ISAUSA, INC.

Integrated Scanning of America

QC-ing the QC

By Manuel Bulwa May 2022

This article is my proposition for a holistic approach to QA/QC for document capture projects. It is also a recommendation to audit the effectiveness of QC methodologies used (or to be used) by a service provider. I tried my best to harness lessons learned and expertise gained during over half a century in the computer industry, half of which into digital document capture. I always tried to defy conventional wisdom using creative new methodologies against a very conservative industry segment once dominated by micrographic entrepreneurs. After a few years assisting several micrographic companies in their transition to digital, I started my own digitization service bureau. 30 years later, 2 billion images later, hundreds of projects later, hundreds of thousands of lines of programming code later, millions of Dollars later, one may think that current projects would be error-free, but no such luck. As in the story of the scorpion and the frog, it is in its nature. Perfection is an asymptotic curve.

Under the premise that profits are proportional to productivity and client satisfaction, I stubbornly continued to innovate and experiment with creative approaches to problems commonly perceived as simple, reinventing QA/QC for each project while being wary of substandard QC procedures represented by other vendors. In hindsight, although I failed more often than I succeeded, the aggregate benefits of successes greatly offset the aggregate cost of failures, which by the way is a price worth paying for valuable lessons learned.

Errors and Omissions (EOs): Out of all challenges involved in production, QC is the most underappreciated and abused task, despite being the most important protection against Errors and Omissions (EOs), the digital counterpart of misfiled, debased or lost documents. Conventional methodologies (some inherited from the micrographic era) offer substandard levels of success in **preventing, monitoring, detecting and remediating** EOs in a document digitization project. To prevent EOs from occurring (or minimize their occurrence) we need a robust production workflow running on production-level hardware and software tools. **Monitoring** EOs require the planting of tracking seeds and gathering profuse, multidimensional(*) production metrics all along the entire production workflow. **Detection** success depends on the results of applying timely analytics on the gathered metrics. As EOs tend to creep up from the very early stages of a project, a late discovery may have devastating, often irreparable consequences. **Remediation** requires methodologies to bridge the gap between the habitats of original and digital documents. There is enough there to write a whole book, but for this article I will try to condense it in a few thousand words.

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

• Paper: staples, clips, bindings, foldups, odd sizes and thicknesses, legibility, bleed-through, fragility, dust, extreme thickness, post-it notes, taped pieces, …
• Microfiche: scratches, odd densities, separation between frames, mixed polarities, poor originals, inconsistent filming backgrounds, overlapping jacket inserts, stapled sheets, mirrored frames, …
• Roll microfilm: lead and trailer sizes, roll size, sliced (cut) sections, odd densities, separation between frames, blip reliability, inconsistent filming backgrounds, splicing defects, …
• Aperture cards: water damage, presence of Hollerith perforations, reliability of Hollerith data, image quality and consistency.
• Large drawings: torn edges, extreme sizes, dust and residuals, abused foldings, fragility, poor originals, blueprints, sepia, Mylar, 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, encryption, integrity issues, format compliance issues, proprietary formats, security restrictions, …
• Diverse media: A mix of two or more of the above requiring the coordinated use of parallel production lines, each with a different type of scanning equipment (sheet feed, overhead, flat bed, large format, microform, etc.)

Some industries tolerate Errors and Omissions (EO's) in their document digitization projects more than others. In fact, some clients irresponsibly assume that their Digitization Service Provider (DSP) will reasonably comply with an explicit or implicit error tolerance declared in a Service Level Agreement (SLA). Clients should always ask DSPs to explain their QC methodologies and demand solid answers, then mistrust and verify.If clients tolerate weak answers, egg is on their face. Answers by DSPs may include a combination of:
• **Random sampling**: Although useful in most projects, random sampling overpromises and underdelivers. It helps, but it is often not enough.
• **100% QC:** an abstract notion sometimes attempted at a very high cost. However, with ingenuity we may get close enough.
• **Counting pages:** Error prone, unreliable. Matching counts is frequently misleading due to undercounts offsetting overcounts. Reconciling human counts against computer counts often yields false positives and false negatives.
• **Dual blind verification:** Only useful in indexing, often imperfect.
• **Control totals:** Crosscheck of input and output metrics at every workflow step. Crucial and effective, but usually one-dimensional (*).
• **Lookup tables:** Crosscheck against user supplied data. Although mostly useful in indexing, it could extend way further.
• **Page By Page Video QC:** A video is taken showing each and every page before scanning. The system then presents a graphical interface that allows an operator to compare still video frames against scanned images, looking for errors and omissions. This is an effective but expensive solution to a select set of extreme QC circumstances,
• **Good documentation:** Statement of Work (SOW), Key Performance Indicators (KPI), Service Level Agreement (SLA), Acceptance Test Criteria and Production Instructions are crucial in containing EOs, if properly adhered to.
• **Capture methodology:** Batch oriented methodologies (the entire workflow is applied against a batch of documents) is used in large volume backfile conversion projects, while transactional methodologies (the entire workflow is applied against each document) is used in low volume day-forward capture. Used otherwise signals trouble.
• **Bill of Health:** A comprehensive multidimensional (*) post-mortem report of all raw metrics and analytical results. My favorite and highly recommended. On a side note, I am currently experimenting with Machine Learning (ML) to enhance the analytic component.
• **Other methodologies.**

