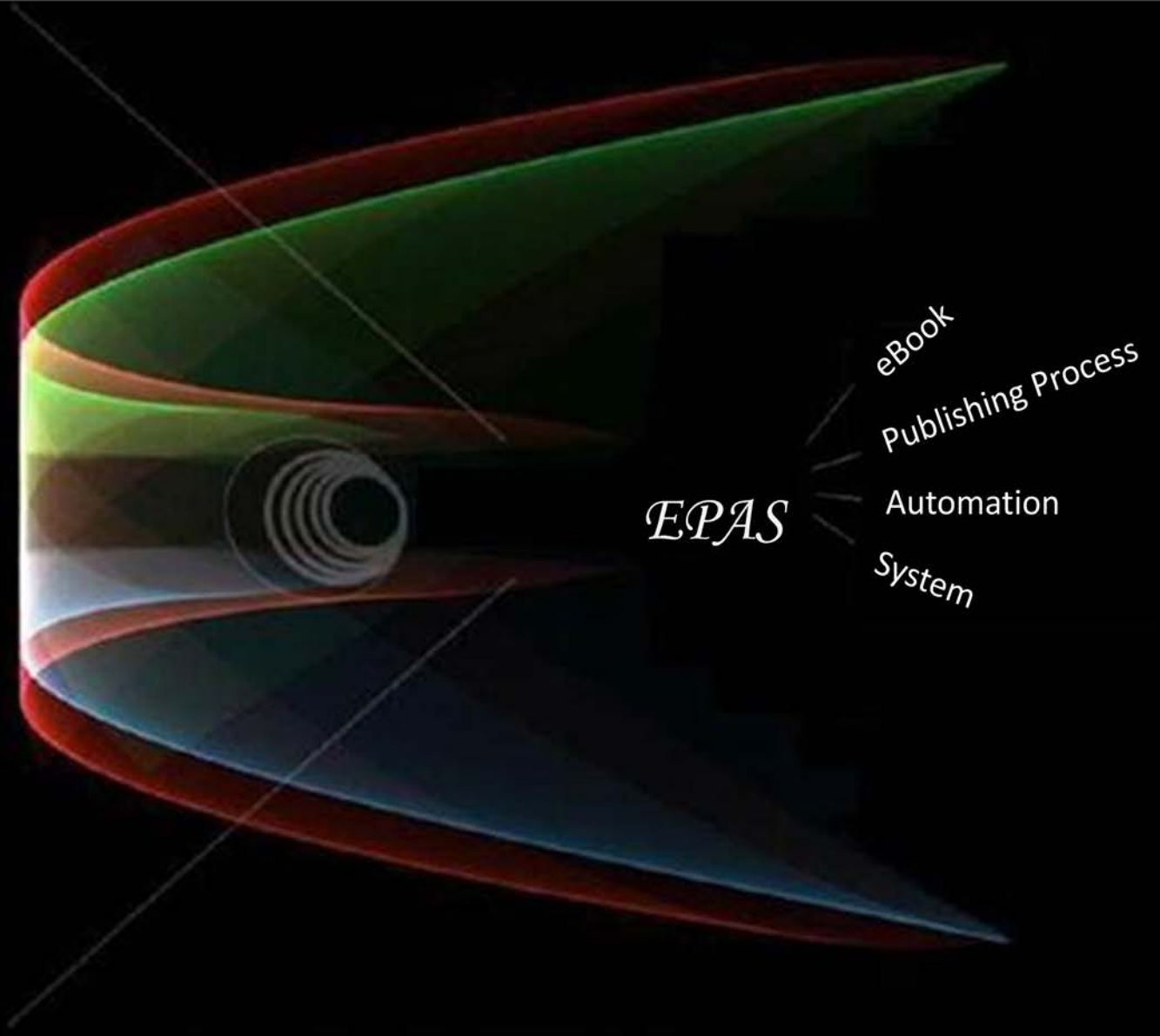


EPAS – eBook Publishing Automation Process System

The study proposes a new solution *EPAS* that meets the end to end requirement of eBook Publishing process and provides effective automation solution for processing huge data sets. To come up with this solution we analyzed existing eBook Publishing process and created few algorithms/scripts to process the data without any manual intervention.

