Redaction Best Practices

Adopted by the PRIA Board on December 10, 2014

http://www.pria.us
PRIA Copyright Notice, Disclaimer and End-User License

Version 1.1 November 2003 (the “PRIA License” or the “License”)

This document or software (the “Work”) is published by the Property Records Industry Association (“PRIA”).

Copyright © 2014 - writers listed in the Work (collectively or individually, a “Licensor”). All rights reserved.

Subject to this License, Licensor hereby grants any user of this document or software (“Licensee”) a worldwide, royalty-free, irrevocable, perpetual, non-exclusive license to reproduce the Work in copies, to prepare proprietary derivative works based upon the Work, to distribute copies of the Work to the public by sale or other transfer of ownership, and to display the Work publicly.

If the Work is software published by PRIA as codes in source and binary form, the License includes the right for Licensee to distribute copies of, and use, the codes in source and binary forms, with or without modification.

Any distribution of copies of the Work, or of a derivative work based upon the 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 works developed by PRIA and some of its contributors, subject to PRIA License, Version 1.1 November 2003 published at www.pria.us/license.htm or any subsequent applicable version of such License.” Any software application developed by Licensee based upon the Work shall include the following notice in its end user documentation and in its codes: “This software product includes software or other works developed by PRIA and some of its contributors, subject to PRIA License, Version 1.1 November 2003 published at www.pria.us/license.htm or any subsequent applicable version of such License.” Upon publication of a derivative work, Licensee shall inform PRIA of such publication and address to PRIA a copy of Licensee’s derivative work and any relevant documentation.

“PRIA” is a trade name of the “Property Records Industry Association.” No derivative work or altered versions of a Work by Licensee may be trademarked or labeled in reference to PRIA or any of its trademark(s) or service mark(s) without PRIA’s prior written approval. No reference to PRIA or any of its trademarks by Licensee shall imply endorsement of Licensee’s activities and products.

DISCLAIMER: THIS WORK IS PROVIDED “AS IS.” PRIA, THE COPYRIGHT HOLDER, THE AUTHORS OF THIS WORK AND ANY STANDARD-SETTING BODY CONTRIBUTORS TO THIS 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 WORK ARE FREE FROM ERROR OR SUITABLE FOR ANY PURPOSE; NOR THAT IMPLEMENTATION OF SUCH CONTENTS WILL NOT INFRINGE ANY THIRD-PARTY PATENTS, COPYRIGHTS, TRADEMARKS OR OTHER RIGHTS. IN NO EVENT WILL PRIA, THE COPYRIGHT HOLDER, ANY AUTHOR OF THIS WORK, OR THE STANDARD-SETTING BODY CONTRIBUTORS TO THIS WORK BE LIABLE TO ANY PARTY FOR ANY DIRECT, INDIRECT, SPECIAL OR CONSEQUENTIAL DAMAGES FOR ANY USE OF THIS WORK, INCLUDING, WITHOUT LIMITATION, ANY LOST PROFITS BUSINESS INTERRUPTION, LOSS OF PROGRAMS OR OTHER DATA ON YOUR INFORMATION HANDLING SYSTEM OR OTHERWISE, EVEN IF PRIA, THE COPYRIGHT HOLDER AND/OR ANY AUTHORS AND/OR ANY STANDARD-SETTING BODY CONTRIBUTORS TO THIS WORK ARE EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.
# Table of Contents

1 EXECUTIVE SUMMARY ................................................................................................................. 1

2 INTRODUCTION ............................................................................................................................. 5

3 BACKGROUND ............................................................................................................................... 7

4 DETERMINING THE SCOPE OF THE PROBLEM .......................................................................... 8
   - Redaction of Master or Authoritative Documents ...................................................................... 8
   - Redaction of Signatures ............................................................................................................ 9

5 REDACTION DEFINED .................................................................................................................. 10
   - Manual Redaction .................................................................................................................... 10
   - Automated Redaction ................................................................................................................ 12
   - What percentage of documents contains data to be redacted? ................................................. 14
   - How will the existing workflow be affected by automated redaction, both in actual tasks as well as the additional time to perform them? ........................................................................... 15
   - How does maintaining two versions of a record affect image storage? .................................... 15
   - What are the pitfalls of automated redaction tools which are unable to handle poor quality images, handwritten PII, etc.? .................................................................................. 15

6 WHAT IS A SOCIAL SECURITY NUMBER ..................................................................................... 17
   - Structure of a Social Security Number – Truncating SSNs ....................................................... 17
   - Social Security Number (SSN) Redaction for Notice of Federal Tax Lien .................................. 18
   - What happens when the SSN is masquerading as another number? ....................................... 18

7 KEY QUESTIONS TO IDENTIFY & ESTABLISH SCOPE .......................................................... 20

8 BALANCING PRIVACY & SECURITY WITH ACCESSIBILITY ..................................................... 22
   - Discussion Questions ................................................................................................................ 23
     - Should redacted images be reviewed by humans for accuracy? ............................................ 23
     - Who can view the original version of a redacted document? ................................................ 23
     - What is the risk to a Recorder when records are in the process of being redacted? ............ 23
What steps should a Recorder take to assure the collection of personal information through the redaction process is secured with even greater vigilance than the source documents? ........................................24

What are the practical security issues and pitfalls surrounding the practice of making both redacted and un-redacted images available online to different users based on logon permissions? ..................................................................................................................24

Which items should be included in legislation affecting the redaction of PII? ..................................................................................................................24

TECHNOLOGY: CHOOSING THE RIGHT SOLUTION ..........................................................26

Image Formats ...........................................................................................................27

OCR Redaction ........................................................................................................27

Automated Redaction Costs .....................................................................................28

Risks Associated with Manual Redaction ................................................................29

Manual vs. Automated Solution Accuracy ..................................................................30

“Redact on Demand,” “Just-In-Time” and "Request-Deferred" Redaction ..................30

Tie to Land Records Management System or Stand-Alone? ....................................30

CALCULATING ACCURACY ..........................................................................................31

Random and Targeted Sampling to Test Accuracy ..................................................32

Manual Validation ....................................................................................................33

Validation Passes ......................................................................................................33

Accuracy Computations ............................................................................................33

Accuracy on Large Repositories ................................................................................35