(*) Multi-dimensional analysis: A concept that enables valuable checks and balances based on four distinct workflow perspectives (dimensions):

• **At Origination:** Raw data captured when the service provider accepts custody of a collection, and a manifest/inventory is produced.
• **During Production:** Contextual data captured at the beginning or the end of each task throughout the production workflow.
• **At Publishing:** Contextual data captured after documents are classified and indexed.
• **At Submittal:** Final data captured when deliverables are subject to preliminary or final acceptance.

A simplified production workflow includes tasks such as: boxing and inventory, chain of custody, logistics and transportation, manifest, prepping/de-prepping, batching, coordinating parallel production lines, scanning, image processing, QC/repair, classification, coding, indexing, lookups, formatting, publishing, de-prepping, testing, on-demand work in progress (WIP) requests handling, deployment, reporting, submittal and final acceptance.

The following is a partial list of things to watch when assessing answers from a DSP:

Planning, Testing and Training: Well documented SOW, defined KPIs, Service Level Agreement (SLA), Acceptance Criteria and strategic reporting milestones.
Chain of custody and check-in/check-out forms: Should be signed and dated each time a DSP picks up and/or returns documents, boxes, media…must show identification of box/container labels
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.).

Boxing: If drawers, cabinets and boxes are carefully labeled (barcoded stickers do help) then boxing should be mostly uneventful. The boxing report must identify the office, cabinet, drawer and box as well as what has been placed on each box (by from-to range or enumerating every folder tab or binder title). A creative inventory method we use consists of a person wearing smart glasses exposing each tab folder or binder title in a box. This allows a video capture of each and every (identified) document on each box by just one person in a couple of minutes. The video so captured is later turned into still frames that a special software program uses to search and find any document label in a handful of clicks…without the cost of data entry. If you wonder why smart glasses, the answer is that it allowed us to cut labor cost in half when we no longer needed one person handling a camera and a second person fingering the folder tabs. A more traditional inventory process will simply enter the pertinent label data on computer or paper forms.

Manifests and Inventory Catalogs: An inventory catalog is a document level list of document labels (metadata), while a manifest is a box level list of box labels. A manifest may also include observations on each box, such as "damaged", or "half full", or "empty", etc. These observations are used to reduce the number of false positives and false negatives during QC defect flagging. It also offers many other protections, such as making sure each and every box safely travelled through the entire production workflow. An Inventory Catalog does not need to be manually created if a digital record exists of most or all of the documents to be scanned. Although users are prone to say that they do not have a reliable digital record of what is in their filing cabinets, chances are that they do exist somewhere in the organization. This data can be used to print unique barcoded sheets that show the document label on each, so they can be dropped in each folder during prepping. This catalog methodology has multiple benefits, including:

- Widows and Orphans QA/QC: At the end of the capture project, any unused barcoded sheet is an indication that a physical record was either mishandled during the project, or the digital file is pointing to a record that may be in someone's desk, or briefcase, or elsewhere unaccounted for or lost, or not part of this particular capture job. A physical record that did not find a matching barcode indicates that the record is not being managed by any computer system for any purpose, or that its matching barcode was mistakenly placed in the wrong folder.

- Inexpensive, accurate indexing: All documents scanned and recognized will be linked with the metadata contained in the digital files used to print barcode sheets. This makes indexing more accurate and more reliable than manually creating the inventory catalog.

Batching: Every batch should be preceded by an Begin Batch barcoded sheet and ended with an End Batch barcoded sheet. Every box must start with a Begin Box barcoded sheet and ended with an End Box barcoded sheet. These sheets will prove invaluable when assessing the integrity and completeness of batches.

Prepping: Paper documents kept in traditional filing drawers and cabinets may be stapled, fastened, bound, folded, kept inside envelopes, taped, glued, clipped, marked with sticky notes, etc. All these conditions create capture challenges that can only be handled through tedious, monotonous but necessary human labor. Prepping is a production activity that facilitates scanning by removing all these obstacles as much as possible or practical. However, they were placed there for good reasons such as adding contextual boundaries and meaning (for example: a document may be defined by stapled sheets). Removing them facilitates scanning, but we need to capture the fact that they exist if we do not want to lose the value of such contextual boundaries. This may require dropping barcoded sheets for conditions such as the ones just mentioned above under "Batching" as well as others such as: "Begin Staple", "End Staple", "Begin loose sheets", "Operator Alarm", "Twin", "Tab", "Begin Doc", data driven barcode, etc.

Deprepping: If the DPS is asked to return originals, they should be neatly packed and possibly restored back to their exact original status (i.e. restapled, reclipped, reinserted in envelopes, etc.) if so indicated in the SOW. During de-prep it is possible to "catch" EOs, as the de-prep operator may have a second chance to pair data driven barcodes to matching folders and catch prepping errors. This somehow acts like a "dual blind" control.

