

**P | T | F | S |**  
CONTENT MANAGEMENT AND LIBRARY SOLUTIONS

3

5

6

C-H-N-S-V

2

7

U

**TOP SECRET**



**ArchivalWare** **DX**

## ENTERPRISE DECLASSIFICATION

*Technology Assisted Declassification Processing for inter/ intra Government Agency Reviews*

## EXECUTIVE SUMMARY

President Clinton ushered in a period of more open government by signing Executive Order (EO) 12958 in 1995 requiring more active declassification programs by all agencies with classified repositories. Subsequent EO's executed by Presidents Bush and Obama have clarified, reinforced, and expanded many of the original tenants in Clinton's EO. Today, the current EO 13526 still requires agencies to review all classified documents 25 years old or older to determine whether continued classification is warranted.

The methods, processes, and tools deployed by federal agencies today to determine whether all or part of a given document can be released to the public are archaic, manpower intensive, expensive, inefficient, and prone to error. In many cases the documents are processed through all stages as hardcopy media. This processing is particularly challenging since there are varying media formats such as paper, microfilm, and audio. In most cases the review is completed manually by scarce and expensive cleared human reviewers. Tools in place today utilize hardcoded workflows which are frequently changed by an army of programmers trying to keep up with structured and unstructured process changes.

PTFS has built new functionality into their flagship product ArchivalWare DX which enables agencies to improve declassification workflow, increase productivity, reduce costs, and increase accuracy. The core elements include the ability to electronically search across documents for dirty words and concepts using a specially calibrated Adaptive Pattern Recognition (fuzzy) search providing for a very low tolerance to any false negatives. The application also has flexible workflow based on the jBPM, a Java based workflow tool, which enables system administrators to make workflow changes quickly on the fly rather than wait for programmers to make hardcode changes. The application is 100% web based and requires no client side software; a feature that will facilitate more efficient processing for both inter and intra agency declassification processing.

## A BRIEF HISTORY

### Executive Order 12958

In 1995, U.S. President Clinton signed EO 12958 which created new impetus for the process of declassifying documents and led to an unprecedented effort to declassify millions of pages from the U.S. diplomatic and national security history. EO 12958 and subsequent amendments require that all classified documents 25 years old or older be reviewed to determine whether continued classification is warranted.

### Executive Order 13291

In 2003 President George W. Bush signed EO 13291 replacing the soon-to-expire Clinton-era EO relating to the automatic declassification of federal government documents after 25 years. The new EO retained the essential provision of the Clinton order—automatic declassification of federal agency records after 25 years—but with some notable caveats. In general, it gave the government more discretion to keep information classified indefinitely, especially if it falls within a broad new definition of "national security." The EO made it easier for government agencies to reclassify documents that have already been declassified, and it made it easier for agencies to classify what is characterized "sensitive" material. The EO also

expanded the list of exemptions of information from future automatic declassification: information that would "assist in development or use of weapons of mass destruction," reports such as "national security emergency preparedness plans," and information relating to "weapons systems." Finally, the order created a 3-year delay for requiring that all agencies comply with the Clinton EO 25-year targeted declassification date.

### Executive Order 13526

On December 29, 2009, President Obama issued EO 13526 that would dramatically change the way the executive branch handles classified material, reduce over-classification and expedite the release of formerly classified materials to the public. Federal agencies would be required to eliminate a 400 million page backlog of materials awaiting declassification by December 31, 2013.

In addition, the President issued a memorandum to heads of departments and agencies that directs additional steps agencies should take to implement the order. The White House also released a Presidential order entitled "Original Classification Authority. This order designates those agency heads and officials as having the authority to classify information as "Top Secret" or "Secret" under the executive order.

Among other major changes, EO 13526 established a National Declassification Center at the National Archives to centralize and streamline agency reviews of classified materials. The Archivist of the United States is now charged with developing declassification priorities with input from the general public and after taking into account researcher interest and the likelihood of declassification. On December 30, Archivist of the United States David Ferriero announced the immediate establishment of the NDC within NARA.

Under the direction of the NDC, agencies will be required to take steps to eliminate the backlog of more than 400 million pages of accessioned Federal records. These Federal records were previously subject to automatic declassification, and now the NDC must take steps to permit public access no later than December 31, 2013. These include archival records related to military operations during World War II, Korea, and Vietnam.

For the first time, the EO establishes the principle that no records may remain classified indefinitely and provides enforceable deadlines for declassifying information exempted from automatic declassification at 25 years. The new EO strengthens the standards agencies must meet to exempt any record from automatic declassification at 25 years; nine well defined categories are specified. EO 13526 also establishes a second automatic declassification period at 50 years with even tighter standards for exemption; only documents revealing human sources or key design concepts may be exempted past 50 years. This 50 year provision takes effect on December 31, 2012. Finally, the EO prohibits classification beyond 75 years except in extraordinary cases.

For the first time, it requires agencies to conduct fundamental classification guidance reviews to ensure that classification guides are up-to-date and that they do not require unnecessary classification. It also directs that information not be classified (or be classified at a lower level) when "significant doubt" exists about the need to classify it. Finally, the EO significantly tightens restrictions on reclassification of information after its declassification.

