Errors and Omissions in Document Digitization Projects

By Manuel Bulwa

5/7/2020

DRAFT

I devoted over 50 years of my life to the field of information sciences, about half of that to digital document capture. Back then, the document management field was dominated by the micrographics industry, which was understandably reluctant to embrace the inescapable digital transformation. After a few years assisting many micrographic entrepreneurs to transition to digital, I started my own digital capture service bureau. 30 years later, over 2 billion images later, hundreds of projects later, hundreds of thousands of lines of code later, millions of Dollars later, one would think that new projects would be near error-free, but to my surprise and frustration it rarely happens. Errors and Omissions (EOs) tend to creep from the very early stages of a project all the way through the entire production workflow up to final client review and acceptance.

I looked up existing studies of QM/QA/QC/BP (Quality Management, Quality Assurance, Quality Control and Best Practices) standards, but what I found was either too narrow (for microfilmed, scanned and check image quality) or too broad (for generic Project and Risk Management articles). I could not find anything worth studying specific to major digitization projects for small to large format paper, microforms and born digital documents.

This paper is my attempt to lay a foundation for a model that could help prevent, detect, remediate and monitor EOs in a document digitization project. It succinctly mentions some methodologies with proven success record to illustrate how certain challenges were handled in projects where I participated.

The basic concept is to identify critical points where EOs may occur and establish preventive and corrective methodologies around certain metrics (Risk, Production and Integrity metrics) which play a role similar to the traditional Key Performance Indicators (KPI) used in Project Management. The metrics defined will propagate throughout the entire production workflow allowing checks and balances at strategic points. I use a “quadrangulation” concept that enables valuable checks and balances based on four workflow points in time:

1. **Origination** (Raw Capture Data): When the service provider accepts custody and creates manifest (inventory).
2. **Production** (Physical Context Data): At beginning or end of each task throughout the production workflow.
3. **Published** (Logical Context Data): When documents are classified and indexed.
4. **Submittal** (Final Data): When deliverables are subject to preliminary or final acceptance.

By cross-referencing, comparing and analyzing aggregate numbers based on data represented in these four coordinate systems, certain EOs become conspicuous, allowing for remediation processes to be effective.

A simplified production workflow includes tasks such as: custody logistics, manifest/inventory, prepping, batching, managing parallel sub-production lines, scanning, image processing, QC/repair, classification,

pagination, coding, indexing, segmentation, lookups, formatting, publishing, de-prepping, testing, deployment, reporting, submittal and final acceptance. Some of these tasks may be handled by one or more parallel and/or concurrent production lines and usually consist of several sub-tasks such as scanning on diverse scanners, barcode recognition, OCR, image enhancement, BPO subcontracts, table lookups, manual indexing, etc. Sub-tasks may be a combination of attended and unattended processes.

What could (and will!) possibly go wrong?

1. **At Origination:** Some documents, even entire batches, may not have made it entirely through the production workflow. They may be stuck in the middle of it, or mishandled, or believed present when they were not. A reliable manifest/inventory is the only way to initialize useful metrics. If the manifest was unsuitably created, these M.I.A. documents may linger in the dark for a long time or forever.
2. **During Production:**
    a. **Physical**: Pages can be missed, mutilated, obstructed, out of sequence, out of scale, illegible, wrongly split or merged, overlapped, files corrupted, files not compliant to SOW, defective blank page detection…
    b. **Logical**: Page groups can be wrongly classified/indexed, not indexed, duplicated, "buried"…
3. **Once Published**: Incorrect structures (missing or extraneous sections), truncated documents, missing documents, "made up" documents, …
4. **At Submittal**: Same as "published" above but affected by EOs introduced thereafter.