Scanning: Production lines may include sheet-feed, flat-bed, conveyor transport or overhead scanners. Sheet feed scanners are prone to double feeds, which carries the risk of missing pages through double feeds. Production level scanners use multiple dual feed sensors. Flat-bed scanners will not double feed unless the scanning operator is negligent or careless, but flat-bed scanning is highly inefficient, and its labor is cost-prohibitive. Conveyor transport scanners allow very fast manual feed with a minimum chance of double feed as the operator drops each page on a conveyor belt, which moves each page to the cameras while still detecting potential double feeds. Overhead scanners offer the convenience of capturing stapled, clipped and bound documents at the expense of some cosmetic sacrifices and other minor imperfections. If the source documents include large format drawings folded along with business size sheets, solid objects as in product samples, CDs, bound material that cannot be cut apart, etc. then the solution is to use and coordinate multiple production lines, each with the appropriate scanning equipment. In this case the best option is to use "Twin Barcodes". These are pairs of identical unique barcode numbers lying on a stack by the prep table. The unscannable segment is removed from the set preceded by one of the twins, while the identical twin remains in place as a placeholder. This methodology allows portions to be scanned on separate production lines, but the logical and physical reconciliations are perfectly handled. Finally, do not ask your DSP to skip certain documents (selective scanning). The perceived benefits frequently die away with high risks and costs.

Avoid Transactional Capture: 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 for large volume projects.
Adopt Batch Capture Methodologies: Batch Capture and Processing is the best way I know to gain control over proliferation of EOs.
Avoid Selective Scanning: When a person that is not a subject matter expert (SME) is asked to decide what should skip capture, things can go south real fast.More often than not it is better to scan it all and make "go/no-go" decisions digitally. This allows remote working by multiple staff including SMEs, provides a recourse to mistakes, and more.
Use Ribbon Methodologies for Microforms: 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.
Classification/Indexing: Classes of documents (document types) should be captured by automated, manual or semi-automated means involving barcode/patchcode recognition, OCR/ICR, table lookup, text analytics, manual key from image, forms id, etc. If indexing is performed manually, accuracy targets may be attained using "dual blind" key from image, where one operator enters data as visualized on the images, while a second operator does the same without knowing what the first one did. A referee (human or otherwise) determines the course of action when the two differ.
Gather and analyze profuse 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.
Publishing and Deployment: Once classified and indexed, records may be published into a temporary or permanent repository or content management system. While data is published as database structures, images may require formatting as dictated by the ECM of choice. Formatting should include validation of PDF/A compliance and file integrity checks to flag corrupted and suspicious images. Although extremely popular, publishing digital documents as named folders and files under Microsoft Windows, is contrary to best practices and is plagued with pernicious side effects. If forced to keep it simple and low budget, a better alternative is to publish them using hyperlinked tables of contents and refrain from naming files and folders after metadata.
QC/Repair: Once a defect is detected, the remediation may or may not require physical rescans. If it does, we need to bridge the gap between the habitats of original and digital documents, i.e. we need to physically locate the container where the originals are, then locate the targets to re-digitize and adjust all digital representations of it after a rescan. This could be aptly handled by "Smart QC barcodes", printed out by the production software showing metadata useful in locating the container and document, as well as a thumbnail of the target itself. By preceding the rescan with this barcoded sheet, the system automatically posts the rescan where it belongs and updates all related databases and logs.
Reporting: Reports are generated at various stages, including: production control, billing, QC, change of custody, manifest, inventory, CAPA (Corrective Action Preventive Action) and RCA (Root Cause Analysis), progress (burn up, burn down charts), KPI/SLA (Service Level Agreement) compliance, OCR accuracy reports, and more.
Bill Of health (BOH): Must show comprehensive, thorough metrics of all image and data attributes. A couple of examples of how to use BOH metrics follow:
Averages and StdDevs per document and per box of page sizes, page counts, index counts, blank pages, barcode types, image densities, words OCRed, etc. This can help flag potential defects such as images too dark/too light, boxes too packed/not packed enough (check against manifest notes), odd presence of too many/too few small/large scanned pieces, unindexed documents, under/overpopulated with barcode sheets, missed back blank page removals, odd number of not right side up/graphic pages.
Bad correlation between twins at the box level against twin batches: This is used to flag missed or mishandled twins (pieces that could not be scanned using the main scanner and needed to be separately scanned elsewhere0.

In conclusion: The quality of results from a large document digitization project depends on the robustness of the production infrastructure and on the effectiveness of the QC methodologies used. It is the end user's responsibility to timely and thoroughly assess the former before the project starts and the latter (QC the QC) before, during and after the project concludes.

Questions and additional levels of detail are available by request emailing the author mailto: mbulwa@isausa.com Some of the concepts explained here are illustrated in this 5 minute video: https://youtu.be/Wd3XM7Nxs9g

ISAUSA, INC.
2363 Teller Road Unit 102 Newbury Park, CA 91320
Phone: (805) 375-0422