FACULTAD DE COMUNICACIÓN E INFORMACIÓN

XML-JATS (Journal Article Tag Suite)
Características y utilidad para la edición de revistas académicas en versión electrónica

Sesión sincrónica | 24 de abril de 2024 (10:00 a.m.)

Carlos Suárez Balseiro
Coordinación Latindex (Puerto Rico)
Seminario de Editores Académicos (UPRRP)
Facultad de Comunicación e Información
Universidad de Puerto Rico, Recinto de Río Piedras

Centro para la
Excelencia Académica

UPRRP

➢ Explicar las características del estándar JATS (*Journal Article Tag Suite*)

➢ Discutir la utilidad de utilizar XML-JATS para el trabajo con revistas académicas en versión electrónica

Organización de la información en el medio electrónico

Documents, Metadata, and Links

The thing which you get when you follow a link, when you de-reference a URI, has a lot of names. Formally we call it a **resource**. Sometimes it is referred to as a document because many of the things currently on the Web are human readable documents. Sometimes it is referred to as an object when the object is something which is more machine readable in nature or has hidden state. I will use the words document and resource interchangeably in what follows and sometimes may slip into using "object".

One of the characteristics of the World Wide Web is that resources, when you retrieve them, do not stand simply by themselves without explanation, but there is information about the resource. Information about information is generally known as **Metadata**. Specifically, in the web design,

Definition

Metadata is machine understandable information about web resources or other things

The phrase "machine understandable" is key. We are talking here about information which software agents can use in order to make life easier for us, ensure we obey our principles, the law, check that we can trust what we are doing, and make everything work more smoothly and rapidly. Metadata has well defined semantics and structure.

(Berners-Lee, 1997)

[…] due to the ever increasing amount and diversity of information objects, changes in technologies, and growing expectations from users to access instant, reliable and complete access to information, libraries are facing an unprecedented challenge to rise **with the need to fully and semantically enrich information objects with appropriate metadata so as to improve the findability and discoverability of information objects** […] (Alemu et al., p. 312)

The idea of providing richer metadata is quite important because, **if the metadata is not built accurately (e.g., in a manner that employs users' terminologies) then information objects, could remain undetected or obscure, and hence, inaccessible, even though they could be present in library databases or on their shelves**. Such an outcome clearly has negative implications for the fulfillment of the goals of digital libraries. (Alemu et al., p. 338)

XML (eXtensible Markup Language)

➢ XML is an open standard that is used to serialize (i.e., encode and describe) structured data and to facilitate the maintenance, organization, sharing, and reuse of these data by computer applications. (Cole & Han, 2013)

➢ Es un metalenguaje que permite la creación de lenguajes de marcado para ser utilizados en diferentes contextos con diferentes tipos de información.

¿Cuál es su origen?

➢ **Década de los 60**. GML (*Generalized Markup Languaje*). IBM Research crea el primer lenguaje de marcado, que se pudo utilizar para crear otros lenguajes. Su desarrollo llevó al metalenguaje SGML (*Standardized Generalized Markup Languaje*) como un estándar internacional (1986)

➢ **Década de los 90.** HTML (*Hypertext Markup Languaje*). Se desarrolla HTML como aplicación directa de SGML

➢ **Década de los 90.** XML (*eXtensible Markup Language*). El Consorcio Web (W3C) busca un lenguaje que combinara la flexibilidad de SGML y la amplia aceptación de HTML. A finales de la década se crea la primera versión de XML
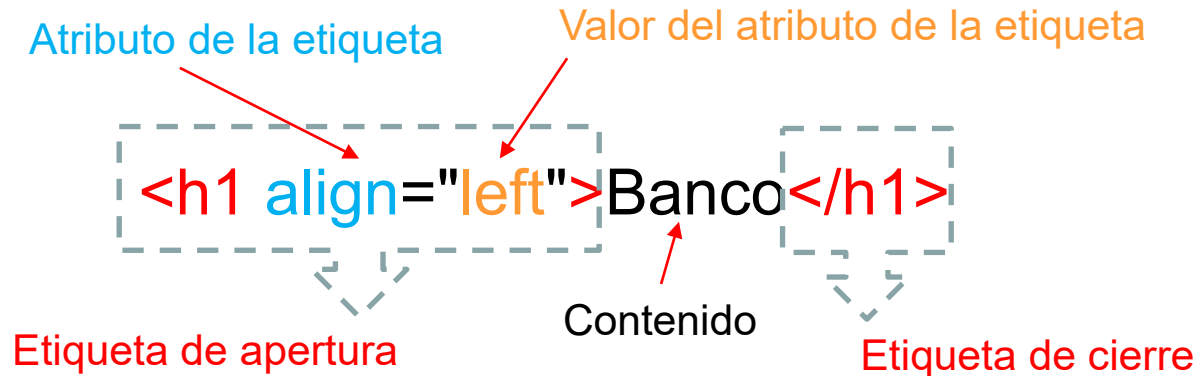
¿Cuáles son sus características?

➤ Extensible: Permite crear lenguajes derivados, con la inclusión de etiquetas (*tags*) definidas por el usuario (hay esquemas XML para diferentes contextos).

➤ A diferencia de HTML, hace marcado estructural, se enfoca en el contenido del documento, en el significado del metadato y no en su presentación o apariencia (los estilos se trabajan con las denominadas *Cascading Style Sheets o XSL (eXtensible Stylesheet Language*))

HTML y XML (apariencia y significado)

Atributo de la etiqueta

Valor del atributo de la etiqueta

`<h1 align="left">Banco</h1>`

Etiqueta de apertura

Contenido

Etiqueta de cierre

`<h1 align="left"> <font family="sans-serif" size="22" color=Red> <b><i>Banco</i></b> </font> </h1>`

(ver en w3school:
https://www.w3schools.com/html/tryit.asp?filename=tryhtml_default)

➢ **¿Qué se interpreta de esta codificación?**
Es un encabezado alineado a la izquierda, con tipo de fuente *sans-serif*, de tamaño 22 y color rojo, en negrita e itálica, cuyo contenido es "Banco"

➢ **¿Qué significa "Banco"?**: ¿Un mueble para sentarse, una institución bancaria, un banco de peces en el mar?

HTML (Apariencia)

```
<!DOCTYPE html>
<html>
<head>
<title>Revista A</title>
</head>
<body>
<h1 style="font-size:300%; text-align:center">Editorial</h1>
<p style="text-align:right"><i>Autor: Carlos Suárez, FaCI, Universidad de Puerto Rico, Río Piedras</i></p>
<p>Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book.</p>
</body>
</html>
```