LEGISLATIVE MANDATES ..........................................................................................37

SOURCE DOCUMENT CONSIDERATIONS ..................................................................38

REDACTION POLICY CONSIDERATIONS: ASKING THE RIGHT QUESTIONS ..........39

CONCLUSION ...............................................................................................................42

APPENDIX A: BEST PRACTICES ..................................................................................43

APPENDIX B: GLOSSARY OF TERMS .........................................................................45

APPENDIX C: RELATED DOCUMENTS ........................................................................50
1 EXECUTIVE SUMMARY

There are continuing concerns and challenges balancing the public's access to information with an individual's right to privacy. To produce this paper, PRIA's Records Access & Privacy Policy Committee consulted and collaborated with industry partners and stakeholders. Industry partners and stakeholders include, on the one hand, Recorders nationwide, as the legislated custodians providing access to the land records, and, on the other hand, those dependent on the accuracy and availability of these records, including the American Land Title Association® (ALTA®), title insurance underwriters, title searchers, lenders, recording software vendors, and redaction software vendors. ¹

This paper addresses the availability of information about individuals contained in land records maintained by Recorders and the technical solutions of redaction. Land records by definition are public records, and in many jurisdictions, there are few restrictions on the release of public records. Accordingly, this paper sets forth redaction best practices for policymakers and Recorders. The paper is specifically intended as a resource tool for legislators before passing legislation and for Recorders as they begin to address redaction.

Redaction is the most common solution for concealing private data in the public record from public view. This solution allows sensitive, private, or confidential information which is exempt by law from disclosure to be covered in a manner that does not distort the meaning of the record. Redaction can most accurately be described as the process of removing sensitive data from the publicly viewable portion of an official government record.

PRIA's Social Security Number and Privacy Protection Act (SSNAPP Act, available on the PRIA website) provides guidance to both the Recorders and legislators, defining Personally Identifiable Information (PII) as “one or more of the following specific unique identifiers when combined with an individual's name: (1) Social Security Number, (2) driver's license number or state identification card number, (3) financial institution account number, credit, debit or charge card number, and (4) date of birth.” Some states have added additional items to this list such as mailing addresses, email addresses, veteran’s administration numbers, and medical IDs.

---

¹ In the United States, land document recording may take place at the State, City, Town, County, Borough, or Parish level. Depending on the jurisdiction, the Office of the Recorder may also be known as Recorder of Deeds, Registrar-General, Register of Deeds, Registrar of Deeds, Registrar of Titles, Deeds Registry, Auditor, or Deeds Office. In some states, the recording function is part of the county clerk's responsibilities. Throughout this paper, the term utilized for this role will simply be “Recorder.”
It is important for stakeholders to understand that when government services and available data are not contributing to the erosion of privacy or personal security. As busy legislatures grapple with redaction laws, they are wise to avoid costly and potentially ineffective blanket solutions that will not appease a worried public.

We offer a cautionary note that the redaction of property addresses will have dire consequences when securing financing and acquiring title insurance to purchase or encumber real property. Title companies cannot guarantee or insure title if they cannot confirm ownership and properly trace the chain of title for the property.

Lawmakers should also be mindful to ensure that legislation does not adversely impact constructive notice elements in their jurisdictions so as to impair the chain of title for real property. Some information is widely available online, and embarking on an expensive redaction effort may not actually result in protecting individuals’ privacy.

As a best practice, legislators should consult their Recorders, as well as PRIA and other like organizations such as the National Association of County Recorders, Election Officials, and Clerks (NACRC), the International Association of Clerks, Recorders, Election Officials, & Treasurers (IACREOT) and ALTA®. These organizations and individuals provide a broad, systemic view of the purpose of public records and would be directly affected by

Bearing in mind the daily potential for identity theft, information security is of paramount concern to government officials. Recorders understand the necessity of court-ordered protection for victims’ addresses, and other concerned groups regarding personal safety. To comply with current statutes that provide open access to public records without exposing individuals’ PII, Recorders are responding with the utilization of both automated and manual redaction tools.

As a best practice, PRIA recommends the use of automated redaction software for covering up data from digital images.

Automated redaction software enables counties to continue to allow open access to their land records documents while guarding the PII that may be included in those documents. Advanced redaction solutions can be utilized to save labor costs by eliminating a significant percentage of tedious human review, while expediting the availability of the records for public access.
PRIA disagrees with the practice of making both redacted and un-redacted images available on the Internet to different users based only upon logon permissions. There are simply too many security issues and potential pitfalls.

The next two best practices address a critical element of any redaction project: “Image Quality.” Three important terms concerning image quality are: Optical Character Recognition (OCR); Intelligent Character Recognition (ICR), most often associated with recognition of handwritten data; and Dots Per Inch (DPI), used to measure the resolution of an image. Most document scanners allow for setting the default DPI as part of the scanning process.

OCR is a critical element of redaction technologies. Good image quality reduces the number of images requiring human review and increases accuracy of the redaction technologies. DPI is not the only consideration when evaluating image quality. It has been a common practice when indexing land records to underline, circle, or add checkmarks to identify information that needed to be entered into the land records systems or simply as a check mark process to indicate the document being recorded contains the required information.

Since documents scanned at 300 DPI result in larger computer file size, and where recording offices routinely store images at less than 300 DPI, some redaction technologies have the ability to allow scanning and processing at the higher 300 DPI resolution using temporary files, while maintaining permanent storage at the lower resolution. The software automatically reduces the image file in a process called file compression to require less storage once OCR has been performed.

Although 300 DPI is the recommended best practice for day-forward scanning, if historical documents have been scanned at a lower resolution, it is not necessary or practical to rescan for redaction. An image scanned at 200 DPI meets the minimum requirement for
redaction processing. If, however, older documents have not been scanned, 300 DPI is recommended as the cost of storage continues to decrease.

In selecting a redaction process, either manual or automated, effective evaluation of the entire solution is necessary. An analysis of automated redaction processing includes both the automated software costs as well as the labor costs for manual review of the results. If a manual only process is planned, several layers of manual validation and their associated costs must be taken into consideration. Results in either redaction processing scenario are measured by the accuracy of the process and the precision or repeatability of the accuracy, measured over time. This paper defines accuracy methodology and benchmarks given a set of best practices within the industry.

As a best practice, conduct accuracy benchmarks prior to purchasing an automated redaction solution and apply the same benchmarks throughout the use of the system to ensure the redaction processes are producing expected results.

As a best practice, for evaluating accuracy in “go forward” file processing, perform a separate controlled vendor accuracy benchmark, providing the vendor with documents that have not been supplied previously.

Unless redaction technology is purchased as a full service, the Recorder’s staff will be reviewing potential redactions of the documents output by the technology vendor’s solution. The least amount of validation for the highest level of accuracy produces the lowest total cost of ownership. This aspect is often overlooked in the selection process. To prepare for an accurate evaluation, a quality control document set should be gathered, the documents of which are specific to land records and include challenging documents, such as non-form documents, handwritten data located on documents, data that requires redaction without machine-printed keywords, mislabeled data (e.g., SSM rather than SSN), and data requiring redaction without keywords or cues.
2 INTRODUCTION

The Information Age, more commonly known as the Computer Age or Digital Age, is characterized by the ability of individuals to transfer information freely and to have instant access to information that would have previously been difficult or impossible to find. The concept of a digital age or digital revolution carries the ramifications of a shift from an economy based on goods and services brought about by the industrial revolution, to an economy based on the acquisition and manipulation of information, i.e., an information society.

In this type of age, personal information, if available online, can be stolen or exploited for criminal purposes, often with devastating consequences for the victim. Government records are public records to enable citizens to monitor their government and to ensure accountability in a democratic society. Land records, for example, are public to impart constructive notice of the rights of individuals and entities in real property. The challenge for policymakers is to balance the public's right to access information with the individual's right to privacy.

This issue is a concern for governments that make public documents available online. The Federal Trade Commission (FTC) has reported that identity theft is the fastest growing crime in the United States. The FTC noted that Government document/benefits fraud was the most common reported form of identity theft, followed by credit card fraud, phone and utilities fraud, and employment fraud. The evolution of technology requires today's public officials to take a modern approach to record-keeping. While banks, title companies, political parties, and other entities use public documents for legitimate non-governmental purposes, unscrupulous users can do harm.

Every year states pass new or amend existing legislation to address the growing concern of identity theft. To date, most states have passed some legislation that addresses privacy concerns ranging from what PII can be included in public government documents to legislatively requiring the redaction of PII from the public records.

To facilitate informed decision making, this paper provides some key questions and considerations when addressing the scope of, and the solution to, PII contained in the public land records. This paper does not attempt to make specific policy recommendations on the substance of what information should be public or what should be private.

In the United States, Recorders are the custodians of the public land records and it is they who are tasked with providing public access to these documents. From adopting new legislation to increasing fees and updating record-keeping procedures, states are seeking new and innovative ways to protect citizens from identity thieves, stalkers, and other threats. Accordingly, Recorders must now address concealing PII, some of which may have been previously mandated by prior law as “public information.” PRIA has assembled a list of land records redaction laws by state for your use and information. You may access the

---

list at:  

In 2006, PRIA created and adopted the SSNAPP Act. The SSNAPP Act is model legislation designed to help guide states to create informed legislation. This model legislation is now the standard for SSN redaction nationwide.

The purpose of this paper is to educate about redaction, to identify project considerations, to identify the questions that should be asked, and to provide a source of reference. This paper is also designed to provide guidance to policymakers and administrators as they consider whether a problem exists and whether redaction is needed.
3 BACKGROUND

Florida was the first state to require redaction of documents in the land records, which are called Official Records in Florida.

In May, 2001, Florida passed a law requiring that by January 1, 2002, each Recorder must have made available on the publicly accessible Internet the complete index data for all documents recorded, at least as far back as 1990. The same law also required that, no later than January 1, 2006, images of all recorded documents at least as far back as 1990 must have been made accessible and viewable on the Internet.

Following the attacks on the United States on September 11, 2001, the Florida legislature convened in early 2002 and passed modifications to the previously enacted law. Death certificates, military discharge documents, probate documents, mental health documents, juvenile court documents, and family court documents (divorces, child support, etc.) were specifically restricted from being viewable on the publicly accessible Internet. Further, the legislature added a requirement that by January 1, 2006, all recorded documents on the publicly accessible Internet have five numbers redacted (removed from public Internet view):

1. Social Security Numbers
2. Bank Account Numbers
3. Credit Card Numbers
4. Debit Card Numbers
5. Charge Card Numbers

Extensions of time for compliance were provided by the legislature over the years as not every Recorder was able to complete redacting the images, primarily because the images first had to be digitized from microfilm or paper. Other counties chose to digitize images further back in time and those images needed to be redacted as well. Florida law first addressed only land records; it was later amended to also include court case records/documents. Subsequently, some states followed Florida’s lead; others did not.

Land records by definition are public records. In many jurisdictions, there are few restrictions on the release of public records. For example, information from public records is frequently obtained by direct marketers. Public records may also be used by private investigators, attorneys, land title professionals, credit reporting bureaus, law enforcement officials and other government agencies. In fact, as more public records are posted online, anyone with a computer and Internet access can easily compile detailed profiles on individuals.

As a best practice, custodians of land records should follow their specific state statutes regarding their duties to redact and provide access to the records.
4 DETERMINING THE SCOPE OF THE PROBLEM

The most basic question policymakers must answer in pursuing a strategy for redaction is: What is the scope of the problem?

Initially, this question requires identifying what constitutes “sensitive” information. Most states define this data in terms of “personally identifiable information” or PII. But even PII is a broad term with no universally-established legal definition.

PRIA, in its SSNAPP Act, defines PII as “one or more of the following specific unique identifiers when combined with an individual’s name: (1) Social Security Number, (2) driver's license number or state identification card number, (3) financial institution account number, credit, debit or charge card number, and (4) date of birth.” A few states have added additional items such as:

- Mailing address
- Professional License Numbers (e.g., doctors’ DEA Numbers)
- Email addresses
- Veterans Administration Numbers
- Medical IDs – Medicare/Medicaid

As a best practice, property addresses should NOT be redacted from land records.

Once policymakers have identified the types of PII that are to be redacted from the publicly accessible record, it is necessary to determine the most practical, cost effective strategy for removing that information. In light of the massive amounts of historical and current data that is available in government databases, policymakers must narrow down the documents and timeframes that will be the focus of redaction efforts. In cases where government resources are limited, this process can be particularly challenging.

It is also vitally important that stakeholders understand when government services and available data are not contributing to the erosion of privacy or personal security. As busy legislatures grapple with redaction laws, they are wise to avoid costly and ineffective blanket solutions that may not totally appease a worried public. Lawmakers should also ensure that legislation does not adversely impact constructive notice elements in their jurisdictions so as to impair the chain of title for real property. Some information is widely available online, and embarking on an expensive redaction effort may not be the best solution.

Redaction of Master or Authoritative Documents

It is generally accepted that, for land records, authoritative or master/legal documents should NOT be redacted, but should be retained as originally recorded. Redaction should be applied only to the publicly accessible (viewable and printable) copies of the documents.
Redaction of Signatures

It is generally accepted that signatures on property record documents should NOT be redacted even when redaction is applied only to the publicly accessible (viewable and printable) copies of the documents.

We offer a cautionary note that the redaction of property addresses will have dire challenges when securing financing and acquiring title insurance to purchase or encumber real property. Title companies cannot guarantee or insure title if they cannot confirm ownership and properly trace the chain of title for the property.
5 REDACTION DEFINED

As previously stated, redaction can be defined as the act of covering or otherwise removing from the public record or public view any sensitive, private, or confidential information. For hardcopy documents, this can be an overwhelming task. For each occasion a person researches and gathers information or makes or requests copies of public records, the Recorder’s office risks that PII will be released.

Consider the volume of land-related records. For every transaction where a parcel of land or real property changes ownership, a voluminous collection of documentation follows. Title information, mortgage/loan information, historical information such as liens, court information, and related documents all have some influence on the disposition of any parcel of land or real property. Varying types and amounts of PII can be found within these documents. Transaction documents such as deeds, mortgages (or deeds of trust), and assignments are sent to the Recorder’s office for recordation both to provide constructive notice and to provide long term or permanent archiving. Part of the Recorder’s responsibility is to make the documents accessible – and that may require redacting any PII present in those documents.

Today, both the public and private industry generally demand electronic information exchange. As a result, more and more recording jurisdictions are adopting electronic technology to manage their core business operations. Based on this shift, communication is becoming more and more electronic. Document sharing no longer consists of sending volumes of paper documents, but allowing access via online computer applications.

Bearing in mind the potential for identity theft, information security is of paramount concern. So how does a Recorder comply with statutes and laws to provide open access to public records without simultaneously exposing individuals’ PII, especially via the Internet? The answer is Redaction.

As a best practice, PRIA recommends the use of automated redaction tools for digital images.

Automated redaction software can help Recorders make their land records available without disclosing PII. Advanced technology can be utilized to save labor costs and eliminate a significant percentage of tedious data review and manual tasks. The following will provide descriptions of redaction methods and methods of measuring accuracies.

**Manual Redaction**

A typical manual redaction process includes:

1. Printing the document.

---

3 For electronic documents, information to be removed may be present not only on the visible image, but embedded as metadata “tags” within the document or the TIFF header. PII should be removed from the metadata tags or TIFF headers, if any is found there.
2. Finding and redacting (typically by black Magic Marker®) all of the desired areas, possibly noting the reasons for each redaction in a manual log or a note in the margin.
3. Photocopying the page(s) once to make sure that no one can read through the marked redaction.
4. Photocopying the page(s) again to make sure no one can read the redactions from the back side of the redacted page.
5. Scanning the redacted page(s) back into the system.
6. Repeating Steps 1-5 above.

When done properly, manual redaction can be effective. However, it has several drawbacks:

1. It is time consuming.
2. It is prone to human error.
3. It will likely require the document to be scanned back into the document management system.
4. If additional changes are required, one must start over.
5. It requires a significant amount of paper, ink, toner and energy.
6. It may not be completely secure.\(^4\)

![Figure 1 Example of manual redaction on a legacy hardcopy document](http://www.cse.lehigh.edu/~lopresti/Publications/2005/spie05a.pdf)

Depending on volume, source documents, and funding, manual redaction may be completely acceptable as an option. Regardless of “how” redaction is accomplished, business practices and policies for redaction should be established and documented.

When redaction is completed with a black marker, the underlying text is not actually removed. Instead, the text is hidden under a layer of ink. To the naked eye, it may appear to obscure the SSN or other piece of sensitive information, but in reality the text and the covering ink could be two slightly different variations of black. For example, the SSN appearing on the document could be typed in a particular shade of black (for simplicity let’s call it black #2), and the redaction marker could be another shade of black (let’s call it black #4). While very close in color, there is a slight difference. A computer scanner may
recognize the difference in the darkness of the two black colors and produce an electronic version with the two shades of black intact. An individual is then able to electronically remove the one shade of black from the image to reveal the SSN or other piece of sensitive information.

**Automated Redaction**

There have been significant technology advances since the year 2000 which offer an automated option to the redaction process. In lieu of using a black marker on paper documents, many Recorders are now using OCR to process scanned images of the paper documents. OCR technology converts scanned images into a digital format which then allows rules-based search engines to locate sensitive information contained within the document. The search engine uses a combination of words, phrases, patterns of text, proximity, and location to identify potentially sensitive information. For example, the engine may find the clue word “SSN” within a document followed by a pattern of numbers such as xxx-xx-xxxx. The combination of the clue word with the format of text provides high confidence that the potentially sensitive information needs to be redacted.

It should be noted that automated redaction processes do not work on paper, microfilm, microfiche, or photostat documents without first converting these formats to scanned/digitized images.

A typical automated redaction process includes:

1. Digitize the document.
2. Automated software processes the document.
5. Production of redacted image.
6. Storage of redacted image back into repository.
Some automated redaction applications perform redaction by creating overlaid blackout regions (annotations) on a new layer in a TIFF or PDF document, while the underlying PII continues to be accessible.

a. The advantage of choosing to annotate images is the Recorder does not have to store a second copy of the image. Instead the “x” and “y” coordinates
where the sensitive information is located are stored and applied based on permission requirements.

b. The disadvantage of annotations is the redaction zone is not secure, and hackers are able to quickly and easily remove annotations to access the PII.

c. Another disadvantage of annotating an image as opposed to burning the redaction into the image is, if the image is later manipulated, the annotation could appear in the wrong position on the image.

d. If choosing to apply redaction zones as annotations to reduce storage requirements, it is recommended that the Recorder also have the ability to burn the redaction zone into the image when the image is to be viewed by unauthorized personnel.

e. If the choice has been made to burn the redaction into the documents rather than utilize an overlay or layer, storage requirements will increase based on the percentage of redacted documents in the underlying image library. Only the documents containing redactions will create the need for an increase in existing storage requirements. For example if 6% of the documents in a given library contain redactions, 6% more storage would be required to maintain the redacted copy.

**What percentage of documents contains data to be redacted?**

The percentage of PII contained in land records requiring redaction is uncertain, although it can be expected to be relatively low. A survey of three PRIA members revealed the following statistics/details:

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Documents Recorded</th>
<th>Images Recorded</th>
<th>Documents Redacted</th>
<th>% Documents Redacted</th>
<th>% Images Redacted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orange County FL</td>
<td>12,901,821</td>
<td>33,264,327</td>
<td>847,611</td>
<td>6.57%</td>
<td>2.55%</td>
</tr>
<tr>
<td>Notes (Orange)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Washoe County NV</td>
<td>1,526,627</td>
<td>6,343,931</td>
<td>14,018</td>
<td>0.09%</td>
<td>0.02%</td>
</tr>
<tr>
<td>Montgomery County PA</td>
<td>601,950</td>
<td>2,640,780</td>
<td>1,160</td>
<td>0.02%</td>
<td>0.04%</td>
</tr>
<tr>
<td>Notes (Washoe and Montgomery)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Redacting 5 numbers: social security numbers, plus bank account, credit, charge and debit numbers
- Date Range: 01/01/1970 through 07/31/2007
- Redacting social security numbers
- Date Range: 01/01/2004 through 12/31/2011
How will the existing workflow be affected by automated redaction, both in actual tasks as well as the additional time to perform them?

Existing workflow is beneficially affected by automated redaction, including:

- Automatically finding and removing PII in land records
- Making data protection a seamless part of everyday workflows, not a separate process
- Meeting legislative requirements for privacy and disclosure
- Protecting PII from unintentional disclosure, theft, or other misuse
- Replacing tedious, manual redaction with highly efficient automated processes

How does maintaining two versions of a record affect image storage?

Storing redacted images may not take up significant space. Land Records Management Systems (LRMS) are able to implement storage of a redacted image by storing the image that contains the redacted field(s) as an additional image associated with the previously stored un-redacted document. Then, when called upon to display the stored document, the LRMS “assembles” the images comprising the document, such that redacted images take the place of un-redacted images within the document. Depending upon the privilege level of the user requesting the display of the document, the LRMS may display the un-redacted document without substituting redacted image(s). In the case of the chart above for Orange County, FL, only 6.57% of the documents contained redactions so the Recorder only had to increase storage capacity for the back file minimally.

What are the pitfalls of automated redaction tools which are unable to handle poor quality images, handwritten PII, etc.?

The primary pitfall of automated redaction/OCR solutions is that they are unable to effectively process poor quality images and handwritten information. This inability to accurately capture information results in an increased amount of manual review and an associated decrease in accuracy. While the cost of a more sophisticated redaction technology solution is greater upfront, it can save Recorders a significant amount of money in the long run by decreasing the number of images the staff needs to manually review. The most basic desktop redaction tools require 100 percent manual verification of images compared to the most sophisticated solutions which typically require less than 10 percent of images to be manually verified. It should be noted that there are many solutions in between these two extremes.

OCR output is critically important to any automated redaction process. Better quality images result in a higher quality OCR output file, and thus improve the success of the redaction process. A basic redaction software package may lack image cleanup utilities to handle poorer quality images. These utilities are usually found in customized redaction/OCR solutions.
Some of the image cleanup tools that can be deployed with customized OCR solutions include:

- Character thickening (dilation) for faint characters
- Character thinning (erosion) for blurred characters
- Character smoothing (opening) for fuzzy and dot-matrix characters
- Removal of:
  - "Salt and pepper" noise
  - Speckles and spots (single pixels)
  - Ink blobs and blotches (small pixel clusters)
  - Page borders
  - Punch holes
  - Blank pages

Some of the OCR solutions also perform fuzzy text searches to match all possible approximate results for a word, name, or number pattern, despite spelling mistakes and number transpositions.

Basic redaction software solutions typically do not come with a handwriting recognition engine or ICR. That service is normally an additional cost. Since ICR is processing intensive, it can significantly slow the redaction processing time if it is applied to every image page. A customized redaction/OCR solution can deploy ICR selectively, applying it only as needed, and thus minimizing the impact on the overall redaction process.

Many redaction/OCR solutions may be sufficient for Recorders that have the resources available to review all the images. Recorders should consider all internal verification costs and the overall accuracy rate of the software when making a decision.
6 WHAT IS A SOCIAL SECURITY NUMBER

In the United States, a Social Security number (SSN) is a unique nine-digit number issued to U.S. citizens, permanent residents, and temporary (working) residents under section 205(c)(2) of the Social Security Act, codified at 42 U.S.C. § 405(c)(2). The number is issued to an individual by the Social Security Administration (SSA), an independent agency of the United States government.

The original purpose of the SSN was to track individuals' accounts within the Social Security program. The SSN has since come to be used as a unique identifier for individuals within the United States, although in rare instances errors occur and duplicate numbers exist. Employee, patient, student, and credit records are sometimes indexed by Social Security number.

Structure of a Social Security Number – Truncating SSNs

The Social Security number is a nine-digit number originally designed in the format "AAA-GG-SSSS." The number is divided into three parts.

- The Area Number (AAA), the first three digits, is assigned by geographical region. The Area Number may represent the area where the applicant lived, where the card was issued, or a zip code. However, it does not necessarily represent the state of residence of the applicant.
- The middle two digits are the Group Number (GG). The Group Numbers range from 01 to 99. Group Numbers are not assigned in consecutive order, but in a predetermined order. The SSA periodically published a list of High Group Numbers specifying the highest Group Number that had been validly issued for any given Area Number.
- The last four digits are Serial Numbers (SSSS). They represent a straight numerical sequence of digits from 0001-9999 for any given Group Number.

On June 25, 2011, the SSA changed the SSN assignment process to "SSN randomization." SSN randomization affects the SSN assignment process in the following ways:

- It eliminates the geographical significance of the first three digits of the SSN, currently referred to as the Area Number, by no longer allocating the Area Numbers for assignment in specific states.
- It eliminates the significance of the highest Group Number and, as a result, the High Group Number List will be frozen in time and can only be used for validation of SSNs issued prior to the randomization implementation date.
- Previously unassigned Area Numbers will be introduced for assignment excluding Area Numbers 000, 666, and those in the 900-999 range.

---

5 The SSA formula alteration was a result of two Carnegie Mellon researchers being able to predict an individual’s SSN from other publicly available data using the prior number assignment method- www.ssnstudy.org.
**Social Security Number (SSN) Redaction for Notice of Federal Tax Lien**

Social Security Numbers have been used for many years on the publicly recorded Notice of Federal Tax Lien. To protect taxpayer PII, the Internal Revenue Service (IRS) issued a notice to state and local Recorders in December 2005 regarding its new practice of partially redacting SSN information. Effective January 2, 2006, new Notices of Federal Tax Lien included redacted SSN information in the format XXX-XX-NNNN, where only the last four digits of the taxpayer’s SSN appear. The IRS stated that the last four digits could not be changed (or redacted) by the Recorder. Significant discussions followed regarding pre-existing state legislation and mandates requiring the redaction of SSNs in full.

Federal tax lien documents, as presented by the Internal Revenue Service, must be readily available to the public. Recorders must maintain, in some form, lien documents with partially redacted SSNs. Maintaining these documents ensures the public is obtaining information on the correct person and eliminates the need for Recorders to perform additional tasks related to lien documents as provided to their offices.

The following disclaimer should appear on any website or other system containing lien documents where the information does not exactly correspond with the IRS document.

> This website contains information on Notices of Federal Tax Lien and other lien documents that are not identical in content to documents presented to this office by the IRS. Therefore, the results of a search of this indexing system cannot be relied upon in judicial or administrative proceedings related to title or transfer of property, including but not limited to searches relative to sales of property and foreclosure proceedings, in determining ownership interests in property on which a Notice of Federal Tax Lien attaches.

> For assistance in obtaining a copy of the required document as filed by the IRS, please contact (name) of this office at (phone number). Copies of recorded documents may be obtained only from this office for which the search was conducted.

For those offices that have fully redacted the SSN and the electronic website is the “official site”, every effort must be made to ensure that the public is able to obtain an official version of the recorded Notice of Federal Tax Lien.6

**What happens when the SSN is masquerading as another number?**

Social Security numbers were originally developed as unique identifiers, well before the advent of sophisticated methods of using and combining data files became available. Although the main legislated uses of SSNs are by the SSA and the IRS, SSNs have become ubiquitous as identifiers in both credit files and even health-related files. Social Security Numbers have masqueraded as other numbers, for example, when many states utilized SSNs as Driver’s License numbers. This practice has ceased due to privacy concerns and

---

6 Internal Revenue Service, Publication 1468.
implementation of the Real ID Law.  
https://www.dhs.gov/xlibrary/assets/privacy/privacy_pia_realidfr.pdf

The Social Security number has also traditionally been a part of the Medicare Health Insurance Claim number, although inconsistently so. Surviving spouses of Medicare beneficiaries who do not themselves have Medicare coverage, for example, may have their late spouse’s SSN as their Medicare Health Insurance Claim number. The use of the SSN in the Medicare context is tied to reimbursement. However, inaccuracies in the recording and reporting of the claim number constitute the number one reason that Medicare claims are returned to the provider, suggesting that another, perhaps less cumbersome, numbering system might be more appropriate. Virtually all algorithmic identification systems make some use of Social Security numbers, and none work as well without the SSN.

In California, the state Franchise Tax Board uses a SSN with the addition of numbers and/or letters to turn a SSN into an account number.
7  KEY QUESTIONS TO IDENTIFY & ESTABLISH SCOPE

The following questions will help policymakers identify if there is a problem, as well as the scope of the problem where one exists:

Identity Security: Is the available information subjecting citizens to an increased risk of victimization from identity theft?

Policymakers should consider:
- Whether threats are real or perceived.  
- Whether the personal information that is available is being directly or indirectly obtained and used by identity thieves, and if so, how frequently/ commonly.
- Whether removing the information from government records sufficiently remedies the risk or threat to privacy or personal security.

Privacy: Is the available information subjecting citizens to unnecessary loss of privacy?

Policymakers should consider:
- Whether the information includes inadvertently collected data that alone, or in combination with other data, is posing an increased risk to citizens.
- Whether any deleterious effects on privacy are understood and outweighed by the benefits of collecting and making the information publicly available.

Personal Information: If there is a problem, what pieces of PII are causing the problem?

Policymakers should consider:
- Social Security Numbers (SSNs)
- Driver’s License Numbers
- Date of Birth
- Bank or Financial Account Information
- E-mail addresses (if used as a form of identification)
- Medical ID numbers
- Mother’s maiden name

Magnitude: How many and what types of records contain the information?

Policymakers should consider:
- Are the documents digital, paper-based, microform-based, or a combination of the three types?

---

7 Note that in Ostergren v. McDonnell, No. 3:08-cv-00362 (E.D. Va. April 13, 2011), Betty Ostergren, a private citizen, is permitted to specifically post SSNs she locates on the Internet, but only those of “public citizens” (e.g., legislators, executive officers, and Recorders) in her attempt to reform the practice of governmental offices making PII available on the Internet.

8 E.g., Wash. Rev Code § 65.04.045(3)(c), which restricts the submission of documents containing the maiden name of a person’s parent so as to be identified with a particular person.
At what government level (state/county/city) are such documents located? Most often land records are located at the county level; Hawaii and Alaska keep the records at the state level; Rhode Island, Vermont and Connecticut keep the records at the town level.

How is the data available (e.g., single record download versus bulk download)?

How many images/documents containing this information are in the Recorder’s archives? In what media are the documents (digital, microform, paper)?

How many images/documents containing this information are processed in a week? A month? A year?

How many types of documents contain information to be redacted?

What is the date range of the data that is in the Recorder’s possession?

Scope: Is the problematic information already “out there” and publicly available in areas beyond government control?

Policymakers should consider:

- How available from other sources (e.g. the Internet) is the data you are trying to protect?
- How much data has already been provided or sold to entities that re-sell the data for legitimate business purposes, and what impact does redaction have on this data?
- What are the short and long-term impacts on government services?
- What other government agencies control and make available records containing the same personal information?
- What other non-governmental, publicly accessible records contain the same information?
- What, if anything, can governments do to coordinate efforts with other policymakers and custodians who control access to such records?
- Will removing the personal information from government documents conflict with provisions of open government (sunshine) laws?

Call to action: In the case where regulations do not exist to mandate the redaction of PII, policymakers should consider:

- Whether and what information can be redacted to provide protections for citizens.
- What process can citizens utilize to search for publicly available PII on their records?
- How can citizens easily make requests for removal of that PII on their records?
- Recorders may want to follow the best practices of Recorders in other states where redaction is required when their state has not mandated redaction and the Recorder is willing to redact upon a resident’s request.
8 BALANCING PRIVACY & SECURITY WITH ACCESSIBILITY

The central problem facing many governments is the paradox created by the rise of the Internet: while making information easier and more accessible to the public, the increased access to public records presents a whole new set of challenges for citizens and government officials alike. The ideal goal is to create policies and practices that protect the public interest in privacy and security without compromising the spirit and letter of open government laws.

Today's searchable databases on government websites offer anyone with a computer and Internet connection the ability to instantaneously search for and download in minutes what formerly could take hours or days. Seasoned government recording staff like to point out that, before the advent of the Internet, those public records languished in "practical obscurity," requiring those who wanted to view them to drive to the Recorder's office and sift through endless stacks of files, books, and rolls of microfilm – often in dimly lit, damp basements. In addition to being time-consuming, costly (photocopying fees were a major expense), and limited to the rigid business hours of the government office, it was enormously challenging to aggregate public records data from different agencies without great effort and expense. Web access to public documents sweeps away these traditional barriers.

Privacy has emerged as a hot-button issue in many recording jurisdictions where the perception exists that documents with personal information are posted online. But the Internet has made the work of government record custodians less labor intensive, providing greater ease in collecting and maintaining information. Additionally, copying and sharing public records in electronic format is much easier and less expensive. However, the cost of transferring legacy documents to an electronic format can be considerable and the hardware for making it accessible can also be costly. There are other issues that can be troubling. For example, the majority of Recorders do not typically track who is accessing the information online. Other issues arise regarding adequately protecting websites from hackers.

Given the growing number of identity theft cases that are reported each year, millions of Americans are understandably concerned about their vulnerability to identity theft. In a 2007 Zogby poll, a majority of Americans said they worry about identity theft. In fact, 91% said they worry about their identity being stolen.

When it comes to public records, government officials are rising to the challenge of protecting the privacy and security of citizens (e.g., keeping SSNs out of insurance and student records). Government entities must be able to distinguish when the need to act in favor of the public interest for protecting privacy outweighs the need for providing access and transparency in government processes. Ultimately, however, some services and transactions require the exchange of personal information. Legitimate commercial users, such as title companies and credit reporting agencies, have for decades born the expense of acquiring and aggregating public land records to support the efficiency of services provided to the community. Policymakers will need to wade into this gray area to decide
what limits, if any, should be put on those exchanges, and at what cost to services or transparency.

**Discussion Questions**

**Should redacted images be reviewed by humans for accuracy?**

Redaction software can be very accurate in catching PII. However, many factors affect accuracy. Human review can increase the accuracy of the solution, but humans are not 100% accurate. Each additional pass of human verification will lead to increased accuracy, but at some point there is a diminished return on investment.

It has been argued in some states that passing documents through a software program that is specifically designed to identify and redact PII is sufficient – also known as “best effort” redaction.

**Who can view the original version of a redacted document?**

Security protocols are set up in many of the Land Records Management Systems in use today. The Recorder generally determines by user login who can access original (i.e. un-redacted) records. In some recording jurisdictions, only a court order will permit access to the original record; in other jurisdictions, certain categories of people (such as law enforcement officers and judges) can access the original record.9

Many Recorders allow redactions to be applied by the Recorder’s staff, but the same staff often do not have the security level which allows them to see the original version once redaction has been applied. Only a supervisor (often only the elected official and chief deputy) can view the data which has been redacted.

The same restrictions also apply to printing. Recording staff can print with redactions, but only supervisors can print un-redacted copies. In some jurisdictions, authorization of the courts or county District Attorney (DA) is required to provide un-redacted data to any requesting party. For example, if the courts make the request, some jurisdictions still require the DA to approve providing un-redacted information.

Different recording jurisdictions set different rules related to the provision of copies to the recording party. In other words, if a document is recorded and PII information is then redacted, most Recorders still will NOT provide an un-redacted copy unless the steps above are followed.

**What is the risk to a Recorder when records are in the process of being redacted?**

Risk is minimal provided government officials have trusted relationship with their vendors and vendor employees are properly screened. Minimizing risk can be accomplished

---

9 Recorders should comply with the requirements of their state laws. If state law indicates only redacted copies are to be provided, that mandate must be followed.
through the request for proposal process and contracting terms. Establishment of a trusted relationship between government employees of the land records office and the third-party redaction vendor minimizes risk.

**What steps should a Recorder take to assure the collection of personal information through the redaction process is secured with even greater vigilance than the source documents?**

The redaction process itself may create a collection of personal information which must be secured. If using a third-party redaction vendor, utilization of encryption methods when transferring data between Recorders and their redaction vendor is recommended. Any abstracted files containing OCR/redacted data stored on servers at the government location, the vendor's site, or any other off-site location should also be encrypted. In this way, encrypted abstraction files containing the grid coordinates of the redaction and PII data are rendered useless in the event of a security breach.

**What are the practical security issues and pitfalls surrounding the practice of making both redacted and un-redacted images available online to different users based on logon permissions?**

*As a best practice, only allow access to the redacted version of a document.*

**Which items should be included in legislation affecting the redaction of PII?**

The following items should be included when legislation is being considered:

**Definition of PII** – what comprises "Personally Identifiable Information" can be subject to interpretation. PRIA's Records Access & Privacy Policy Committee has focused on four primary unique identifiers which, when combined with an individual's name, would be considered PII:

- Social Security Number.
- Driver's License number or state identification card number.
- Financial institution account number, credit, debit or charge card number.
- Date of Birth.

**Inclusion of PII** – Effective legislation includes provision(s) to discourage document preparers from including PII in a document which is to be presented for recording. Legislation could also give the Recorder authority to reject the document if PII is present, but not penalize the Recorder if the document is not rejected. Further, effective legislation will hold document preparers (not submitters) accountable for including PII within documents to be recorded (see PRIA's SSNAPP Act).
Liability of Recorder – Effective legislation will include a provision indicating the Recorder is not liable for any claims arising from the inclusion of PII in a document presented for recording.