Case Study Core Group –

Study Initiator – Mukesh Sharma
Analyzer and Implementation – Manik Mahajan, Hemmanshu Sethh, and Prashant Dhaka

Acknowledgement

Our CEO, Mukesh Sharma initiated a discussion with one of our prospective clients for a better solution for existing eBook Publishing process through automated tools and scripts. Our Quality group managers, Prashant Dhaka, Hemmanshu Sethh and Manik Mahajan have helped throughout this study sharing their valuable inputs at various stages for the study and helping resolve any queries.

What is this Case Study about?

This case study is about a new proposed eBook Publishing Automation System (EPAS). This solution would cater to end-to-end eBook processing before publishing on production environment for customer sales. EPAS provides capabilities of managing and validating huge data sets from different publishers that partner with our clients. This solution is extensible across clients with similar requirements.

Source of Case Study: EPAS tool integrates with eBook Publishing system to validate and process PDFs in specific formats to publish on eBook database and web production environment. This was necessitated due to:

- Existing PDF eBook delivery system which required manual efforts to maintain various publishers' datasets
- Heterogeneous processes to maintain publishers' datasets
- Varied client eBook processing models and requirements
- Ongoing discussions with QA and engineering teams

Why this Case Study?

Problem Statement: There was a need for an effective solution to help our clients process voluminous publisher datasets reducing manual efforts and bringing in an effective eBook delivery model for their customers. Clients were in need of an automation solution that offers end to end eBook processing capabilities such as PDF validation, PDF processing, metadata extraction, huge datasets maintenance, schema conversion, data auditing, web delivery of processed eBooks. Without such a solution the process required significant manual efforts and had dependencies on different vendors to deliver eBooks to their customers.

Some questions that formed the basis of this study include –

- Is there a market out of box or an in – house solution to provide:
  - E2E eBook Publishing process
  - Dataset management across publishers
  - PDF validation, data extraction and data auditing
  - A processing model for eBook delivery with minimal manual efforts

Rationale: The purpose of this study was to analyze one of our client's existing eBook processing/delivery model to enhance the process and make it more effective. This will help us determine whether few changes to the existing model will sufficiently fulfill the deltas that exist or if we would need a new eBook delivery solution to:

- Effectively manage volumes of different publishers' datasets (source PDFs)
- Build a solution to replace existing manual activities and increase the quality standards
- Reduce cost and increase ROI

Scope: Analyze the prospect's eBook processing/publishing model and identify the activities/tasks (PDF validation, PDF processing, metadata extraction, maintain huge datasets, schema conversion, data auditing, web delivery, etc.) to be automated to replace manual efforts. Additionally, perform a feasibility study to confirm that the proposed automation solution has the potential to maintain the client's quality standards, reduce cost and dependencies on vendor teams

Approach adopted: After preliminary analysis of existing eBook publishing model of prospective client and in-house CDL tools we conducted:

- An exhaustive study of three extensively used CDL tools (PDF validation, PDF metadata extraction and schema conversion tools) to understand the functionality offered by each of them
- An analysis of the benefits and drawbacks of available CDL automation tools
- Iterative brainstorming sessions within the CDL core team, our tool experts and R &D team to better understand the implementation of these tools
- Based on this study we were able to conclude that a new tool is not required and that customization on our existing tool (to be named EPAS) should adequately meet the prospect's requirements

EPAS: EPAS stands for "eBook Publishing Process Automation System" proposed with a vision to design an end to end solution to fulfill our prospect's eBook publishing requirements.. This system will have different modules (envisioned to have 8 modules) and will provide a platform to our prospects to trigger processing in an automated manner. List of modules that were developed include –

- PDF validation
- TOC Extraction
- References extraction and analysis
- Data auditing
- Inventory processing
- Schema Conversion processing
- Ingestion on QA or production environment
- Reporting module

PDF validation: This module will validate PDF eBook files received from various publishers to ensure delivered PDFs have no issues such as PDF corruption, password protection, encryption, compatibility, version, bookmarks, attachments, water marks, etc.

TOC Extraction: This module will extract articles and related meta data (author info, article type, article name, publishing year/date, etc.) from various PDF files and create TOC for individual articles.

Reference Extraction: This module will extract author references of different articles to analyze the royalty points of the individual author, based on article search/hits/sale.

Data auditing: This module will ensure all articles (along with meta data) exist in database and are searchable. This will provide flexibility to customers to filter articles based on various search terms.