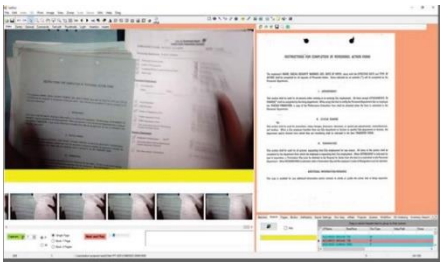
What are the causes of EOs?: Human errors, equipment malfunction, inadequate technology, software bugs/defects, poor project management, poor originals, inadequate reconciliation of parallel sub-production lines, etc. Media challenges such as the following also contribute to the proliferation of EOs:

- Paper: staples, clips, foldups, odd sizes, legibility, bleed-through, fragility, dust, bleed through, extreme thickness, post-it notes, taped pieces, …
- Microfiche: scratches, odd densities, separation between frames, mixed polarities, poor originals, inconsistent filming backgrounds…
- Roll microfilm: lead and trailer sizes, roll size, sliced (cut) sections, odd densities, separation between frames, blip reliability, inconsistent filming backgrounds …
- Aperture cards: water damage, presence of Hollerith perforations, reliability of Hollerith data, image quality and consistency.
- Large drawings: torn edges, extreme sizes, hazardous dust and droppings, extreme dust, creased foldings, fragility, poor originals, blueprints, sepia, vellum, dual side translucent drawings…
- Newspapers: Bleed through, brittleness, small fonts, fragmented text, …
- Bound books: Gutter issues, brittleness, foldups, taped pieces, …
- Ingested Digital Files: Password protection, integrity issues, format compliance issues, proprietary formats, …
- Diverse media: A mix of two or more of the above requiring the use of parallel production lines, each with a different type of scanning equipment (sheet feed, overhead, flat bed, large format, microform, etc.)

<u>Who is responsible for EOs</u>?: If you got this far reading this paper, I suspect that the answer to this question may be you. A thin line separates excuses from valid reasons when it comes to accountability and your Organization sometimes makes it even harder: The IT Department should not be responsible for help beyond installation and maintenance of equipment and systems software. The Records Management department seldom has enough C Management authority to secure adequate budgets and priorities, as it is frequently considered a cost center. This is then a job for a superhero like you, a project **champion**.

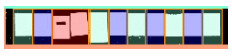
What can be done to reduce the proliferation of EOs?: Although each project has different idiosyncrasies demanding different approaches to each problem, there are common areas worth studying and considering, namely:

- **Planning, Testing and Training**: Document the SOW, define KPIs and profuse metrics, draft a Service Level Agreement (SLA), design a preliminary workflow, establish strategic reporting nodes, train, test, rehearse, strategically plant booby traps to stress test your defenses.
- **Timing**: It is of paramount importance to define early strategic QC tasks. Detecting an EO too late in the game can be disastrous and very costly (originals no longer readily available, defects found by client before you do, costly rework possibly needed, etc.).
- **Use Production Level Equipment**: Low cost equipment, including the so-called departmental scanners, are the main cause of mishandled double feeds and other nuisances related to paper thickness, size and quality. What you may have saved purchasing that equipment you are likely to spend in rescans and EO consequences.
- **Use Customizable Production Level Software**: I am yet to learn about the existence of a universal, "*one size fits all*" capture software solution. If you plan to use a third party capture solution, still allocate some budget to perform custom coding in virtually every major project.
- **Beware of the Human Factor**: A large project that includes intense human labor is susceptible to the negative effects (EOs) of human fatigue and monotonous, highly repetitive tasks. Despite dramatic recent advances in AI, RPA and software automation, human participation is still inescapable. A good compromise is to creatively combine both as in the case of the "Smart Inventory" where a person wearing smart glasses performs "hands free" video recording of folders in a box of documents. The goal here is to make very short pauses when the folder label is exposed to the video camera. The downloaded video will be later analyzed by software to produce still frames with the "best" images displaying the labels content. These still images are useful for inventory/manifest purposes and (surprisingly) much more. We inventoried boxes of medical records on an average of two to three minutes per box. A second example worth mentioning is when we video recorded every page of every folder for a Human Resources project. The goal here was to perform page by page visual QC against the scanned images. The system distributed video batches across multiple local and remote operators using an efficient and user-friendly interface where the side by side image comparison was semi-automated, and all exceptions flagged and audited.

