya no la tienen, generar una paginación ficticia de los artículos. En la FaHCE, optamos por otorgarle una numeración consecutiva de 2 páginas en total, haciendo que el primer artículo del número sea 1-2, el segundo 3-4, etc.

3 La extensión .ods también sirve para archivos de Open Office, pero al estar este paquete descontinuado, no podemos tener la seguridad de que funcione perfectamente.