XML Publishing and PKP

PKP Barcelona 2019

Who are we?

Vitaliy Bezsheiko:

- Senior Editor for Psychosomatic Medicine and General Practice
- Grobid improvements, JATS Parser Plugin, DOCXTOJATS plugin, more

Dulip Withanage:

- Developer for Heidelberg University Publishing
- meTypeset, heiMPT, OJS typeset & lensGalleyBitsPlugin, Texture

James MacGregor:

Coalition Publica ops/technical teams, PKP|PS

Who are we?

Substance Consortium: Substance, eLife, EMBO/Sourcedata, Érudit, PKP, SciELO, Substance

Coalition Publica: PKP, Érudit

Semantic Extraction Group

Friends of Manuscripts Press Group

JATS4R (JATS for Reuse) Working Group

Where are we now?

We are now working on a more distributed/decentralized ecosystem:

- Creation/conversion: DOCXTOJATS, Grobid, meTypeset, from scratch
- Editing/Typesetting: Texture
- Publishing/Display: Lens Reader, JATS Parser Plugin

What follows is a discussion and demonstration of ***in-progress*** developments.

Not all components are complete yet!

Structure of the session

XML workflow demonstration:

- Conversion
- Editing
- Typesetting/publishing tools

Open Discussion:

- Your use cases and questions
- Timelines, next steps

Demo/playtime!

Conversion: Grobid

Grobid is a machine learning software developed for data extraction from scientific articles in PDF format. Originally aimed at metadata extraction, it’s now also capable to extract article’s full text. The main problem of extraction of any data from PDF is that it’s only “visible” as raw and unstructured text. Grobid with machine learning techniques allows to put this data into a structure, like sections, tables, citations, etc. Grobid receives PDF files as an input and converts them to TEI XML, which has similar to JATS XML structure.

Conversion: Grobid

Current development plans:

- Enrich data extraction for fulltext model:
  - Lists. The support for lists is already implemented and pushed on the dev GitHub branch.
  - Tables. Considering the best approach to determine the table structure (rows and cells):
    - Tabula integration. Already implemented, more training data is needed to evaluate accuracy or
    - Determine structure based on PDFAlto results.
  - Blockquotes.
- Provide more annotated training data for machine learning models.
- Develop a converter from TEI XML (Grobid's output) to JATS XML or consider integration of an existing one.

Conversion: Grobid

Next steps

- Integration tool like the Typeset Integration Plugin, so that Grobid can be used directly in OJS.
- More data!

Conversion: DOCX to JATS XML Converter Plugin

The plugin converts DOCX files which correspond to OOXML format to JATS XML. It's based on a library written in PHP and doesn't have any other dependencies. The output is compatible with the Texture Plugin. As an input can be used DOCX files produced by MS Word, LibreOffice Writer and Google Docs.

Current development plans:

- Extend the support for article elements, like figures, formulas, citations.
- Integration with tools that can help to extract data that aren't regularly present in OOXML.

Conversion: DOCX to JATS XML Converter Plugin

Differences from Grobid (besides input format):

- Doesn’t require anything beyond PHP interpreter.
- The output is restricted to how data is structured in the source DOCX file, although should work well in most of the cases.
- Other limitations for XML to XML conversion.

Readme and demo: https://github.com/Vitaliy-1/docxConverter/blob/master/README.md

Online demo:

- Article: http://xml.test.publicknowledgeproject.org/index.php/docxtojats/article/view/1
- Backend: http://xml.test.publicknowledgeproject.org/index.php/docxtojats/workflow/index/1/5

Conversion: meTypeset command line tool

Description

1. Docx / Openoffice -> TEI XML -> JATS converter
2. Python3 based command line tool with saxon XSLT processor
3. Configurable individual steps
4. Chains XSLT stylesheets
5. Reference parsing (automated / interactive)
6. Zotero/Mendeley integration to extract structured references
7. Stable, Used in monograph production for MS-Word
8. Integrated regression test-suite