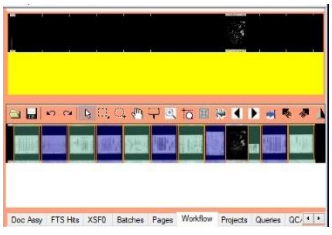


- **Avoid Transactional Capture**: By "transactional" I meant capture processes that perform cycles of prep, scan, index, QC and publish document by document, on the fly, by one operator. This obviously in sharp contrast against batch capture methodologies. I do not have enough fingers and toes to count the number of projects I declined participating in because of a mandatory transactional capture nature of these projects. In addition to the absence of multidimensional checks and balances I mentioned before, the fact that under transactional capture one person must undertake all tasks increases the risk of undetected EOs. For micrographic media, transactional ("on demand") desktop equipment is not recommended as part of a major project, or a project sensitive to consequences of EOs.
- **Adopt Batch Capture Methodologies**: Batch Capture and Processing is the best way I know to gain control over proliferation of EOs. As part of it, I use and abuse barcode based methodologies everywhere, all the time. There are three types of barcoded sheets we print prior to prepping:
    - **User Provided Data Barcode Sheets:** Whenever possible, we obtain a download of user provided files containing most (ideally all) of the document collection metadata. We drop these in the paper documents during prep before scanning. Document classes, categories or sections are also signaled by dropping class-specific barcoded sheets.
    - **Fixed Barcode Sheets:** Examples: Begin Batch, End Batch, Begin Document, Begin Staple, Begin Poor Quality, Begin loose sheets, Begin Multipage Doc, Begin Single Page Docs, Retake, Tab/Bookmark, Operator Alarm, …
    - **Identity Barcode Sheets**: Examples:
        - QC Rescan Barcode Sheets: Contain a unique Id of an image in our workflow database and exposes a miniature of the image to be repaired along with metadata useful to locate that page in a box of paper sheets or microforms. For microfiche, it displays a small image of the entire fiche, including title. Once ingested along with a rescanned image, the system makes sure it is inserted exactly where it belongs.

        - Twin Barcode Sheets: When documents need to split production across separate scanners (as when a box of paper includes folded large format drawings), one twin sheet travels with the drawings to a large format sub-batch, while the matching twin is placed at the point where the drawings were. The system perfectly reconciles all images regardless of where they were scanned, and the matching barcodes greatly facilitate physical return in place at de-prepping time.
        - UDOCS: A combination of the two types just described, but for cases where only one twin is needed, as when multiple pages need to be inserted.



- **Use Ribbon Methodologies for Microforms**: Under similar arguments I used above to prefer batch capture over transactional capture, I strongly advise to embrace ribbon capture and not the alternatives. By "alternatives" I meant capturing frame by frame as opposed to everything inside a ribbon, which will be subsequently segmented under strict control and auditing. Of course, if your goal is to digitize relatively small sections on demand, then these alternatives are perfectly adequate.



- **Define metrics, units of measure and tolerance ranges for each category of defects**: Examples: Under/Oversized images. Folded or torn document corners. Folded or torn edges. Image density, brightness and contrast. Noise. Front-Back Image density ratios. Focus/blur. Counts for: images per batch, indexed records per batch, blank pages detected, pages flagged for inspection, zero-page documents, microform frames, microfiche titles, images per ISO size, ...
- **Apply metrics**: Crosscheck all figures whenever possible and explain potential discrepancies. Involve aggregates such as averages, standard deviations, correlations, regressions, etc. Look for patterns and extremes. Sort each column of figures in both ascending and descending order looking for odd values. Create SQL queries to supplement static figures. Produce Bill of Health (BOH) reports.

Notes: This paper has been heavily edited to comply with size restrictions. Links were used to expand on

certain concepts if additional details are desired. Deeper levels of detail are available by request emailing the author (mbulwa@isausa.com). Some of the concepts explained here can be illustrated by watching this 5 minute video: