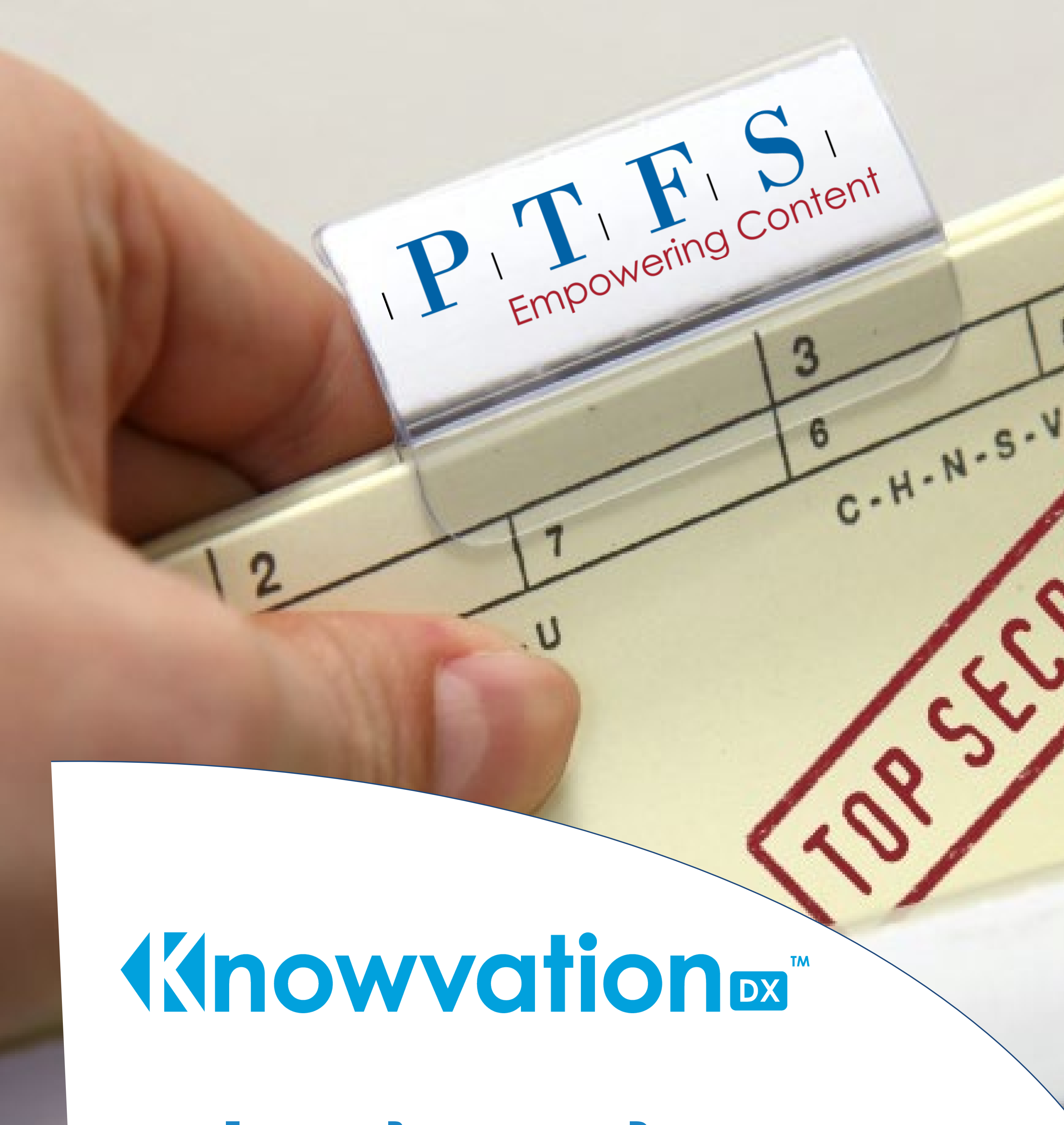

Knowvation DX™

Enterprise Redaction and Declassification

Technology Assisted Redaction and Declassification Processing for Inter/Intra Government Agency Reviews

Executive Summary

Progressive Technology Federal Systems, Inc. (PTFS) has built an application on top of their flagship Knowvation™ enterprise Content Services Platform (eCSP) software to support document redaction and declassification. Known as Knowvation DX™, the application enables a semi-automated process that reduces costs and improves accuracy. Core elements include the ability to search across documents electronically for sensitive words and concepts using a specially developed fuzzy (pattern) search capability that maintains low/no false negatives while minimizing false positives. Knowvation DX has flexible workflow that enables system administrators to make workflow changes quickly without the use of expensive programmers. The application is 100% web based, cloud ready and requires no client-side software -- a feature that facilitates more efficient processing for both inter-agency and intra-agency reviews. Knowvation deployments using Top Secret and Secret cloud services allows for further flexibility and savings.

