110 Horizon Drive, Suite 210, Raleigh, NC 27615
919.459.2081

Automated Redaction Technology White Paper

Adopted by the PRIA Board on June 16, 2021

http://www.pria.us

PROPERTY RECORDS INDUSTRY ASSOCIATION

Copyright Notice, License, Disclaimer
For
PRIA Completed Work Product

July 2021

A. **COPYRIGHT NOTICE**: Copyright © 2021 – Property Records Industry Association ("PRIA"). All rights reserved.

B. **LICENSE:** This completed PRIA work product document (the "Completed Work") is made available by PRIA to members and the general public for review, evaluation and comment only. This document is under development and not a final version.
PRIA grants any user ("Licensee") of the Completed Work a worldwide, royalty-free, non-exclusive license ("License") to reproduce the Completed Work in copies, and to use the Completed Work and all such reproductions solely for purposes of reviewing, evaluating and commenting upon the Completed Work. NO OTHER RIGHTS ARE GRANTED UNDER THIS LICENSE AND ALL OTHER RIGHTS ARE EXPRESSLY RESERVED TO PRIA. Without limiting the generality of the foregoing, PRIA does not grant any right to: (i) prepare proprietary derivative works based upon the Completed Work, (ii) distribute copies of the Incomplete Work to the public by sale or other transfer of ownership, or (iii) display the Completed Work publicly. Comments on the Completed Work must be sent to PRIA.

Any reproduction of the Completed Work shall reproduce verbatim the above copyright notice, the entire text of this License and the entire disclaimer below under the following header:

This document includes Completed Works developed by PRIA and some of its contributors, subject to PRIA License. "PRIA" is a trade name of the "Property Records Industry Association." No reference to PRIA or any of its trademarks by Licensee shall imply endorsement of Licensee's activities and products.

C. **DISCLAIMER: THIS COMPLETED WORK IS PROVIDED "AS IS." PRIA AND THE AUTHORS OF THIS INCOMPLETE WORK MAKE NO REPRESENTATIONS OR WARRANTIES (i) EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE OR NON-INFRINGEMENT; (ii) THAT THE CONTENTS OF SUCH COMPLETED WORK ARE FREE FROM ERROR OR SUITABLE FOR ANY PURPOSE; AND, (iii) THAT IMPLEMENTATION OF SUCH CONTENTS WILL NOT INFRINGE ANY THIRD-PARTY PATENTS, COPYRIGHTS, TRADEMARKS OR OTHER RIGHTS. IN NO EVENT WILL PRIA OR ANY AUTHOR OF THIS COMPLETED WORK BE LIABLE TO ANY PARTY FOR ANY DIRECT, INDIRECT, SPECIAL OR CONSEQUENTIAL DAMAGES FOR ANY USE OF THIS COMPLETED WORK, INCLUDING, WITHOUT LIMITATION, ANY LOST PROFITS, BUSINESS INTERRUPTION, LOSS OF PROGRAMS OR OTHER DATA ON ANY INFORMATION HANDLING SYSTEM OR OTHERWISE, EVEN IF PRIA OR THE AUTHORS OR ANY STANDARD-SETTING BODY CONTRIBUTORS TO THIS COMPLETED WORK ARE EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.**

Table of Contents

Executive Summary

The purpose of automated document redaction software is to provide for removal of varying types of privacy information from documents with as little human intervention as possible. Information to be redacted is often thought of as being a finite set of data fields within a given industry's document universe (land recordings, court filings, patient records, etc.). Data fields and redaction requirements often vary per installation and over time, which requires the redaction software to be adaptable. The goals of automated redaction technology are to achieve the highest levels of accuracy, take the least amount of time to process, and minimize human labor for document review.

This white paper supplements other PRIA redaction resources. Visit the PRIA Resource Library on the website (www.pria.us).

Technical Challenges

Traditionally, document formats processed within land records were Tagged Image File Format (TIFF). With inclusion of Portable Document Format (PDF), redaction technology now needs to address both visible image layers, as well as non-visible PDF text layers and metadata objects.

Data to be redacted has also grown in complexity. Simple redactions such as social security and credit card numbers now also include more complex data such as addresses and names. Complex redactions often require not just the presence of data, but the role of the data elements to be determined prior to redaction. It is not just a matter of redacting all addresses found in a document. Redaction algorithms must now accommodate knowing if the address is the appropriate one to redact. Redacting all addresses in a document when only a law enforcement officer's home address needs to be redacted presents a new set of challenges that advancements in redaction technology solve.

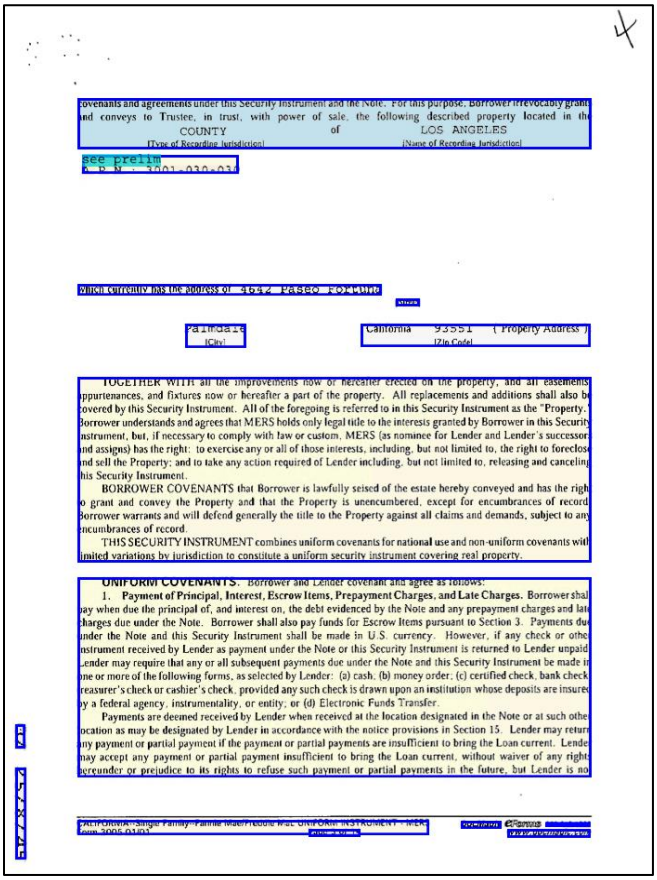
The Process of Redaction

Present day automated redaction processing workflows are well established. First, electronic document pages are converted to searchable text via optical character recognition (OCR) processing.

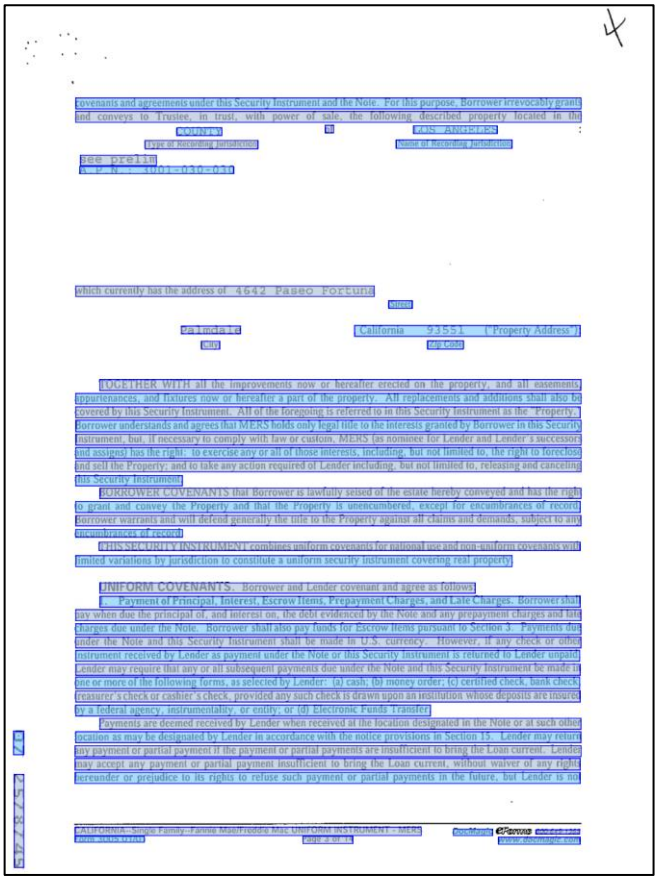
Optical Character Recognition

The basics of OCR technology are as follows. OCR software first performs a “full page read” of each document page, identifying all graphical objects and their boundaries on the page.

Full-page read object detection

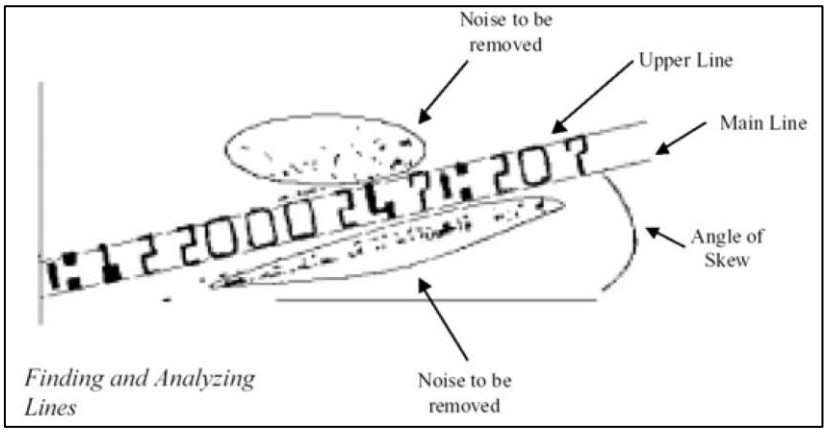


Text line detection



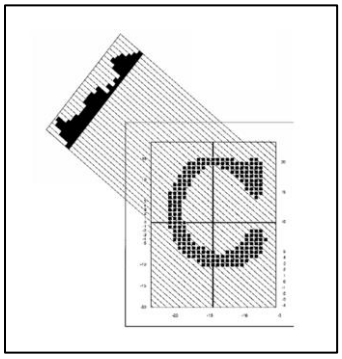
Then, for each graphical object discovered, OCR attempts to establish boundary lines within the graphical object where text may be located. Then, within each bounded line of text, OCR attempts to establish locations of words, and the location of graphical character objects within each word.

OCR word segmentation



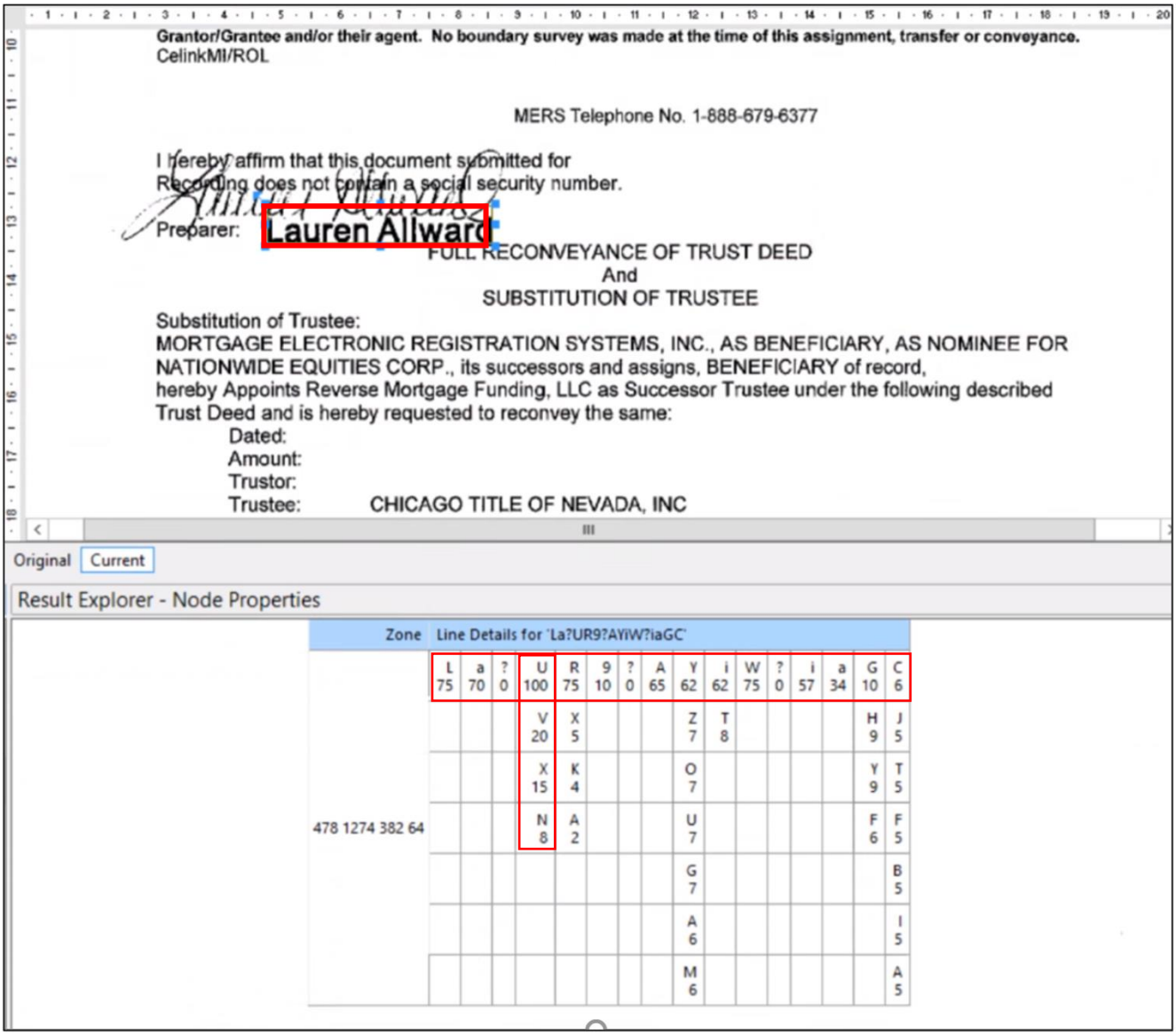
For each graphical character object found, OCR software attempts to determine the corresponding text character that most closely represents the image symbol. It does so using "best fit" pattern matching algorithms that take into consideration the symbol's size, shape, and curvature data points. Graphical character data points derived from the image are matched against a set of previously trained character symbol data points stored in the OCR engine.

OCR histogram character matching



The matching process typically produces several text character candidates, with each having an associated “matching” confidence value. The character candidates and their confidence values are then used by additional OCR statistical algorithms to further refine the selection of a correct character. This additional processing first takes into consideration surrounding text characters and their probability of occurrence (trigram analysis) for each candidate character. For alpha text, character candidates are further evaluated as being correct in their use producing correct or incorrect words (dictionary lookups).

Alternative OCR values for “Lauren Allward” text

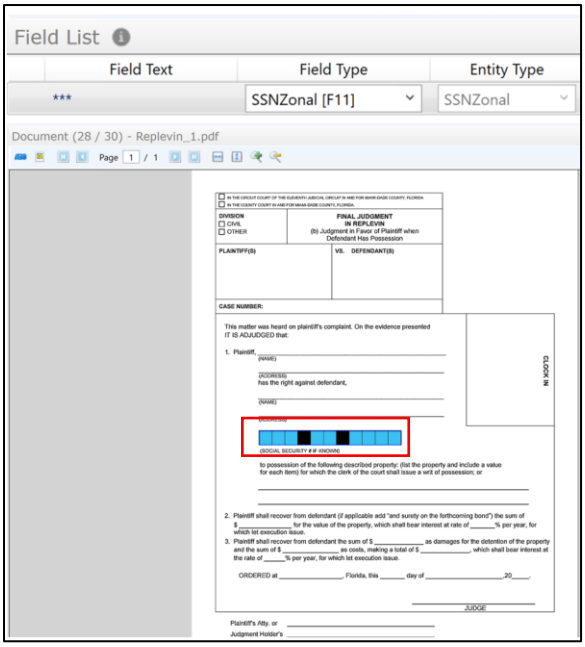


In a world of perfect document quality, OCR is still a challenging task given varying font types and image resolutions (DPI), as well as image artifacts for which OCR engines must accommodate.

Intelligent Character Recognition

While OCR technology is responsible for the transformation of machine-printed image text, Intelligent Character Recognition (ICR), technically similar to OCR, addresses transforming the handprint on documents to useable text. Unlike OCR, traditional ICR does not produce accurate handprint results on full-page reads. To be successful traditional ICR must be directed to specific areas on a page where handprint text may be located.

Definition of ICR Handprint Detection Zone



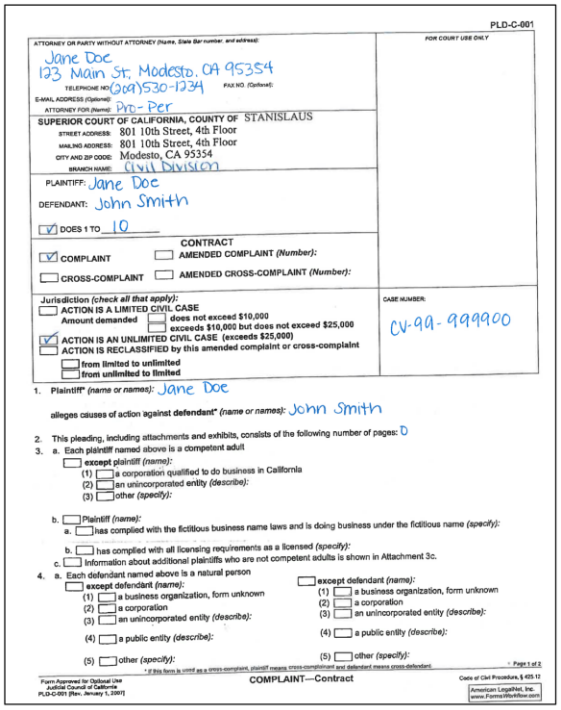
This process is problematic, as unless redacting structured form documents, one does not know in advance where handprint text may be located. Recently, OCR technology has seen the introduction of neural network engines from several OCR software vendors. Results have shown neural network OCR to produce accurate full-page handprint reads. Neural network OCR solves the problem with traditional OCR missing redactions of free form handprint that often occurs by human annotation of Personally Identifiable Information to documents. Neural network OCR has been shown to improve the recognition rates of poor-quality machine-print documents as well.

Below are results of testing with 50 documents comparing three OCR engines on a per character recognition basis. The error rate includes both incorrectly recognized and missed characters.

Engine	Error Rate	Expected Characters	Detected Characters	Wrong Characters
Neural Network OCR	.48%	77034	76811	368
FineReader 12	.90%	77034	76918	695
Tesseract 4.1	1.82%	77034	77141	1500

Examples of performing OCR full-page read on handprinted text.

Handprint problem – Zero Results



PLD-C-001

ATTORNEY OR PARTY WITHOUT ATTORNEY (Name, State Bar number, and address): Jane Doe 123 Main St, Modesto, CA 95354 TELEPHONE NO: (209)530-1234 FAX NO. (Optional): E-MAIL ADDRESS (Optional): ATTORNEY FOR (Name): Pro-Per

FOR COURT USE ONLY

SUPERIOR COURT OF CALIFORNIA, COUNTY OF STANISLAUS
STREET ADDRESS: 801 10th Street, 4th Floor
MAILING ADDRESS: 801 10th Street, 4th Floor
CITY AND ZIP CODE: Modesto, CA 95354
BRANCH NAME: Civil Division

PLAINTIFF: Jane Doe
DEFENDANT: John Smith
☑ DOES 1 TO 10

CONTRACT
☑ COMPLAINT ☐ AMENDED COMPLAINT (Number):
☐ CROSS-COMPLAINT ☐ AMENDED CROSS-COMPLAINT (Number):

Jurisdiction (check all that apply):
☐ ACTION IS A LIMITED CIVIL CASE
Amount demanded ☐ does not exceed $10,000 ☐ exceeds $10,000 but does not exceed $25,000
☑ ACTION IS AN UNLIMITED CIVIL CASE (exceeds $25,000)
☐ ACTION IS RECLASSIFIED by this amended complaint or cross-complaint
☐ from limited to unlimited ☐ from unlimited to limited

CASE NUMBER: CV-99-999900

1. Plaintiff (name or names): Jane Doe
alleges causes of action against defendant (name or names): John Smith
2. This pleading, including attachments and exhibits, consists of the following number of pages: 0
3. a. Each plaintiff named above is a competent adult
☐ except plaintiff (name):
(1) ☐ a corporation qualified to do business in California
(2) ☐ an unincorporated entity (describe):
(3) ☐ other (specify):
b. ☐ Plaintiff (name):
a. ☐ has complied with the fictitious business name laws and is doing business under the fictitious name (specify):
b. ☐ has complied with all licensing requirements as a licensed (specify):
c. ☐ Information about additional plaintiffs who are not competent adults is shown in Attachment 3c.
4. a. Each defendant named above is a natural person
☐ except defendant (name):
(1) ☐ a business organization, form unknown
(2) ☐ a corporation
(3) ☐ an unincorporated entity (describe):
(4) ☐ a public entity (describe):
(5) ☐ other (specify):
☐ except defendant (name):
(1) ☐ a business organization, form unknown
(2) ☐ a corporation
(3) ☐ an unincorporated entity (describe):
(4) ☐ a public entity (describe):
(5) ☐ other (specify):

* If this form is used as a cross-complaint, plaintiff means cross-complainant and defendant means cross-defendant.

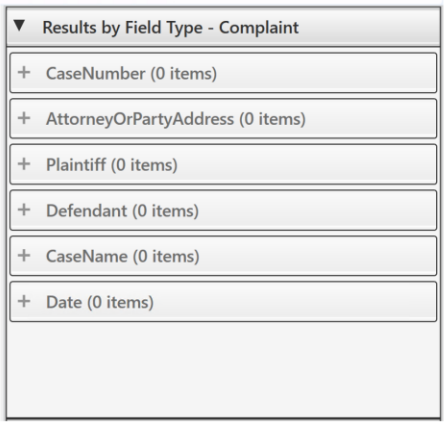
Page 1 of 2

Form Approved for Optional Use Judicial Council of California PLD-C-001 [Rev. January 1, 2007]