Editorial

Autor: Carlos Suárez, FaCI, Universidad de Puerto Rico, Río Piedras

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<section>
<title>Editorial</title>
<author>
<name> Carlos Suárez</name>
<affiliation>FaCI, Universidad de Puerto Rico, Río 'Piedras</affiliation>
</author>
<content>Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book.</content>
</section>
```

Esquemas en diferentes contextos

➢ CML (*Chemical Markup Languaje*, http://www.xml-cml.org),lenguaje XML que permite la representación de estructuras químicas.

➢ MathML (*Mathematical Markup Languaje*, http://www.w3.org/Math) , lenguaje XML que permite la representación de expresiones matemáticas.

➢ MEDLINE®/PubMed® XML Data Elements (https://www.nlm.nih.gov/bsd/licensee/data_elements_doc.html)

➢ JATS (*Journal Article Tag Suite*, https://jats.nlm.nih.gov/), estándar técnico (versión 1.3 ANSI/NISO Z39.96-2021 https://www.niso.org/standards-committees/jats), una iniciativa que define un formato XML para describir estructura, semántica y metadatos para contenido digital científico.

JATS (Journal Article Tag Suite)

¿Qué es JATS (Journal Article Tag Suite)?

- JATS [2003], formally known as ANSI/NISO Z39.96-2015 JATS: Journal Article Tag Suite, is an international standard XML tag set for journal articles. JATS is an XML vocabulary designed to model current journal articles.

- JATS is a named collection of XML elements and attributes that can be used to mark the structure and semantics of a *single journal article*. Thus JATS does not model issues of journals, books, language corpora, patents, legislation, standards, or other document types.

- Originally, JATS was used for STEM (Scientific, Technical, Engineering and Medical) articles, but now journals in the humanities, sociology, economics, and the soft sciences also use JATS XML markup.

(Lapeyre, 2018)

Uso y alcance

- JATS is in use in more than 25 countries world-wide.

- Many Public and private archives accept (or require) JATS, including: PubMed Central (both US and UK), British National Library, Australian National Library, US Library of Congress, Portico/ITHAKA/JSTOR, and many more.

- Conversion vendors worldwide have developed processes for and gained significant experience in converting into and out of JATS. Using a shared tag set means that conversion vendors do not need to learn/customize a journal tag set for each publisher. Tools have been developed to make JATS out of many formats, including Microsoft Word and LaTeX, and to make quality PDF, HTML (HTML, XHTML, HTML5), various accessible formats, and eBooks out of JATS.

(Lapeyre, 2018)

Características

Create Once, Publish Everywhere
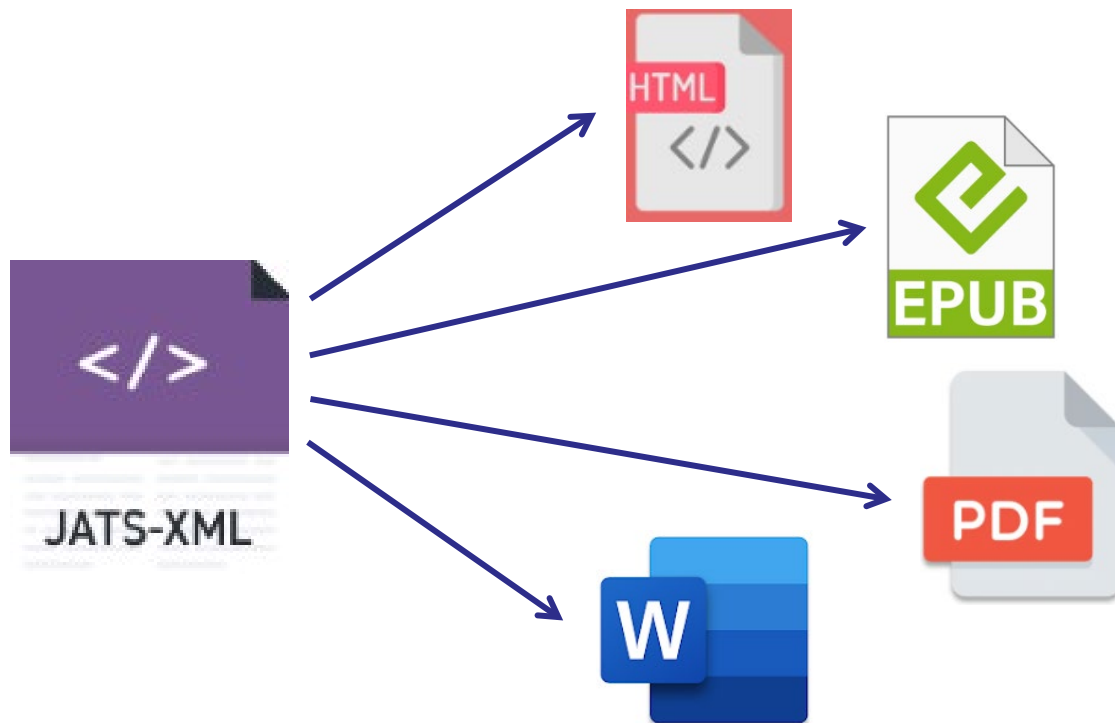
- Criterio 32 de metodología de Latindex.
  Uso de diferentes formatos de edición
    - Para calificar, la revista debe usar más de un formato de edición para el despliegue y lectura de los textos completos de los documentos publicados.
    - Nota de aplicación: La revista puede elegir los formatos de su preferencia según el público lector. Cada documento debe tener más de un formato. Se busca ampliar el uso del PDF, el uso de HTML y otros formatos que faciliten el hipertexto y la lectura.

    Ver: https://alldocs.app/convert-jats-xml-to-html

Características

JATS was originally constructed for interchanging journal articles, providing interoperability of article content and article metadata among publishers and archives. In the early days, it was expected that publishers, hosters, portals, and archives would use their own XML tag set internally, and transform into JATS XML when they wanted to:

➢ exchange information with each other,

➢ put information into a combined repository,

➢ sell/display items on the same hosting platform, and/or

➢ share the development of tools and resources for common use.

(Lapeyre, 2018)

Estructura

Front Matter (<front>)

    journal-level metadata
(journal title and identifiers)

    article-level metadata
(article title, author(s), identifiers like a DOI)

Body Matter (<body>) the narrative text of the article, including, for example:

    paragraphs

    sections

    figures and graphics

    Tables

    equations and quotations

Back Matter (<back>)

(Lapeyre, 2018)

    appendices

    bibliographic reference lists with deeply detailed (but completely optional) citation metadata

Estructura

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE article PUBLIC "-//NLM//DTD JATS (Z39
    .96) Journal Publishing DTD v1.3 20210610
    //EN" "https://jats.nlm.nih.gov/publishing/1
    .3/JATS-journalpublishing1-3.dtd">
<article article-type="research-article" dtd
    -version="1.3" xml:lang="es"
    xmlns:mml="http://www.w3.org/1998/Math
        /MathML"
    xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema
        -instance"
    xmlns:ali="http://www.niso.org/schemas/ali/1
        0/">
<front>⟷</front>
<body>⟷</body>
<back>⟷</back>
</article>
```