The cumulative impact of this legislation presents a dilemma for government agencies: on one hand, protecting sensitive information is critical for national security but on the other hand they must maintain openness, serve U.S. citizens and comply with the EO.

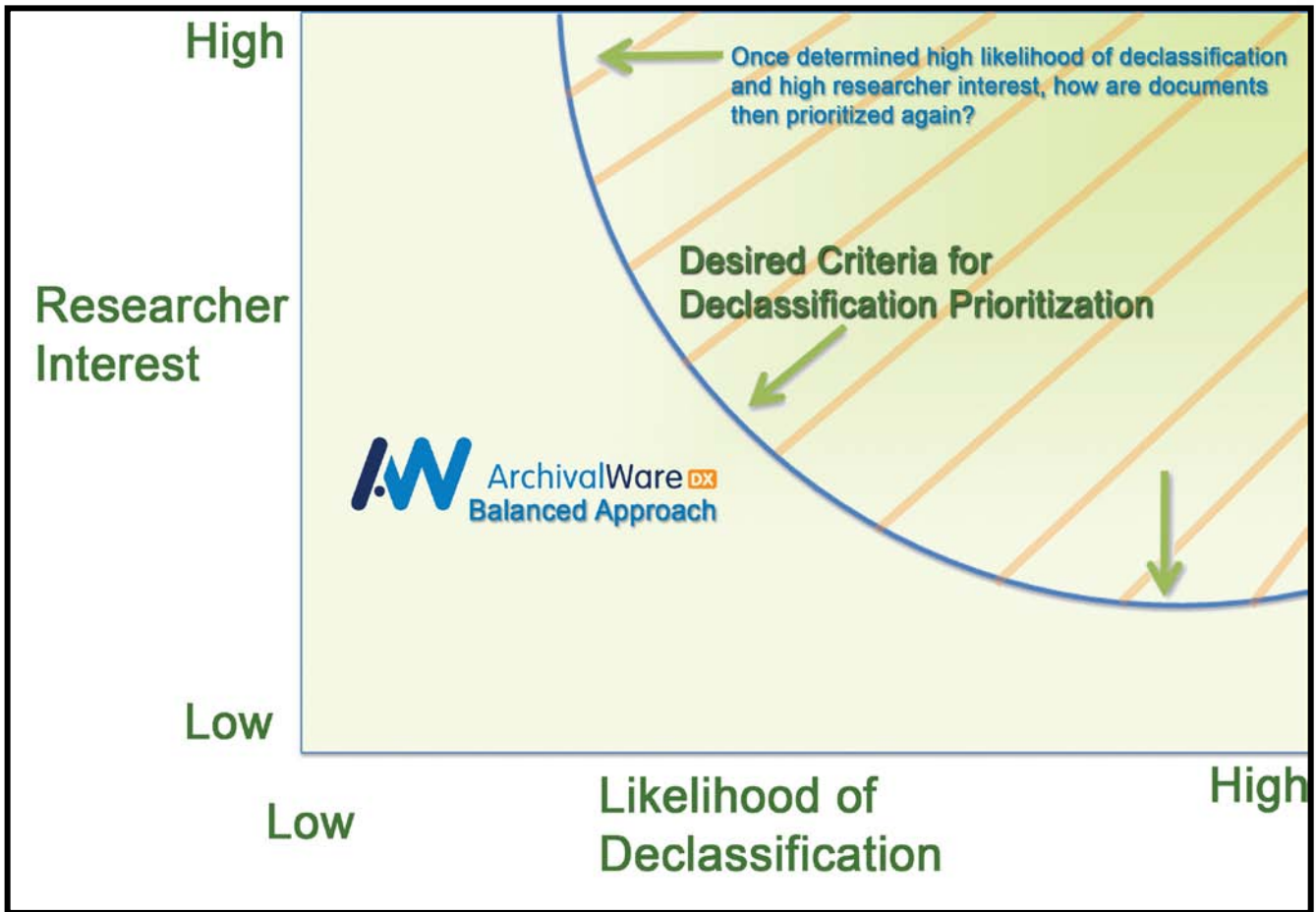
# CURRENT PROCESS ISSUES

The methods, processes and tools employed by Federal agencies today to prioritize documents for review, and then determine whether all or part of a given document can be released to the public are inefficient, outdated, resource intensive, costly, and prone to error. The result is a huge backlog of documents that is growing, and a growing staff of expensive analysts to complete the reviews. The complete dependence on human analysts to subjectively apply classification guides to declassify or exempt documents is naturally prone to errors that, at best, are embarrassing and, at worst, create security risks.

## Prioritization Process and Challenges

Given the enormous backlog of documents, limited resources to review documents and inefficient processes for reviewing documents, prudent administrators first prioritize collections of documents subject to declassification.

Figure 1 - Prioritization



Generally they prioritize documents by evaluating two variables. The first consideration is the relative importance of a particular document to researchers. The document subject, author, time period and document relevance in history all contribute to researcher interest.

Next, leaders of declassification efforts consider the likelihood of declassification. Documents tend to fall along a continuum based on the sensitivity of information in documents; documents that reveal sources and

weapons of mass destruction design details are likely to always remain classified. Ideally, the optimal prioritization is a balanced approach that effectively considers researcher interest and the likelihood of declassification. This approach allows agencies to use their limited resources to declassify the maximum number of documents that also rank high to researchers. The dilemma around prioritization is portrayed in Figure 1.

The challenge today is that all prioritization is made manually without electronic assistance. The process is slow, inefficient, and prone to subjective errors that result in inefficient use of resources focused on documents that are not the highest priority or documents that are highly unlikely to be declassified.

### Declassification Process and Challenges – Intra/Inter-agency

There are two basic processes employed to declassify documents, and both are slow, inefficient, labor intensive and subject to error. Many agencies still review hardcopy documents and make a pass/fail decision at a document level. Currently, the state of the art redaction and declassification processing is performed by scanning hardcopy source materials to a TIFF image, manually reviewing (reading) every word of the document on a computer screen, and redacting words/sentences/paragraphs manually using the electronic equivalent of a black magic marker.

While this electronically assisted method allows manual electronic redaction, it still requires a word-by-word review of each document and manual creation of redaction zones. Since the current declassification process is conducted almost exclusively manually, this process has obvious shortcomings:

- This process requires highly skilled analysts with domain subject matter expertise and the highest level clearances. These resources are expensive and in short supply.
- Analysts must interpret lengthy and complex declassification/classification guides which provide conceptual and nebulous guidance. The analysts must then subjectively apply their interpretation to documents as they perform document level and/or word-by-word reviews. Naturally this process is prone to errors by even the most competent and dedicated analysts.
- The process is naturally monotonous and tiring which is conducive to human fatigue and significantly increases the likelihood of human error. A very low tolerance for errors is the only acceptable level of performance.
- It is challenging to catch duplicate documents, and two different reviewers are likely to redact a document differently. Since the process is not repeatable, releasing duplicate documents is embarrassing and potentially dangerous as it may allow information to be constructed by comparison of two versions of redacted declassified documents.
- The combination of expensive people and a time consuming review process results in low production rates (documents processed per unit of time) and high costs (cost per document processed).
- When a classified document contains multiple agency equities, the document must be referred to that agency for review. Currently the method used to route classified documents to external agencies for review is slow and inefficient. Additionally, the originating agency loses visibility into the referral agency process since there is no electronic tracking.

## Media Diversity, Condition and Evolution Challenges

The content/media subject to review for declassification is diverse and evolving. Currently most of the material for the review varies across a wide set of hardcopy media types. While the focus has been on the backlog of paper documents, it is well recognized that other media types including microfilm, microfiche, video, audio, and other media types are in the queue for review. Additionally, much of the hardcopy media was not recorded well on the original source material (onion skin) or transferred using optimal techniques (microfilm scanning). Agencies need better techniques to process diverse materials or the backlog will grow exponentially. The challenges created by this cumulative change include:

- Each different media type requires different hardware, software, and processes for conversion to digital content for electronic processing. There are different scanners for high volume paper, fragile paper, microfilm, microfiche, video, audio, etc.
- A lot of the material was poorly printed or is in poor condition. The poor text in many documents will require highly tuned software to make electronically searchable documents. Both material challenges require optimized hardware, software, and processes for conversion to usable digital content.

We have also now reached the point in time where the first mass of born digital material is 25 years old. Naturally as time goes on, the stream of born digital files will increase and there will be a corresponding decline in the material only available as hardcopy. With this transition we will see an explosion in the volume of material requiring declassification. The large wave of born digital content designated for declassification requires implementation of new declassification processes and systems.

## Software Challenges

Pioneering efforts have been made to use electronic tools to assist analysts; however, these tools have serious shortcomings and discourage improvements on current productivity and cost metrics. Shortcomings include:

- The tools were primarily developed to support only bi-tonal formats; grayscale and color formats have not been supported. Older material with dirty backgrounds, photos, and poorly formed text requires grayscale processing for satisfactory results. As time goes on, there will be more and more color material which also cannot be processed in a simple bi-tonal mode.
- Tools currently employed do not work well for OCR'd materials. Even with the best OCR engine that has been properly tuned, there will be OCR errors due to dirty backgrounds and poorly formed original text. A technology solution called Adaptive Pattern Recognition Processing (APRP) can be utilized to overcome OCR inaccuracies. This technology also compensates for periodic typos in the document as well as inaccurate and misspelled queries. This capability is critical when searching for entities such as Arabic names with many spelling variations.
- The tools employed today are for digitized formats and not born digital content. As time goes on, there will be an increasing requirement for born digital processing.
- Electronic workflow is not easily configured to meet rapidly changing workflow requirements. The workflow utilized today is hardcoded in a programming language. The processes and procedures reflected in the workflow change frequently. There are both structured and unstructured processes that are part of the declassification process. Changing the hardcoded workflow is expensive, has long lead times and is subject to breaking other code.

- The current tools do not provide the capability to perform semantic relationship concept searching. Many dirty words can be readily expanded into like words: Missile into Exocet for example. The current tools will miss these expanded terms unless each word is specifically identified.

## NEXT GENERATION DECLASSIFICATION

PTFS provides enterprise declassification solutions to protect sensitive information and release content that is properly sanitized for mass consumption. Whether it's hardcopy or digitized, born digital or other forms of electronic content, PTFS can provide full life cycle support from secure digitization to declassification processing and redaction. Benefits from the PTFS Declassification solutions can be found in Figure 2 below.

**Figure 2 – PTFS Declassification Features and Benefits**

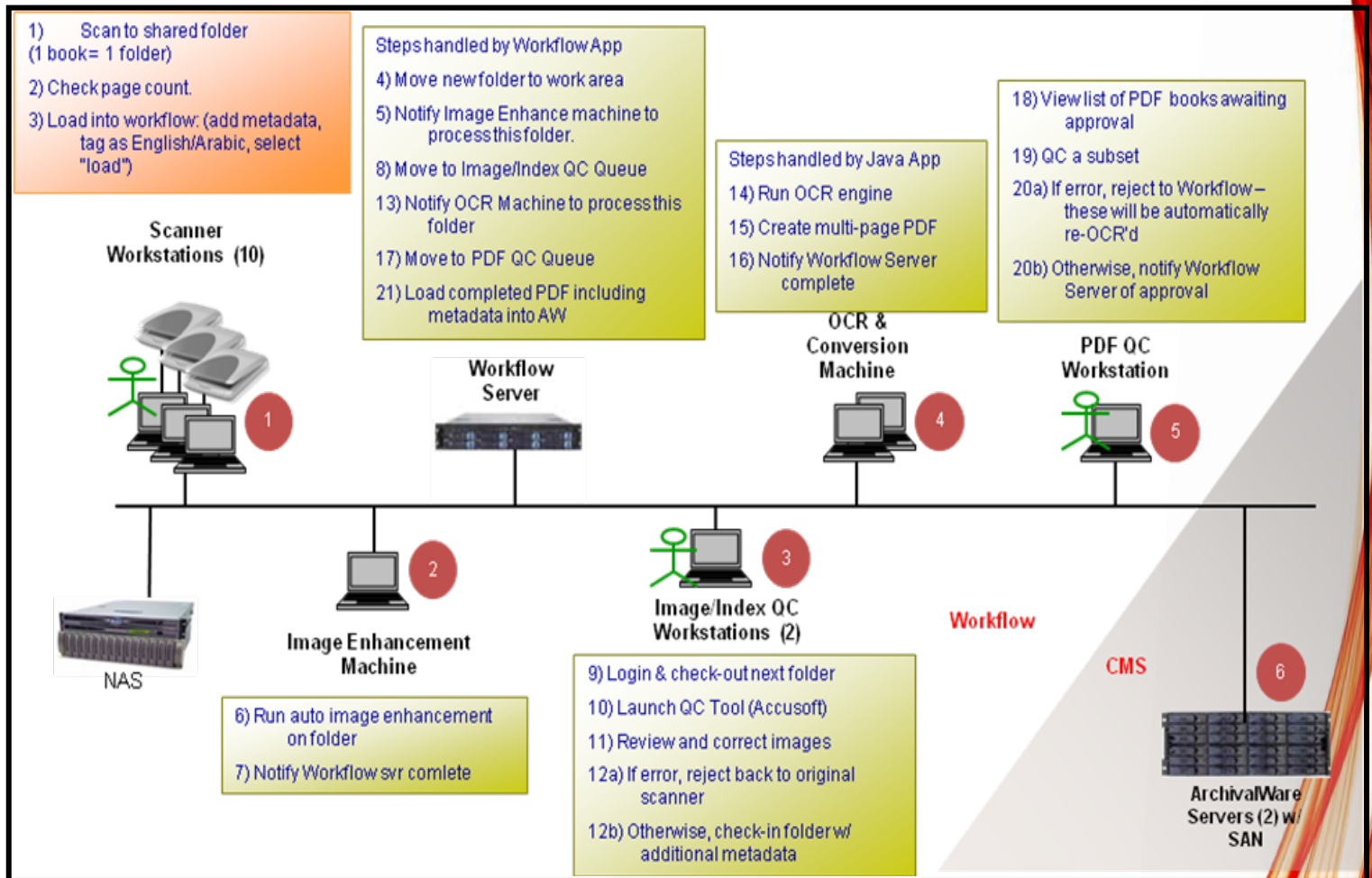
Features	Diverse Content/Media Processing	Increased Process Efficiency	Minimize Errors	Cost Reduction
Automated Search: Dirty Word Glossary is Created and ArchivalWare DX Locates and Highlights Hits	X	X	X	X
Entity Search (proper names and places): Allows for HUMINT Searching	X	X	X	X
Electronic Referral Process	X	X	X	X
Adaptive Pattern Recognition Processing (APRP) Search Technology to Overcome OCR Inaccuracies and "Dirty Word" Misspellings	X	X	X	X
Utilization of Custom Semantic Dictionaries, Replicating SMEs and Classification Guides	X	X	X	X
Best Practices for Creating Hi-Res PDF/A's and Accurate OCR, Out of Poor Quality Source Material	X	X	X	X
Built-in Flexible/Customizable Workflow Management	X			X
Concept Search Capabilities to Locate Synonyms That Could Possibly Be Overlooked	X	X	X	X
Electronically Categorize Documents for Prioritization	X	X	X	X
Eliminate Duplicate Document Processing	X	X	X	X
Effectively Manage All Digitized (Color and Grayscale) and Born Digital Content	X	X		X
File Type Agnostic Platform; 225 Supported File Types	X		X	
Numeric and Wildcard Searching: Allows for Recognition of Standard Identification Formats (Social Security Numbers, Phone Numbers)		X	X	X
Speech to Text Functionality: Enables Declassification of Audio Files	X			
PDF/A and XMP Metadata		X		X

### Digitization

PTFS brings an array of best practices to its cost effective digitization process. Continuously funding R&D efforts, PTFS has determined the best hardware, software, and processes to ensure optimal facilities and techniques are employed for each media type and digitization challenge. Depending on customer preference, PTFS can provide secured scanning (onsite or offsite) to create a digitized version of all hardcopy content or provide hardware and best practices so internal staff can perform the process. Figure 3 depicts a high volume digitization workflow. Digitization provides cost effective approaches for declassification processing when established technologies and best practices are implemented.

The metadata record for each file is an important part of the digitization process. Frequently customers have invested resources to build and maintain metadata records for existing hardcopy files. Digital conversion projects frequently require metadata migration from legacy repositories to preserve the existing metadata investment. PTFS often develops new processes for creating metadata records that are low cost and efficient, particularly when personnel familiar with the document content are scarce. Figure 4 depicts a “profile sheet” that uses bar coding and mark sense to facilitate metadata creation. If no electronic metadata is available, PTFS can also provide cleared staff to perform the metadata keying process or provide best practices to onsite staff.

**Figure 3 – High Volume Digitization Workflow**



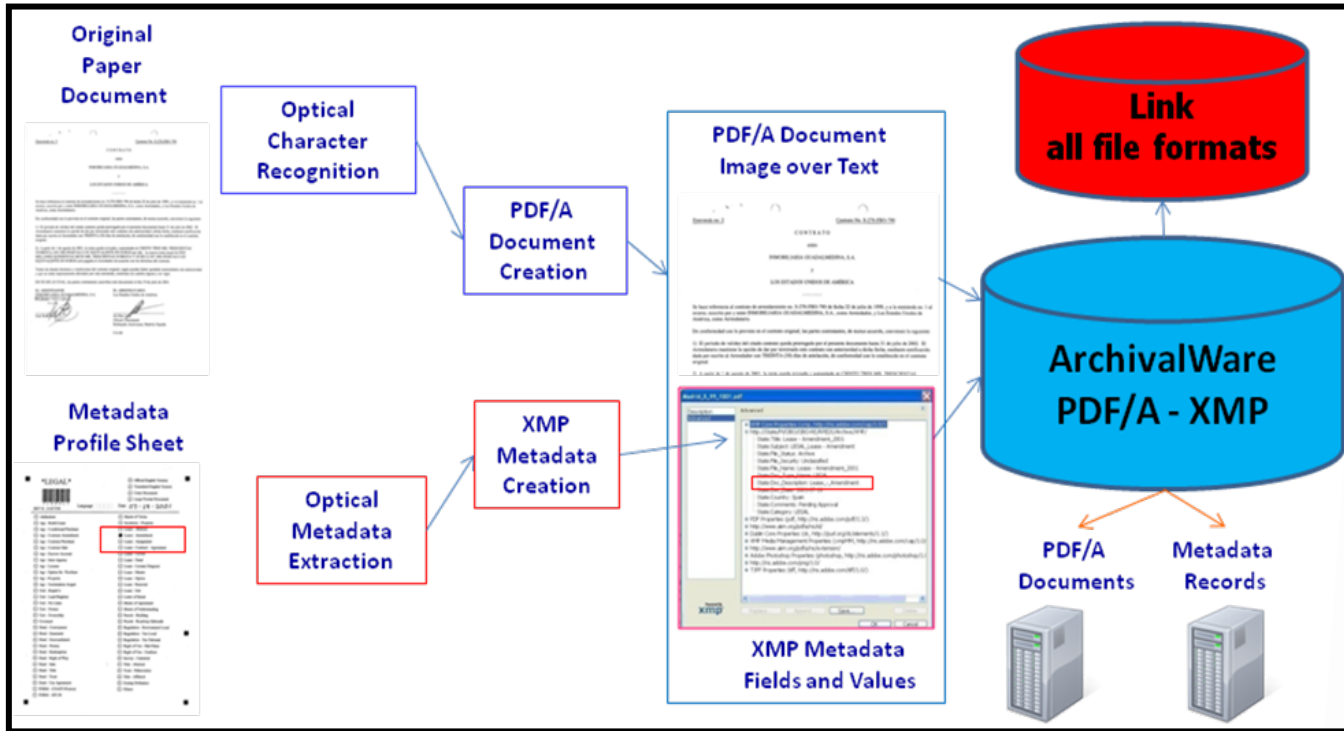


Figure 4 - ArchivalWare DX process for linking Metadata (existing or keyed) to Document

### Ingestion

Once content is in a digital format (scanned or born digital), PTFS will upload and index the entire digital file collection into a client based server, pre-loaded with the ArchivalWare DX application for the development of an automated document review and workflow process. PTFS works with born digital and digitized (scanned) content

If there is a metadata record for each document, they will be loaded and linked to the appropriate documents within ArchivalWare DX. If metadata is present, a GUI is available for authorized users to add or enhance the metadata associated with each document with further detail and applicable information.

During the content ingestion process, ArchivalWare DX utilizes a semantic index process based on terms, expressions and concepts. Compared to simply ranking the frequency of words for indexing, ArchivalWare DX initiates a semantic analysis to discover synonyms and related concepts embedded in semantic networks. This feature can be used to assist the user in identifying other potential words or expressions subject to redaction.

ArchivalWare DX can also ingest a glossary of dirty words (either third party or developed internally by the customer) to be used in a query against the document. The dirty word dictionary can be updated as required when words are added or removed from the dirty word list.

To overcome digitization or hardcopy deficiencies, ArchivalWare DX applies Adaptive Pattern Recognition Processing (APRP) technology to queries, identifying words and phrases that were not accurately OCR'd or

which have been misspelled in the document or in the query. ArchivalWare DX utilizes APRP to reduce false negatives caused by these issues as a part of the ArchivalWare DX's declassification work flow process.

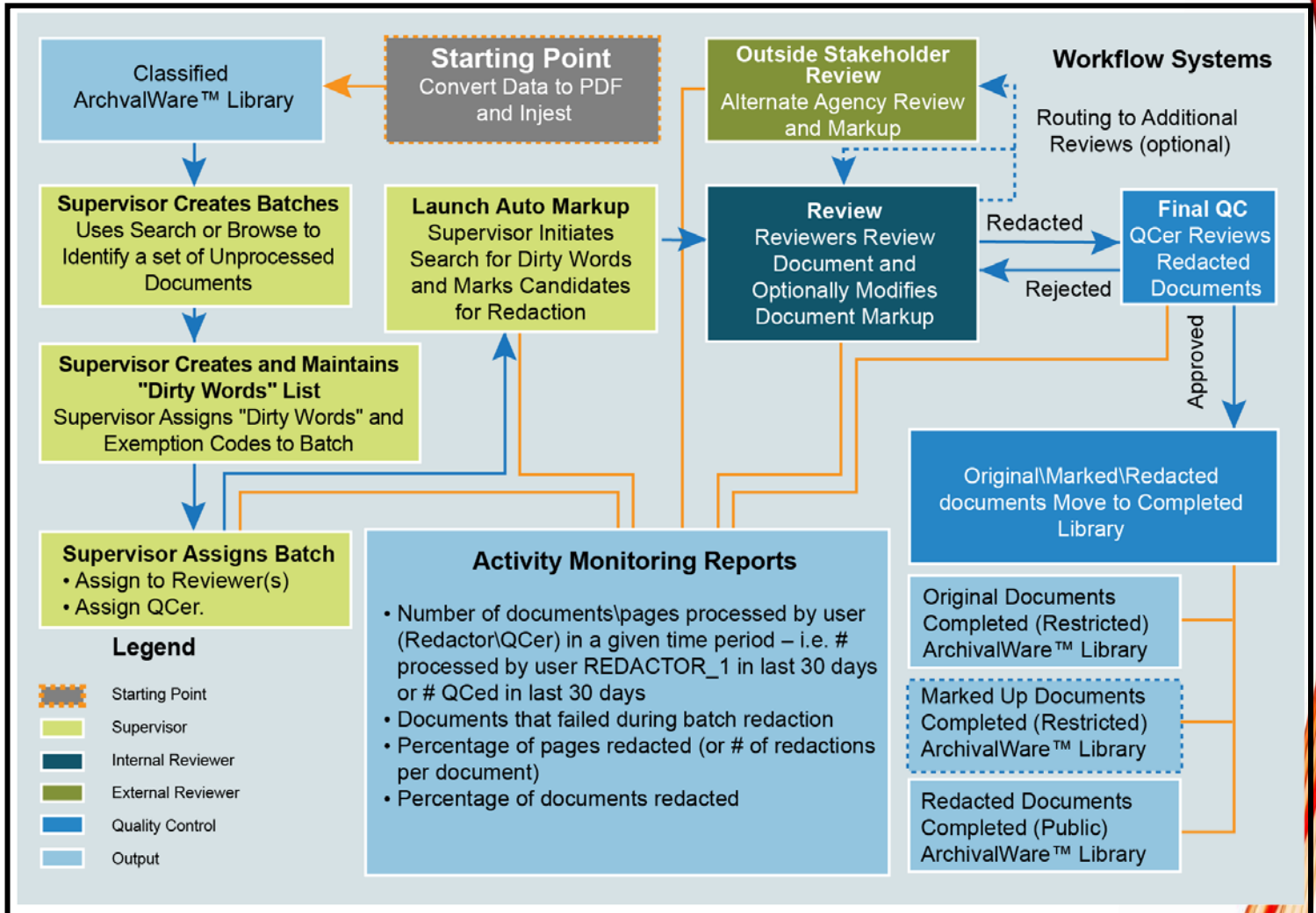
## ArchivalWare DX

### Workflow

ArchivalWare DX is integrated with a robust Java based workflow engine, jBPM, which enables users to rapidly replicate the business process in the application without any programming requirements. This includes the capability for managers to assign documents to declassification analysts, and analysts to accept or reject documents. Others can initiate reassignments, automated document assignment based on dirty word analysis, and automatic forwarding of documents to managers who QC/QA the validation decision of the declassification analysts. Figure 5 displays typical a declassification workflow process.

During this process, metrics can be developed to determine the speed of executing the workflow process using automation.

**Figure 5 - ArchivalWare DX Flexible Workflow**



### Audits and Reports

ArchivalWare DX provides both standard and customizable audit reports for every batch processed with detailed information on status, assigned processor, change types, time and date, priority assignment, productivity metrics, and percentage of documents redacted. Reports can be created or modified to support additional information as needed.

### Searching

ArchivalWare DX provides powerful and effective search capabilities for words or phrases by using a combination of search modes; concept, Boolean and fuzzy text (APRP) and wild cards. This methodology virtually eliminates false negatives which can occur during the data capture process as well as recognizing OCR errors, a common issue in classification initiatives.

### Concept Search (Synonyms)

This type of search expands your search term to include semantically related terms. It uses a network of word associations enabling the expansion of search terms by using variations, synonyms, antonyms, and other relationships to search the entire document text. This allows users to have the most relevant documents delivered to the top of the result list.

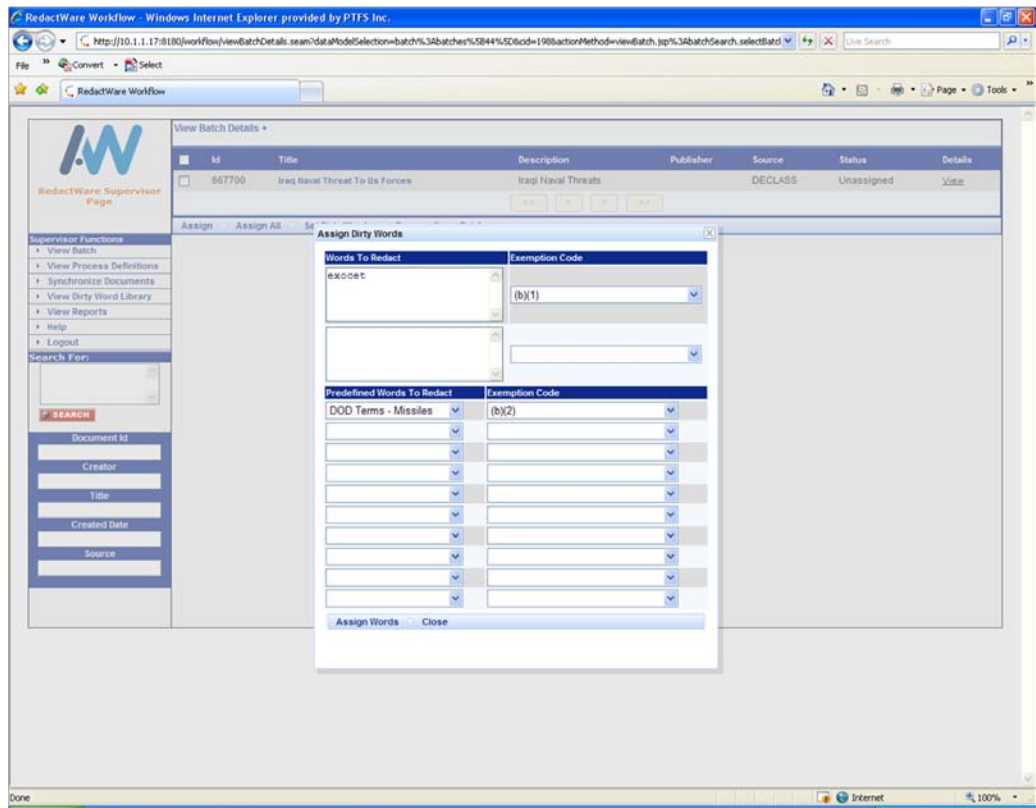
### Pattern Search (Searching OCR Errors)

Pattern APRP searches tolerate spelling errors in either the body of the text or the keyword search. It automatically performs pattern expansion on all keywords based the number of words set by the user, and then ranks the retrieved documents. Pattern searching overcomes spelling differences and deficiencies in OCR quality.

### Dirty Word Glossary

Users can store, retain and modify a list of targeted words or phrases which will be automatically highlighted on any viewed document notifying the user for pending redaction processing. Multiple lists can be viewed, shared, distributed or assigned between users or groups. Figure 6 – Displays Dirty Word Glossary Input screen.

Figure 6- Example of Dirty Word Glossary



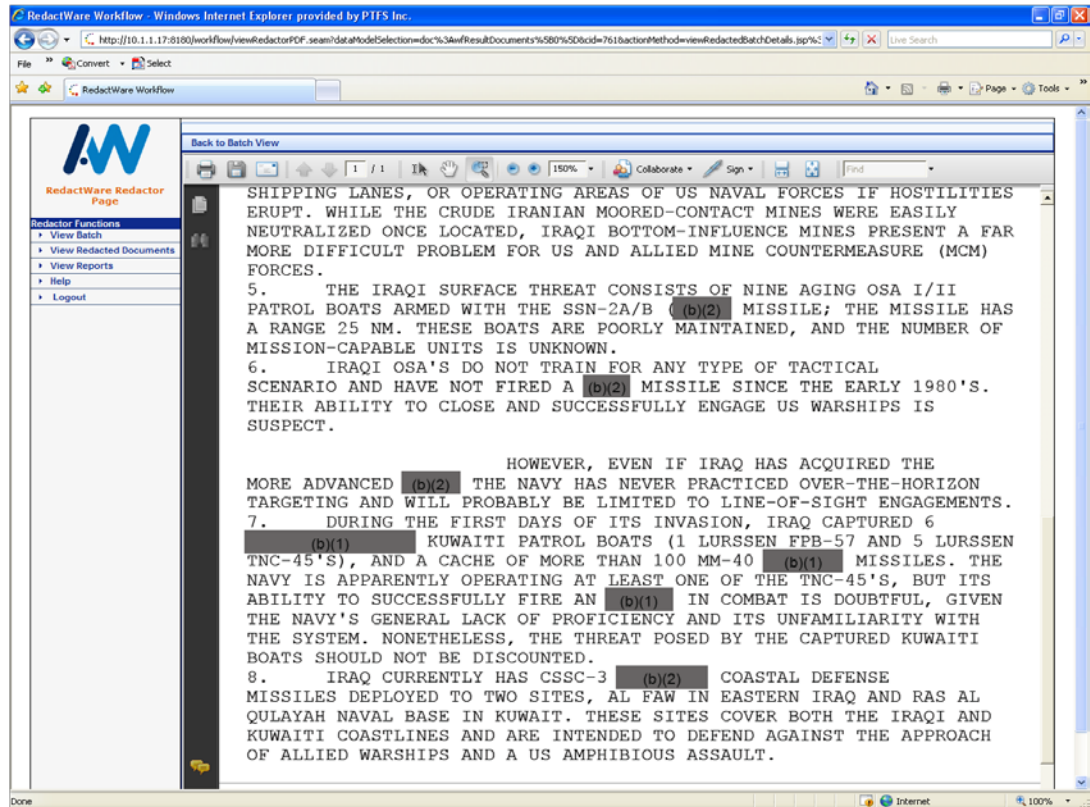
## Redaction

ArchivalWare DX permanently removes any targeted text along with its associated metadata, including both image text as well as hidden text. This ensures against any possibility of identifying or extracting the original word or phrase. Figures 7 displays an example of redacted words with redaction codes.

## PDF/A & XMP Technology

ArchivalWare DX leverages the latest Adobe PDF/A technology that meets ISO standards. ArchivalWare DX is designed to ensure redactions are made to both the image and hidden text layers to prevent leakage of sensitive information. A major advantage of using PDF/A formats is PDF/A's Extensible Metadata Platform

(XMP), a technology that allows you to embed metadata describing a digital file, into the file itself. Using XMP technology, ArchivalWare DX can ingest one file and index both the file content as well as the metadata without a Content Management Solution and a database to connect the metadata to the document. When properly architected, metadata updates immediately modify the XMP data housed in the file allowing portability at any time.



Figures 7 - Redacted Dirty Words with Redaction Codes

## FLEXIBLE SOLUTIONS

PTFS understands the challenges of running enterprise declassification initiatives required by the Presidential Executive Orders. PTFS offers full declassification life cycle support services from hardcopy classified materials to sanitized digitized versions ready for mass consumption. Use of this service transfers the burden of converting content, staffing SMEs with appropriate clearances, manually reviewing every document, and avoiding dangerous errors. This process can be incredibly costly and can strain government organizational resources which may be already operating with reduced funding.

PTFS brings a new approach that can simultaneously minimize cost, reduce errors, increase productivity and effectively process the growing backlog of documents. It doesn't matter where your organization is

during the process, PTFS can fill the gaps or provide total enterprise solutions from the start. For more information visit us online at [www.ptfs.com](http://www.ptfs.com) or contact us at [sales@ptfs.com](mailto:sales@ptfs.com)

## ABOUT PTFS

PTFS is a leading Content and Knowledge Management solution provider. Founded in 1995, PTFS has focused on developing enterprise content management solutions for Federal, state, and local government organizations as well as commercial entities. Staffed with knowledge management experts, PTFS understands the difficulties of managing content across an enterprise, let alone redacting information from content for declassification processing purposes. With these difficulties in mind, five years ago PTFS started developing its declassification solutions. Today PTFS provides full declassification life cycle support, from scanning hardcopy content to implementing ArchivalWare DX redaction and declassification processing software.