Conversion: meTypeset

Current status

- development partners still contribute - no actively known community
- MS-Word feature extensions only on request
- Complex stylesheet structure : enhancement requires knowledge in python and xslts

Official branch: https://github.com/MartinPaulEve/meTypeset

Active branch: https://github.com/withanage/meTypeset

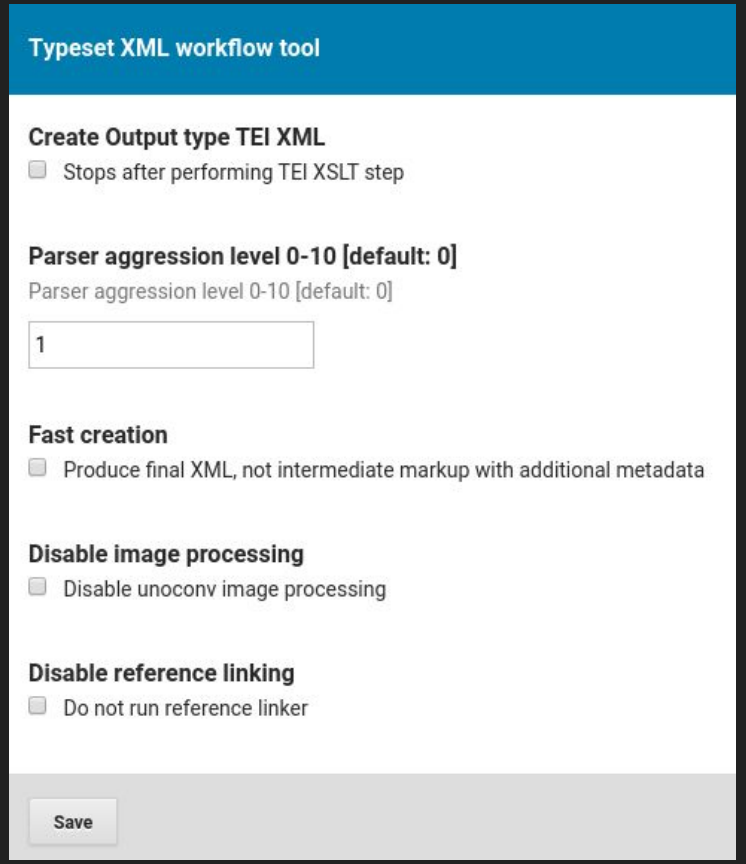
Conversion: ojs typeset plugin

Description

1. Generic plugin for setting a command line tool as a conversion utility
2. Easily extendable for any XML converter (not restricted to JATS)
3. Tool settings can be controlled in the plugin settings
4. Word, Openoffice as Input formats
5. TEI , JATS as output formats
6. Supports : OJS Stable 3.1.2 +
7. Next : Support OMP 3.1.2 +

- http://xml.test.publicknowledgeproject.org/index.php/typeset/workflow/index/3/5

https://github.com/withanage/typeset



Editing: Texture

Functionality

- Texture is a javascript based XML editor for JATS XML
- Integrated in OJS as a plugin in plugin gallery
- Both Standalone and integrated Editing software environment
- DAR as base - Document archive format for research
- Visual Editing, Semantic tagging
- Supports images as dependent files
- New Feature : creates a galley from a given XML file and it’s dependent files

Editing: Texture

Demo

- http://xml.test.publicknowledgeproject.org/index.php/typeset/workflow/index/3/5

Editing: Texture

Current development plans

- Support most-used JATS tags in XML- publishing workflows
- Support for uploading Zipped offline DAR archive
- Integrate UI/UX developments of Texture into OJS
- Discussion with partners to establish a standardized editor-friendly machine readable JATS subset for XML editors for easy machine learning
- Non-Editable areas for metadata fields from Content providers
- Relax strict validation (in discussion with major partners)
- Collaborative editing

Publishing: LensGalleyBits Reader

Description

- XML based web-reader (native support for JATS)
- Early implementation by eLife, forked and extended.
- Client-side rendering, no extra libraries
- Additional support for a subset of BITS for monographs and edited volumes
- Limited mobile support : iPad and above
- Standalone application for offline production preview
- Available in OJS plugin-gallery

Publishing: LensGalleyBits Reader

Demo

- http://xml.test.publicknowledgeproject.org/index.php/typeset/index
- Real world example of an edited volume in heiUP

OJS Plugin

- https://github.com/withanage/lensGalleyBits

Lens source code

- https://github.com/withanage/UBHD-Lens#implemented-extensions

Publishing: LensGalleyBits Reader

Current development plans.

- BITS support constantly improved
- Integration of annotations in the navigation panel
- DOI- generation for annotated paragraphs
- Customizing using journal-based stylesheets
- Support newer features from texture

Publishing: JATS Parser Plugin

JATS Parser Plugin is aimed to convert JATS XML to HTML and PDF and present the article on the front-end. It can be divided mainly into 2 parts:

- JATS Parser library that is written in PHP, thus converts documents on the server side. Current output is HTML (JATS XML -> PHP Objects -> PHP DOM) and PDF produced with TCPDF. Currently, it parses body and references from a given JATS and retrieves article's meta from OJS.
- Front-end part, integrated into the plugin. It relies on Bootstrap 4 and is mobile-friendly.

Requirements: PHP 7.2 or newer, theme that supports Bootstrap 4 (Classic, Immersion, Health Sciences).

Publishing: JATS Parser Plugin

Current development plans:

1. Ensure full compatibility with the output from Texture Plugin.
2. Add compatibility with all OJS themes.
3. Option to display JATS XML on article landing page rather than as a separate galley.
4. Testing and first production release.
5. Extend the support for more JATS XML elements, like formulas and footnotes.
6. Consider the ability to display references from JATS in different citation styles, currently only AMA (similar to Vancouver style).
7. Option to customize PDF output. TCPDF has limited support for styling of produced PDFs, although it’s the fastest and most lightweight library.

Demo: HTML, PDF

Next Steps

1. Release current tools (some maybe still in “beta”)
2. Develop a common test framework for each toolset
    a. Eg., for all conversion tools, develop a “supported elements” matrix
    b. Formalize testing and reporting of results
3. Continue to participate in XML Publishing Community
    a. GROBID, Texture/Libero Editor partnerships
    b. JATS4R Working Groups
    c. Informal community discussion groups
    d. Settle on formal interoperability standards for all applications
4. Publish timeline for project as a whole
    a. Ongoing toolset production releases
    b. Possible service provisions

Resources

GROBID: https://github.com/kermitt2/grobid, https://github.com/Vitaliy-1/grobid/

Texture: https://github.com/substance/texture/, https://github.com/withanage/texture/

meTypeset: https://github.com/MartinPaulEve/meTypeset, https://github.com/withanage/metypeset

Typeset plugin: https://github.com/withanage/typeset

DOCXConverter Plugin: https://github.com/Vitaliy-1/docxConverter

JATSParser Plugin: https://github.com/Vitaliy-1/JATSParserPlugin/

LensGalleyBits Plugin: https://github.com/withanage/lensGalleyBits

Who's who in JATS 2019 (Marc Bria): https://forum.pkp.sfu.ca/t/who-is-who-in-jats-2019/57063

Thank you!

Questions?

James MacGregor: jbm9@sfu.ca

Dulip Withanage: dulip.withanage@gmail.com

Vitaliy Bezsheiko: vitaliybezsh@gmail.com

Test Information: email to jbm9@sfu.ca, or try:

DOCX to JATS journal (DOCX to JATS, Texture, JATS Converter)

- http://xml.test.publicknowledgeproject.org/index.php/docxtojats
- User: pkpadmin / pkpadminpkpadmin

Typeset journal (Typeset, Texture, LensGalleyBits)

- http://xml.test.publicknowledgeproject.org/index.php/typeset/
- User: pkpadmin / pkpadminpkpadmin

Grobid test instance (web interface)

- http://sfulib8.nmsrv.com:8070