➢ **Front**. Contiene los metadatos del artículo (título del artículo, revista en la que se publica, fecha, tipo de publicación, copyright, etc.). Es información bibliográfica del artículo y la revista en la que es publicado.

➢ **Body**. Es el principal contenido textual y gráfico del artículo. Consiste en párrafos y secciones, que pueden incluir figuras, tablas, citas, etc.

➢ **Back**. Contiene información que auxilia al texto principal, como es el glosario, apéndice, y listado de referencias bibliográficas.

Fuente: Redalyc.org

Ver: https://jsonformatter.org/xml-editor

Metadata for Searching, Mining, Extraction

JATS can (always optionally) provide rich metadata needed for semantic interchange, context-sensitive search, and data mining.

- article and article component identifiers such as DOIs and arXiv numbers,
- author identifiers such as ORCIDs,
- institution identifiers (e.g., Ringgold, ISNI), and
- funder identifiers
- article title (with alternative languages if desired);
- article identifiers (DOI, publisher identifiers, etc.);
- copyrights and permissions;

(Lapeyre, 2018)

Metadata for Searching, Mining, Extraction

JATS can (always optionally) provide rich metadata needed for semantic interchange, context-sensitive search, and data mining.

- finding aids such keywords, terms, and subject classifications (which may be given in multiple languages and linked to ontologies/taxonomies), and

- detailed descriptions of the funding behind the research on which the article reports, including:
    - both monetary and non-monetary funding
    - award and grant identifiers
    - funding sources and principals

- links to companion articles or resources (different language versions, errata, updates, related books or web pages);

(Lapeyre, 2018)

Contributors

- unique external identifiers
- the ability to record both initials and full names
- the ability to have only a given name, without a surname, as is the custom in some southeast Asian countries
- alternative names for one individual (both Japanese characters and a romanized version, for example)
- description of the role or roles played by the person (possibly linked to CRediT taxonomy terms)
- multiple affiliations (with alternative names for multiple languages);

(Lapeyre, 2018)

- Detailed bibliographic references for an article allow publication tracking and help establish scientific relevance through citation. The named components of a reference allow software to check references with CrossRef, figshare or other DOI repository to improve reference accuracy.

- JATS can support nearly any type and style of reference tagging including data citation (with access times). For citing journal articles, specific elements include contributor names, article title, journal title and issue, first and last page numbers, day-month-year of publication, DOI, and more. Similarly, JATS has named elements sufficient to cite books, patents, conference papers, standards, and other works.

(Lapeyre, 2018)

XML-JATS en servicios de indización y resumen

➢ PubMed Central Tagging Guidelines
(https://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/article/genprac.html)

➢ Sistema de Información Científica Redalyc
(https://www.redalyc.org/postulacion.oa?q=criterios)

➢ CrossRef Metadata Search
(https://search.crossref.org/)

Consideraciones finales

➢ XML-JATS es una poderosa combinación de XML y JATS, que permite la estandarización e intercambio de información científico-técnica en un formato estructurado y legible por las máquinas.

➢ Proporciona numerosos beneficios, incluida la consistencia, interoperabilidad, conservación y reutilización de la información.

➢ El uso de XML-JATS contribuye al completamiento y enriquecimiento de los metadatos, la precisión de los mismos y la legibilidad de la indización de artículos científicos en el medio electrónico, lo que influye favorablemente en la visibilidad y el aprovechamiento de los resultados de investigación.

Referencias

Alemu, G., Stevens, B., Ross, P., & Chandler, J. (2012). The social space of metadata: Perspectives of LIS academics and postgraduates on standards-based and socially constructed metadata approaches. Journal of Library Metadata, 12(4), 311-344. https://doi.org/10.1080/19386389.2012.735523

Berners-Lee, T. (January 6, 1997). Metadata Architecture. https://www.w3.org/DesignIssues/Metadata.html

Cole, T. W., & Han, M.-J. K. (2013). XML for catalogers and metadata librarians. Libraries Unlimited. https://search.worldcat.org/en/title/862370898

Lapeyre, D. A. (2018, October 12). Introduction to JATS (Journal Article Tag Suite). XML.com https://www.xml.com/articles/2018/10/12/introduction-jats/

Redalyc. Sistema de Información Científica Redalyc (https://www.redalyc.org/)

W3C Extensible Markup Language (XML) https://www.w3.org/XML/)

W3C XML Essentials (https://www.w3.org/standards/xml/core)
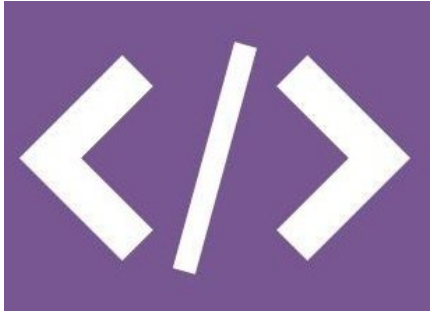
XML-JATS (Journal Article Tag Suite)

Características y utilidad para la edición de revistas académicas en versión electrónica

Carlos Suárez Balseiro

Facultad de Comunicación e Información

Programa Graduado de Ciencias de la Información

Universidad de Puerto Rico, Recinto de Río Piedras

10 Ave. Universidad STE 1001, SAN JUAN PR 00925-2530

Tel. 787-7640000 ext. 85335

faci.uprrp.edu | carlos.suarez5@upr.edu

FACULTAD DE COMUNICACIÓN E INFORMACIÓN

Centro para la Excelencia Académica

UPRRP   LA IUPI