Fluid Article Production: Using AI to Automate XML Creation

Marianne Calilhanna, DCL

1 INTRODUCTION

Consistently good XML post-peer review, delivers advantages to publishers and authors alike.

- streamlines production & single-source management to improve quality
- improves discoverability during production & post-publication
- supports research integrity and AI initiatives
- maintains metadata from peer review process

2 METHODS

DCL architected an automated workflow engine that invokes machine learning algorithms, natural language processing, and named entity recognition (NER) to transform Word-based manuscripts to valid JATS XML.

The system is architected to

- ✓ invoke or bypass activities based on content
- ✓ maintain metadata, status, metrics, issue alerts, and updates
- ✓ improve over time as machine learning and rules-based algorithms are updated
- ✓ provide management reporting and full transparency

Manual intervention comes into play

- ✓ if a manuscript fails a workflow step

DCL
Data Conversion Laboratory Inc.

The dynamic synergy of AI and XML improves the speed of delivering accurate research to the scholarly community.

www.dataconversionlaboratory.com

3 CHALLENGES

Journal articles typically follow a standardized format but variations in layout and style pose challenges for automated processing. Variations in manuscripts hinder the identification and categorization of content components such as headings, paragraphs, citations, tables, and figures.

- ✓ authors are inconsistent with Word styling
- ✓ tables often have a high diversity of layouts, complicated elements such as straddle headings, various textual alignment, the presence of empty cells, and other intricacies
- ✓ references are not uniformly styled and typos in authors' names are ubiquitious
- ✓ equations and chemical symbols are often styled as images embedded in a Word file

4 RESULTS

The system produces an XML file that is valid to JATS 1.2 and adheres to any XML style points required by a publisher.

10 mins average turnaround time from manuscript to XML

40% average reduction in time spent manually updating XML and styling

80% of manuscripts auto-converted and validated with 100% accuracy over time

95% of manuscripts converted with 100% accuracy as machine learning improves with usage

Society for Scholarly Publishing
2024 Annual Meeting
Poster Session

Fluid Article Production: Using AI to Automate XML Creation

Data Conversion Laboratory Inc.

10-minute turnaround time

START
Peer-reviewed manuscript

1 Receive & pre-process

This initial workflow step unpacks the submitted files and securely saves and registers assets and corresponding metadata.

- ✓ Convert non-native math to MathML
- ✓ Convert native Word math to MathML
- ✓ Correct malformed URLs
- ✓ Normalize lists
- ✓ Accept all track change revisions
- ✓ Insert page number bookmarks
- ✓ Handle line breaks
- ✓ Extract & convert images to PNG
- ✓ Convert to HUB with the OASIS table model
- ✓ Convert Word table properties to XHTML/CSS
- ✓ Represent tables as HTML
- ✓ Handle inconsistent table columns, row separators
- ✓ Normalize fields such as symbols to XML-style grammar

2 Autostyle

The autostyle step applies consistency and styling that further enables congruous conversion to XML meeting.

- ✓ Link graphics & graphic captions
- ✓ Handle layered graphics
- ✓ Style captions
- ✓ Accommodate page boundaries
- ✓ Log anomalies such as unconnected drawings
- ✓ Convert tables to CALS model
- ✓ Tag tables with attributes to denote specific stylistic choices authors used
- ✓ Autostyle based on content structure
- ✓ Employ spaCy, industrial-strength natural language processing in Python, is used to detect and autostyle author names & affiliations

3 Conversion

Previous steps contribute to the successful conversion to XML. Verifying autostyling actions and repairing anomalies is the first process.

- ✓ Autostyling verification & repair
- ✓ Author/affiliation linking
- ✓ Special characters handling
- ✓ Table footnotes
- ✓ Retention & propogation of table colors, shading, bolding
- ✓ Bibliographic reference decomposition
- ✓ Figure citations
- ✓ Reference citations
- ✓ Appendix citations
- ✓ Conversion to the project-specific version of the JATS 1.2 XML schema
- ✓ Populating metadata

4 Confidence Analytics

Automated QC checks are performed and the results of the checks are analyzed and weighed to provide a confidence factor and automated recommendation for manual review or straight-through processing for each document.

Low confidence factor
- ✓ Manual review

High confidence factor
- ✓ Straight-through processing
- ✓ Onto next step in automated workflow

5 Editorial

The editorial stage involves human intervention from offshore and onshore resources. Combined review supports time and budget concerns that are factors when architecting a workflow that is in a constant state of movement.

- ✓ Review limited to only the elements the system highlights for review and repair
- ✓ Global QC performed upon receipt of updated XML
- ✓ Verification made that only identified elements were modified

6 Quality Assurance

A model of continuous improvement that both allows the correction of issues prior to final delivery and proactively prevents those issues in the future.

- ✓ Automated differencing
- ✓ Error analytics
- ✓ XML-based QC checks

Machine-learning algorithms evolve & rules-based algorithms are updated with ongoing usage.

7 Final Packaging

Final QC checks performed
- ✓ Verified files (XML & images) zipped and prepped

8 Delivery

API call and delivery to target. Report confirms that all handshakes were successfully executed and that no files were dropped upon retrieval and return.

Customer API
- ✓ Invokes DCL GetJob API
- ✓ Retrieves packaging