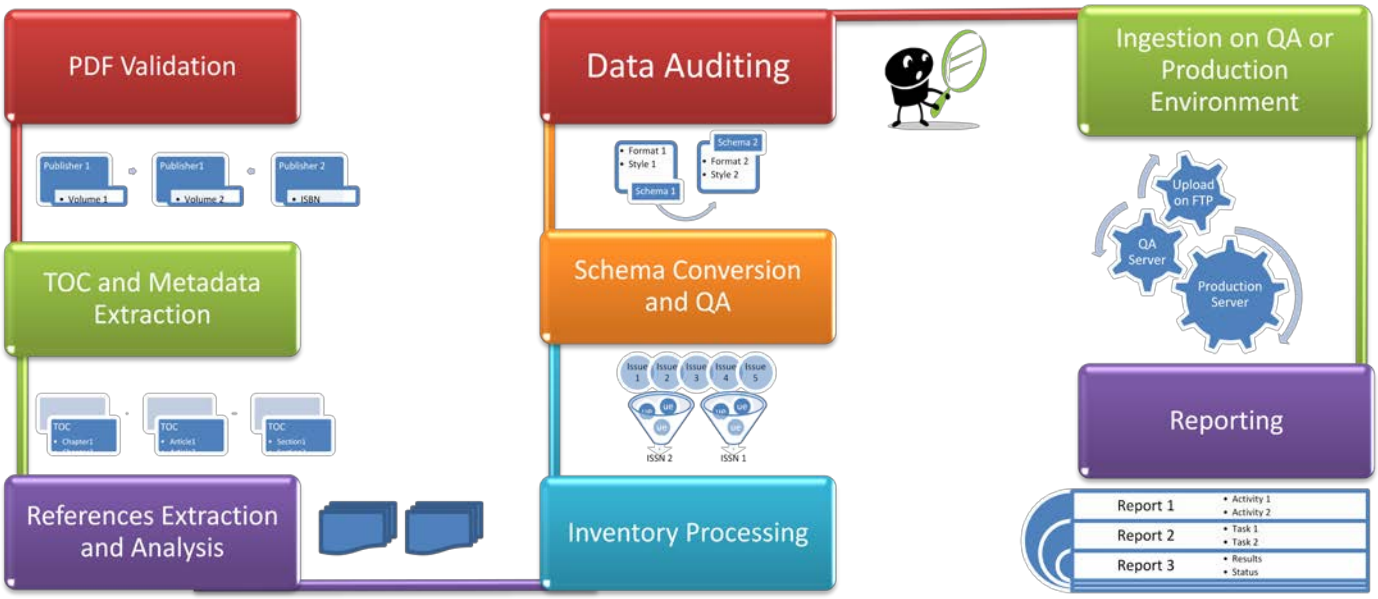
Inventory Processing: This module will manage articles on a shared location indexed by different publishers, article volumes, publishing years. This will also help to maintain data repository of various publishers' datasets.

Schema Conversion: This module will extract metadata from received PDF/xml files and create a schema xml for individual PDFs/xmls per client's system integration requirements. These files will be integrated in the existing system.

Ingestion: This module will help the team in server ingestion process for eBook web delivery on QA and production environment.

Reporting: This module will help the client team to extract a detailed report for the processing/tasks done by EPAS and also provide current status of activities under execution including a brief summary on completed tasks for tracking purposes.

eBook Publishing Workflow



Value Add from the EPAS Study:

- Evaluate existing in-house tool functionality for the prospect's eBook publishing requirements
- Identified key areas of improvements/gaps in existing CDL tools to be bridged to derive an optimal solution
- Designed different modules to meet the prospect's requirements to easily integrate into the existing model
- Identified missing features/functionality in the existing PDF validation tool for further enhancement and development.
- Upgrade CDL tools profile with newly developed tools and further adapted eBook industry wide processing to provide an optimal solution
- Provide an edge to our prospects by enhancing the publishing process's ROI and establishing them as market leaders in the eBook publishing industry

Services offered by QAInfoTech CDL division (Content Conversion and Testing):-

We provide Content Conversion Automation and manual engineering assistance (from our Content Digitization Lab division) which speeds up the existing project processes & ensures the timely delivery of quality content. Our services include:

- ✓ Data Conversion from PDF, DOC, and other soft copies to formats such as XML, ePub, OEB or other client requested formats etc.
- ✓ OCR extraction
- ✓ Proof-reading and editing
- ✓ Pre-application, pre-production, and post-production QA
- ✓ Processed file ingestion across various CMSs

Appendix

E-2-E – End to End solution

EPAS – eBook Publishing Automation Process System

CDL – Content Digitization Lab, a separate division of QAInfoTech Pvt. Ltd.