COMPLAINT—Contract

Code of Civil Procedure, § 425.12

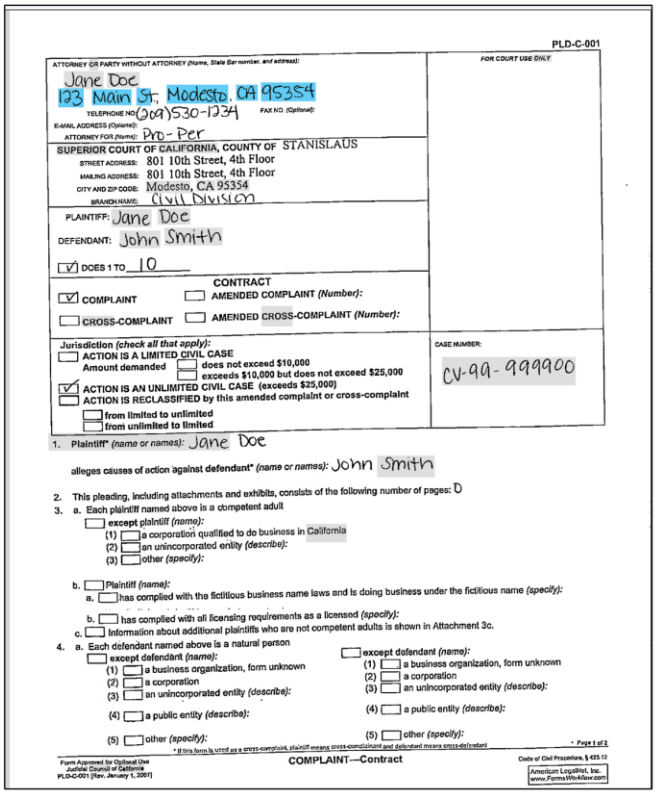
American LegalNet, Inc. www.FormsWorkflow.com



Results by Field Type - Complaint

+ CaseNumber (0 items)
+ AttorneyOrPartyAddress (0 items)
+ Plaintiff (0 items)
+ Defendant (0 items)
+ CaseName (0 items)
+ Date (0 items)

Neural Network OCR – Solves Handprint problem



PLD-C-001

ATTORNEY OR PARTY WITHOUT ATTORNEY (Name, State Bar number, and address): Jane Doe 123 Main St, Modesto, CA 95354 TELEPHONE NO: (209)530-1234 FAX NO. (Optional): E-MAIL ADDRESS (Optional): ATTORNEY FOR (Name): Pro-Per

FOR COURT USE ONLY

SUPERIOR COURT OF CALIFORNIA, COUNTY OF STANISLAUS
STREET ADDRESS: 801 10th Street, 4th Floor
MAILING ADDRESS: 801 10th Street, 4th Floor
CITY AND ZIP CODE: Modesto, CA 95354
BRANCH NAME: Civil Division

PLAINTIFF: Jane Doe
DEFENDANT: John Smith
☑ DOES 1 TO 10

CONTRACT
☑ COMPLAINT ☐ AMENDED COMPLAINT (Number):
☐ CROSS-COMPLAINT ☐ AMENDED CROSS-COMPLAINT (Number):

Jurisdiction (check all that apply):
☐ ACTION IS A LIMITED CIVIL CASE
Amount demanded ☐ does not exceed $10,000 ☐ exceeds $10,000 but does not exceed $25,000
☑ ACTION IS AN UNLIMITED CIVIL CASE (exceeds $25,000)
☐ ACTION IS RECLASSIFIED by this amended complaint or cross-complaint
☐ from limited to unlimited ☐ from unlimited to limited

CASE NUMBER: CV-99-999900

1. Plaintiff (name or names): Jane Doe
alleges causes of action against defendant (name or names): John Smith
2. This pleading, including attachments and exhibits, consists of the following number of pages: 0
3. a. Each plaintiff named above is a competent adult
☐ except plaintiff (name):
(1) ☐ a corporation qualified to do business in California
(2) ☐ an unincorporated entity (describe):
(3) ☐ other (specify):
b. ☐ Plaintiff (name):
a. ☐ has complied with the fictitious business name laws and is doing business under the fictitious name (specify):
b. ☐ has complied with all licensing requirements as a licensed (specify):
c. ☐ Information about additional plaintiffs who are not competent adults is shown in Attachment 3c.
4. a. Each defendant named above is a natural person
☐ except defendant (name):
(1) ☐ a business organization, form unknown
(2) ☐ a corporation
(3) ☐ an unincorporated entity (describe):
(4) ☐ a public entity (describe):
(5) ☐ other (specify):
☐ except defendant (name):
(1) ☐ a business organization, form unknown
(2) ☐ a corporation
(3) ☐ an unincorporated entity (describe):
(4) ☐ a public entity (describe):
(5) ☐ other (specify):

* If this form is used as a cross-complaint, plaintiff means cross-complainant and defendant means cross-defendant.

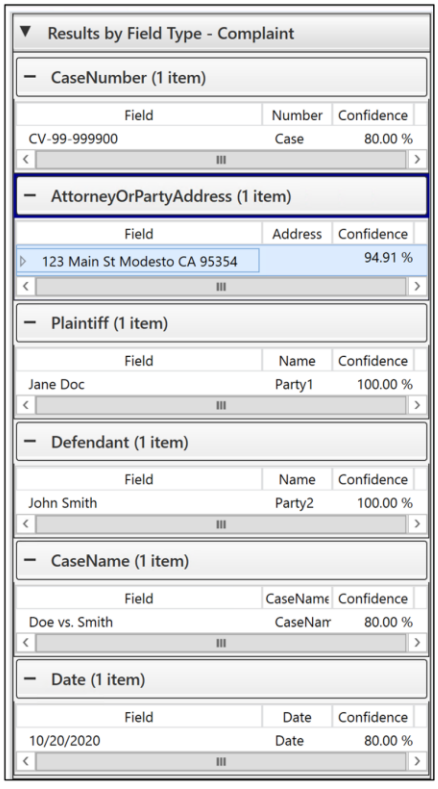
Page 1 of 2

Form Approved for Optional Use Judicial Council of California PLD-C-001 [Rev. January 1, 2007]

COMPLAINT—Contract

Code of Civil Procedure, § 425.12

American LegalNet, Inc. www.FormsWorkflow.com



Results by Field Type - Complaint

− CaseNumber (1 item)

Field	Number	Confidence
CV-99-999900	Case	80.00 %

− AttorneyOrPartyAddress (1 item)

Field	Address	Confidence
123 Main St Modesto CA 95354		94.91 %

− Plaintiff (1 item)

Field	Name	Confidence
Jane Doc	Party1	100.00 %

− Defendant (1 item)

Field	Name	Confidence
John Smith	Party2	100.00 %

− CaseName (1 item)

Field	CaseName	Confidence
Doe vs. Smith	CaseNam	80.00 %

− Date (1 item)

Field	Date	Confidence
10/20/2020	Date	80.00 %

The application of both OCR and ICR technology to document images produces necessary text data for input to automated redaction software systems. The accuracy of OCR and ICR processing has a direct effect on final redaction results.

As for PDF documents that contain existing text layers, state-of-the-art OCR software will recognize the presence of the text layer, perform OCR and ICR and then determine whether the PDF or OCR text has higher accuracy. It might be assumed PDF text has 100 percent accuracy, but only in PDF documents "digitally born" from word processing or other electronic document creation software. As significant quantities of PDF documents are created from Multi-Function Printers (MFP) utilizing embedded legacy OCR, they often contain poor quality text layers.

Locating Redaction Data

After transforming the document page images to searchable text, the next step in redaction workflow is to locate the data to be redacted. Depending on the document types being processed, different data location algorithms are applied.

Structured Documents

For structured form-based documents, zone-specific location is usually the preferred method of locating data. Given that data fields on specific forms will always be at the same location on all pages, redaction software can make use of this knowledge.

In processing structured documents, redaction software first classifies the document as a specific form type. It does this by comparing the present document against a library of forms typically defined during system configuration. Often OCR processing is not necessary, as the form can be identified by image processing using image artifacts such as seals, stamps, horizontal and vertical lines, as well as consistent text blocks.

Having identified the form type, redaction software can then redact the appropriate image zones. Advanced redaction software can also handle different skew and distortion ratios on images to minimize the number of forms required in its library. A significant benefit of form-based redaction is poor quality text, handprint and even cursive script data can be successfully redacted with 100 percent accuracy.

Unstructured Documents

Many documents that require redaction are not structured. Therefore, OCR processing to convert the image to usable text is necessary. Text pattern matching software locates the data to redact. Created patterns provided with the redaction software or defined during system configuration are applied to the document text.

Regular Expression Pattern Matching

It is important to note the standard representation of matching patterns is accomplished with the use of regular expressions. Regular expressions, or regex, is a text-matching language created to simplify pattern matching within text strings. Regex allows for character specifications to be searched without providing the different text variations that can occur, allowing consolidation of search patterns.

As an example, the following regular expression can be used to find numbers that are of the format 999-99-9999 and are also valid social security numbers.

*regex = "**^(?!666|000|9\\d{2})\\d{3}-(?!00)\\d{2}-(?!0{4})\\d{4}$**"*

> ***^** represents the start of the string.*
>
> ***(?!666|000|9\\d{2})\\d{3}** represents the first 3 digits should not start with 000, 666, or be between 900 and 999.*
>
> ***–** represents the string followed by a hyphen (-).*
>
> ***(?!00)\\d{2}** represents the next 2 digits should not start with 00 and it should be any digits from 01-99.*
>
> ***–** represents the string followed by a hyphen (-).*
>
> ***(?!0{4})\\d{4}** represents the next 4 digits can't be 0000.*
>
> ***$** represents the end.*

From our simple, well-formed social security number example, regular expressions are a programming language requiring expertise to construct and maintain. Regular expressions need to adjust for OCR errors in document text, which dramatically increases their complexity. Redaction software vendors typically provide a standard set of redaction data fields, with accompanying regular expressions, and create additional ones as needed.

Entity and Role Detection

There are two components in finding data to redact. For well-formed data, having well defined and known numeric patterns such as social security numbers, all occurrences of such text are first located in the document using specific regexes. This process is known as entity detection, or the finding of specific well-formatted text strings.

Next, the role of the entity must be established. The redaction software uses regexes to look for labels, words, phrases or even sentences, which provide additional clarification to the meaning of the entity located.

In the case of a social security number, entity role detection may include pattern matching to find the label "social security number:" immediately to the left of a previously detected number.

If unsuccessful, the redaction software will attempt matching with any other entity role patterns the software vendor has included in its product. The process of finding entity locations and then entity roles is known as backward searching.

There is an alternative process of forward searching, where potential entity roles are first established, and then appropriate entity detection is performed. Backward searching is more common as it requires fewer processing resources because as there are fewer regex patterns to be processed. Forward searching is generally applied to documents having handwritten text, as OCR errors often prevent backward searching from finding well-formed roles and entities, resulting in missed redactions.

Redaction Confidence Values

During entity as well as role search operations, redaction software typically creates confidence values for each redaction field. This is different from the OCR character confidence values.

Redaction confidence values determine how close the document text matches the vendor's expected redaction patterns. Exact matches are 100 percent, and less than exact matches have corresponding decreases in confidence percentages. These redaction confidence numbers are often used to determine what documents need manual review. They are typically displayed to the end-user to highlight redactions on which they should focus.

Machine Learning

Automated redaction software relies upon the use of regexes to locate text to redact. Traditionally, regular expressions are constructed by software developers after analyzing documents having data present to redact. A significant development in automated redaction software has been the introduction of machine learning technology to replace this human effort. Machine learning algorithms perform document analysis and regular expression creation much faster and more accurately than humans can.

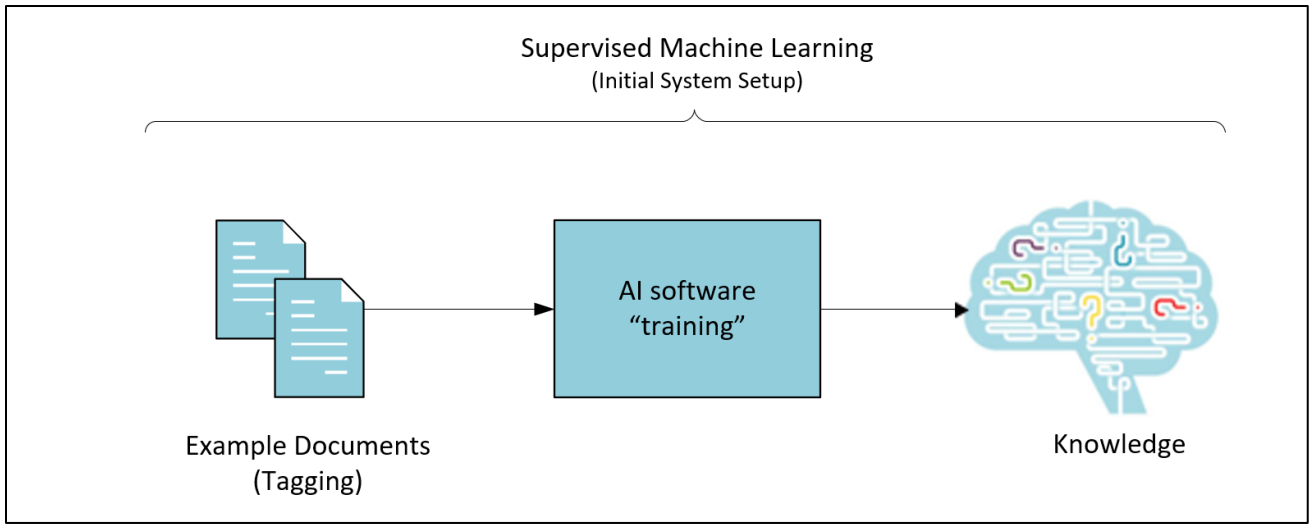
In addition to automatic construction of regexes, machine learning typically incorporates speech decomposition and natural language processing (NLP) algorithms. NLP provides for the meaning of words to be determined. NLP is important in eliminating false redactions for words having the same spelling (located by regex pattern matching) but different meanings.

For example, consider the redaction of the name Sue Smith. In the case of regex, finding a match to the text "SUE SMITH" the presence of the match is not enough to determine if it would be a correct redaction. Is the text found by a regex a person's name that should be redacted? Or, does it have the meaning of suing the Smith party which should not be redacted? NLP solves the problem of finding and redacting words that have the same spelling but different meanings (homographs).

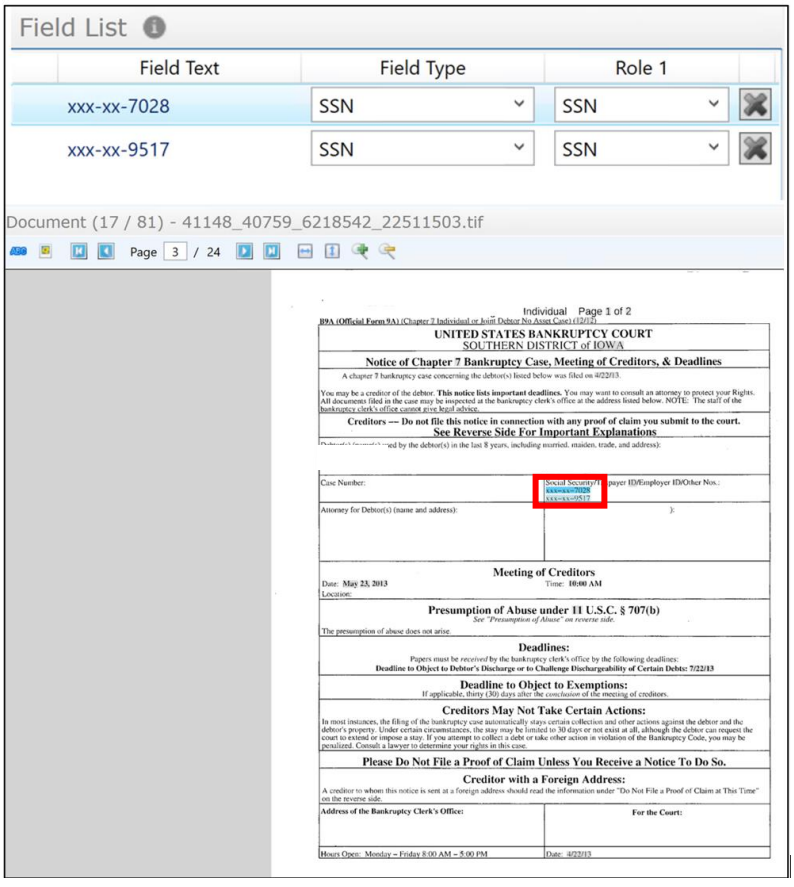
Typical machine learning consists of two stages, supervised learning and online learning.

Supervised Learning

In supervised learning, examples of data to redact are created by identifying or "tagging" data on documents using the redaction vendor's software. Tagging provides input for the supervised learning algorithms, which construct knowledge of how to locate data requiring redaction. Supervised learning is performed on all documents provided as training examples. Providing diverse training documents increases redaction knowledge and accuracy.



Example tagging masked SSN fields



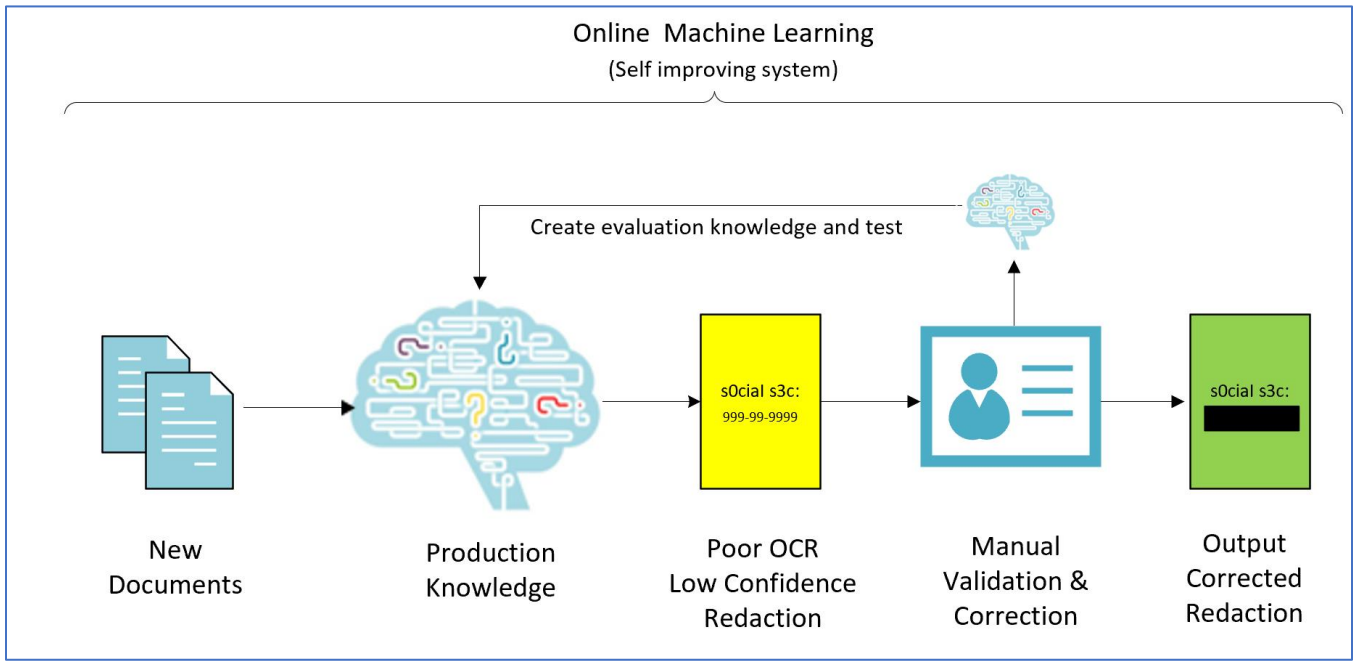
Online learning

Online learning provides for automatic refinement of redaction knowledge during production processing.

In production redaction workflows, documents of low confidence are still presented to users for manual validation. However, changes made by users (adding, deleting, or moving redaction zones) are automatically captured and used to create new versions of knowledge. It is important to note that these new versions of knowledge cannot be immediately used in production processing. It is quite possible validation operators have made incorrect modifications which would result in their mistakes being present in the new version of knowledge.

Vendors differ in their approach to evaluation and use of new versions of knowledge. Some vendors provide for a second manual review of the changes made by the first validator, while others provide advanced automated solutions. Automated solutions employ different mechanisms, but all test the new version of knowledge against additional documents. Automated testing is used to verify the new knowledge increases redaction accuracy. New versions of knowledge that increase accuracy are promoted to production knowledge and the online learning cycle repeats itself. Typical online learning cycles range from every hour to once daily.

Online learning automatically adjusts redaction knowledge, accommodating for changes in documents and refinements made by validation operators. Modern online learning redaction systems can produce consistent accuracy of 95 to 99 percent with little to no human intervention.



Redaction Output

Once document areas to redact have been identified, different processing outcomes exist. Is the redaction system going to permanently remove the data from the document? Does it need to provide redaction co-ordinates that will be used to mask the image when it is produced for display? In cases where redaction coordinates are provided, redaction processing is complete. Note that coordinates are typically useful only for documents that do not contain text layers or have metadata that requires redaction.

In cases where permanent data removal must occur, redaction software must first alter the document's image layer. For bitonal TIFF documents, redaction image areas have all their bits set to either a black (0) or white (1) value. For greyscale images that have "depth" for each displayed image bit, the values are black (0x00) and white (0xff). For documents that also have a text layer, the corresponding OCR text must be removed. This is either performed in-place, or with an additional OCR pass performed on the newly redacted image to generate sanitized text. For in-place removal, redaction systems may provide alternative text to indicate what type of data was removed (i.e., SSN xxx-xx-xxxx replaced with "removed social security number"). This is quite helpful for ADA compliant screen readers to provide context in their text-to-speech translations.

Conclusions

Since the early 2000s, automated redaction software has been applied to land record documents and significant advancements have been made. Multiple redaction algorithms and technologies exist to reliably locate and redact data. Neural OCR network technology now exists to process handprint with the same accuracy as machine print. Machine learning has taken the place of manual document analysis and hand coding of regular expressions used to find data to redact. Natural language processing technology adds additional information to determine what data is relevant. Online learning allows redaction systems to self-adjust in near real time keeping redaction accuracy consistently high with little to no manual intervention.

A benefit of the current redaction technology is that the data location algorithms can be used in both redaction and auto-indexing. Organizations should consider the benefit of auto-indexing technology in addition to providing privacy protection.