Next Generation Redaction and Declassification

Knowvation DX is an enterprise solution to semi-automate redaction and declassification. The primary benefits are increased productivity resulting in lower costs, and improved accuracy reducing the chances of an inadvertent spill. Other benefits include:

- **Captures intellectual capital from Subject Matter Experts (SMEs) before they leave -** Heavy ongoing/future attrition requires capturing institutional knowledge
- **Fast reprocessing -** If/when rules change, material can be quickly reprocessed against new rule sets
- **More efficient near real time referral processing**
  - Referrals can be accessed electronically and processed simultaneously
  - Less expensive and shorter cradle-to-grave elapsed time.
- **Web services platform for the future growth and rapid modifications** - New applications with new algorithms can plug in and feed instructions to the redactor.

Knowvation DX Functions

Searching

Knowvation DX provides powerful and effective search capabilities for words or phrases using a combination of search modes: concept, Boolean, fuzzy text and wild cards. These help to eliminate false negatives due to context, OCR errors, misspellings, and typos.

Concept searching expands search terms to include semantically related terms. It uses a network of word associations to expand 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 results list.

Fuzzy text search overcomes spelling errors in the body of the text or the keyword search. It automatically performs pattern expansion on all keywords based on the number of words set by the user, and ranks the retrieved documents. Pattern searching also overcomes spelling differences and deficiencies in Optical Character Recognition (OCR) quality.

Optical Character Recognition (OCR)

Knowvation provides embedded OCR functionality. The OCR processes an image file, creates text-based characters from the image and records the spatial page coordinates for each character. When a PDF file is generated by Knowvation the full-text content becomes searchable and permits the system to display hit-highlights in the file so the user can easily understand why a file was retrieved. Once OCRed text has been created it can be cut and pasted for repurposing or metadata creation and can be saved in a variety of formats. The OCR process can be automated during ingest or on digital objects already in the repository.

PTFS utilizes a specially developed and tuned best-in-industry OCR engine to process images and create the most accurate OCR currently available in the industry. The OCR process can be applied to a diverse set of file formats including image-only PDFs, JPEGs, TIFFs, and other image file types.

While Knowvation functionality includes Fuzzy Text search to overcome OCR inaccuracies when they occur it is most important to capture images with high image quality and perform image enhancement if necessary, to allow the Knowvation OCR engine to perform as accurately as possible. Images should be scanned at 300 DPI or higher but studies have shown that OCR accuracy improvement for standard page images beyond 400 DPI is minimal.

Sensitive Word Glossary

Knowvation DX can ingest a sensitive word and concept list to be used in a query against the document. Sensitive words and concepts are defined as words, phrases, synonyms, acronyms and other text that might indicate a cause for exemption, that another agency has equity interest, or that a Kyl-Lott review is required. 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.

Workflow

Knowvation DX is integrated with a robust workflow engine based on the Java Business Process Model (jBPM), an open source tool which allows users to replicate the declassification business process rapidly in the application without any programming requirements. This includes the ability for managers to assign documents to declassification analysts, and analysts to accept or reject documents. Other staff can initiate reassignments, automated document assignment based on sensitive word analysis, and automatic forwarding of documents to managers who perform Quality Control and Quality Assurance (QC/QA) on the validation decision of the declassification analysts. A system administrator or supervisor has full visibility into the workflow process, and the captured metrics track individual reviewer's productivity and accuracy.

Audits and Reports

Knowvation 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.

Vision with the Cloud

The advent of Top Secret and Secret clouds opens up the potential for further efficiencies. The diagram below portrays a vision for how Knowvation DX, agency applications and future applications can work when applications are modularly constructed and open architected. PTFS provides full redaction and declassification capabilities in a modular manner for agencies to select their components a la carte based on their requirements and existing systems (See Figure 1).

Hit Highlighting and Redaction

Knowvation DX highlights the sensitive words in the document along with the exemption code for the analyst to review. The analyst can highlight additional areas, including text, charts/tables/figures, handwriting in the margin, and other sensitive content. The highlights can be converted to redactions which permanently remove 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. At the end of the process, the original file is maintained and instruction sets are generated to recreate highlighted and redacted documents.

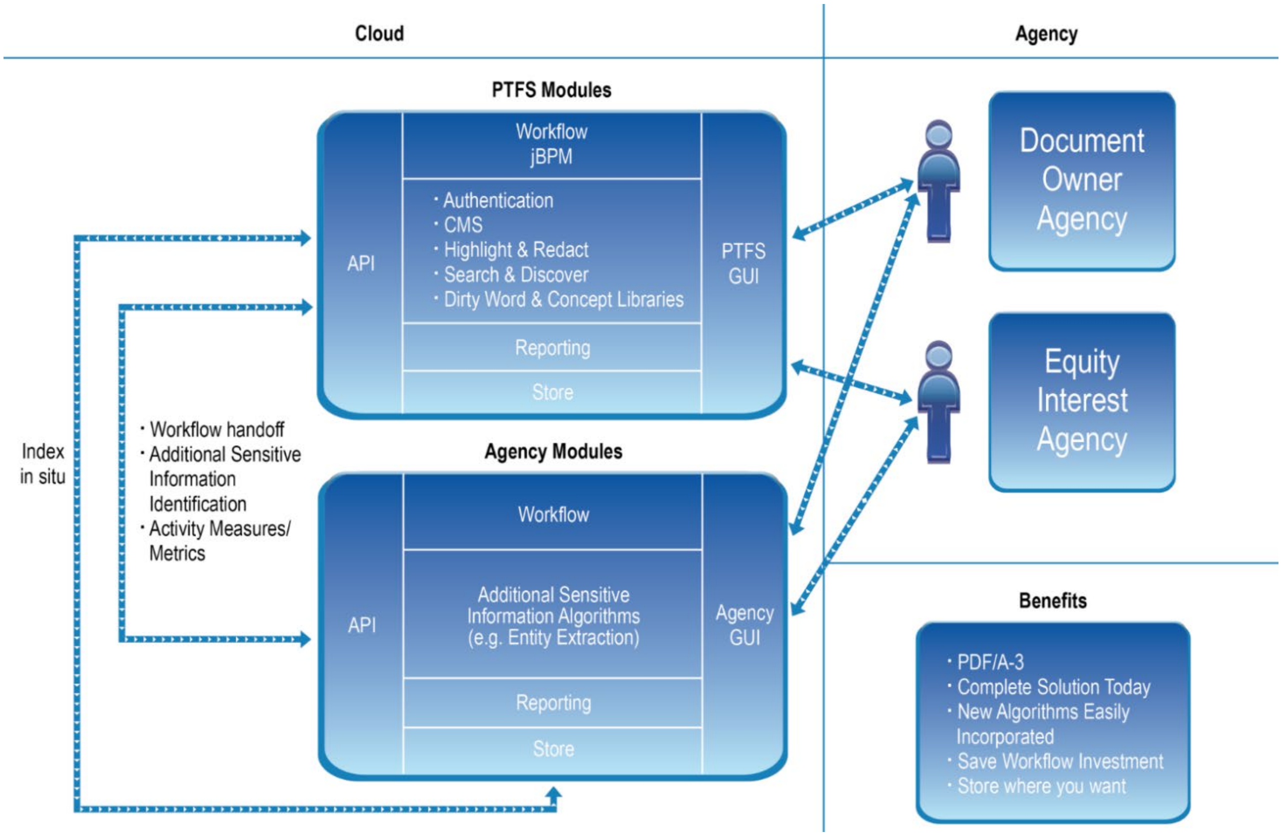

Figure 1: Semi-Automated Redaction and Declassification in the Cloud

PDF/A and XMP Technology

Knowvation DX leverages the latest Adobe PDF/A technology that meets ISO standards. It 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 a rich and extensive metadata describing a digital file to be imbedded in the PDF and contained as a package. Using XMP technology, Knowvation DX can ingest one file and index both the file content as well as the metadata. When properly architected, metadata updates immediately modify the XMP data housed in the file allowing dissemination, portability and NARA compliance.

Machine Learning Add-On

While artificial intelligence systems are the panacea for a redaction and declassification solution, machine learning is a real-world capability that will enhance Knowvation DX's sensitive word and concept library. Information provided by the analyst will enhance the system's ability to assist the reviewer. This capability is planned in the Knowvation DX roadmap and will provide a method for the system to become smarter as highly knowledgeable and trained analysts perform their job. This add-on Knowvation DX capability has been proposed for research and development for the product under a US Air Force Small Business Innovative Research Grant.

Knowvation eCSP Overview

Knowvation eCSP is an enterprise data management and discovery platform. The system allows implementation of a centralized enterprise platform to manage data collected allowing analytical processes to be run on specific data sets anytime. The solution enables methodical processes to be performed to repeat analysis on an ongoing basis as new data is collected. This functionality supports a wide variety of decision support activities. Knowvation provides robust and easy to use ingestion capabilities in addition to federated indexing functionality for data that is located but not ingested by the system. Ingesting, locating, identifying, normalizing, and tagging/selecting an organization's data is a critical requirement before performing any type of analytical process. A continually updated federated index provides near real time data to be served to a wide variety of integrated applications for analysis and business intelligence.

The Knowvation platform is open and designed for flexibility and growth to allow for changes without the need to replace the underlying technology. Open system architecture uses J2EE to provide an open architecture environment with an Application Programming Interface (API) and RESTful services that enables integration with existing and future third-party applications. Knowvation satisfies multiple data, content and knowledge management requirements and will support a number of big data requirements with the following benefits:

1) Easy to learn and intuitive platform allowing data ingestion to manage content and build the organizations data warehouse;
2) Powerful research and discovery functionality across all data types;
3) Robust ability to read/write/tag metadata during or after ingestion;
4) The ability to display and visualize data for a diverse file formats and data types;
5) The ability to enhance and normalize data required for analytical processing; and
6) A federated indexing capability to build a master index of the organization's data, both local and disparate.

The overall Knowvation DX platform application is shown in Figure 2 below.

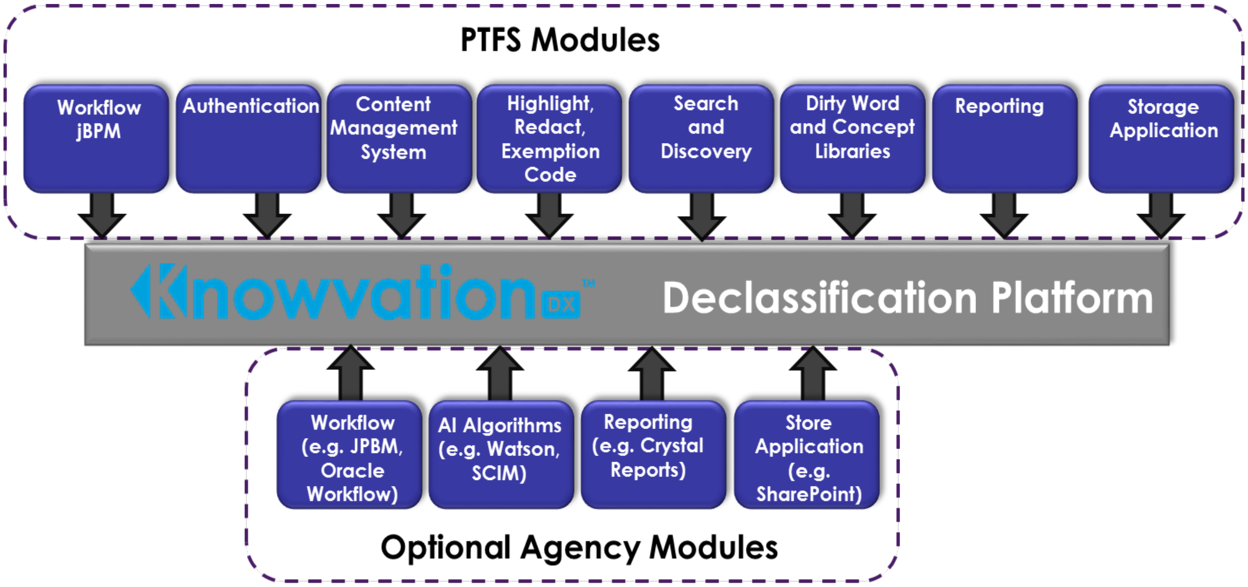
PTFS Modules
Workflow jBPM
Authentication
Content Management System
Highlight, Redact, Exemption Code
Search and Discovery
Dirty Word and Concept Libraries
Reporting
Storage Application
Knowvation DX™
Declassification Platform
Workflow (e.g. JPBM, Oracle Workflow)
AI Algorithms (e.g. Watson, SCIM)
Reporting (e.g. Crystal Reports)
Store Application (e.g. SharePoint)
Optional Agency Modules

Figure 2: Knowvation DX Declassification Platform

Cloud Based SaaS Offering

Knowvation DX is available on the Amazon Web Services commercial EC2 Marketplace as SaaS offering. The product is also available on the government's secure C2S Commercial Cloud Services contract from the Marketplace or via a BYOL mechanism.

Other Knowvation™ enterprise Content Services Platform (eCSP) Offerings

In addition to Knowvation DX, PTFS has built one other vertical market applications based on the eCSP platform (see box to the right).

- **Knowvation GS**™
  Geospatial analyst interface to manage, search, and discover diverse multi-intelligence collection with or without geo-tagging
- **Knowvation RM**™
  Provides a consistent system for organizing and managing regulated content, facilitates retention and dispositioning, and controls access and use.

About PTFS

Founded in 1995, PTFS has focused on developing enterprise content management solutions for Federal, state, and local government organizations, as well as commercial entities. Our experience implementing and integrating Knowvation and related products and services to satisfy customers' unique needs in multiple environments has made PTFS an industry leading solution provider. There are more than 100 employees in the company, including technical experts in database management systems, security, records management, geospatial analysis and data communications, as well as seasoned project managers.

Updated July 28, 2020