Applicability – Effective legislation will include an applicability provision detailing which documents the law does not apply to (e.g., state or federal tax liens, certified copies of death certificates or other documents required by law to contain PII that are filed or recorded in the office of the Recorder).

Funding – Effective legislation will include a mechanism for the collection of funds to facilitate the implementation of redaction. Budget considerations may be necessary for day-forward and back-file redaction projects. “Day-forward” redaction means the redaction of PII from land records that are added to the database of a Recorder after the date redaction legislation was passed or redaction was implemented. “Backfile” redaction means the redaction of PII from land records that existed in the database of a Recorder on and before the date redaction legislation was passed or redaction was implemented. When considering a backfile redaction project, a Recorder must determine how far back in history to go.
9 TECHNOLOGY: CHOOSING THE RIGHT SOLUTION

Commercial redaction providers have made revolutionary advances in the technologies they are bringing to market. Early software solutions relied exclusively on redacting specific zones of a document using a grid-like system once the correct form had been identified. The latest, cutting-edge software now automatically locates and analyzes graphic images to decipher patterns, letters, words, numbers and more, providing for completely unstructured data recognition.

Current software solutions also allow redaction of specific words and numerical patterns (e.g., a string of nine digits, such as a Social Security number); and still others can target handwritten information and special fonts. Capabilities to look for in dedicated redaction software include:

- The ability to work with multiple image file formats, such as PDF and TIFF.
- Text-based searching for specific words or word strings to be marked for redaction.
- Text pattern recognition of predefined text patterns to be redacted, such as credit card numbers.
- The ability to create non-standard pattern searches.
- Zone-based redaction to designate specified areas for redaction that are consistent from document to document.

As pointed out previously, there are two broad types of redaction methods available:

- Manual Redaction: PII is manually redacted by creating a copy of the document and physically crossing out or covering the information (e.g., with a black marker or redaction tape). This method is often combined with a second human review in an attempt to increase accuracy, referred to as a “manual double pass.” A second human review does not necessarily find data missed by the first human review. The concept of accuracy improving with a second human review in redaction is not the same as double blind key data entry, since two different people are not taking specific image data, keying it twice, and comparing for inconsistencies. Rather, two different sets of human eyes are looking at a single image and are attempting to find PII within that image.

- Automated Redaction: Automated redaction procedures rely upon software technology that uses OCR capabilities to convert scanned images into a machine-readable format. One of the many benefits of this approach is that sensitive information can be located within the document automatically using a search engine. ICR software, which is able to recognize a wider range of data types and styles than OCR, allows for more advanced techniques, including feature analysis and artificial intelligence. It is not uncommon for automated redaction tools to include combined OCR/ICR capabilities. Automated redaction may or may not involve a second human-based review. It may include a human-based review only
in specific circumstances that may be triggered by the redaction software. Note that with automated redaction, there is usually a manual review stage for documents detected by the software as problematic and benefiting from a human-based review. The percentage of documents requiring human-based review depends on several variables which include image quality, the type of the documents, the fields to be redacted, and the selected processing technology.

**Image Formats**

When a paper document is scanned, the resulting electronic document may be created in one of many different formats. There are a wide variety of image formats and each type presents its own unique set of characteristics. Some industry standard image file formats and file compression formats include:

- BMP3.LZW (Bitmap image with Lempel-Ziv-Welch compression)
- GIF (Graphics Interchange Format)
- JBIG / JBIG2 (Joint Bi-Level Image Experts Group)
- JPEG / JPEG2000 (Joint Photographic Experts Group)
- PDF, PDF-A, PDF-B (Portable Document Format)
- PNG (Portable Network Graphics)
- TIFF, TIFF Group 4 Single Bit, TIFF Multi-bit (Tagged Image File Format)

The TIFF format is the most often used format for land records documents. This almost uniform format came about as Recorders started scanning and saving the digitized land records documents in the late 1990s or early 2000s.

**OCR Redaction**

Redaction technologies are based on the ability of an OCR system to accurately convert printed text into a coded language that a computer can understand.

Factors that influence the accuracy of the recognition process are type size and font, character sharpness (acuity), and contrast between the characters and the background on which they are written. Highly stylized text or extraneous lines interfering with written characters will likely degrade recognition accuracy.

Another significant factor in OCR accuracy is capture resolution. The number of pixel elements captured by the scanner along the length and width of a page determines the output resolution. Resolution is measured in terms of “dots (pixels) per inch” or “dpi.” Typically, recorded documents are scanned at a resolution of 200 by 200 or 300 by 300 dots per inch. For simplicity, these resolutions are commonly stated as 200 dpi and 300 dpi respectively.

---

10 Note that with modern day computer systems, the automated OCR/ICR process takes approximately 2-3 seconds per image while the search for PII can be between 5 and 15 additional seconds per image. The time consumed varies, based upon the information being located as well as the quality of the image being processed. For large volumes of documents, this process can be very computer-processor intensive.
In the context of OCR, 300 dpi is recommended because most OCR software is optimized to manage the stream of data that this resolution produces. As such, it represents the optimal balance between scanner throughput, OCR processing time, file size, and character accuracy. This recommendation should not be read as implying that perfect recognition will result from any page scanned at 300 dpi. As previously stated, the appearance and size of the text being captured plays a significant role in output accuracy as do the other capture settings on the scanner.

Automated redaction tools are particularly helpful to government offices dealing with backfile conversions, especially considering the variety of document types and layouts in legacy records. Rules-based OCR and ICR redaction software makes it possible to redact information from ever larger volumes of images with increasing speed and accuracy. Automated redaction can also help government entities perform time-of-filing redaction to paper-based or online filings with or without human interaction.

Documents scanned from microfilm can be particularly difficult to OCR because character acuity is degraded and contrast variations are more likely. Expect a lower accuracy rate from these documents even when scanned at 300 dpi. If the film is scanned at less than 300 dpi, expect the accuracy to be significantly lower than source document scanning at the same resolution.

In summary, the OCR process should be viewed holistically where the appearance of the text, the quality of the source media, and the setup of the scanner all play a role in producing an outcome that will meet the requirements of the application being used.

As a best practice, when OCR is a part of the process, 300 DPI TIFF Group 4 is the desired file format for best character recognition accuracy.

**Automated Redaction Costs**

In addition to time and cost, the requirement to redact PII may leave Recorders open to unreasonable liabilities, not to mention unrealistic outcomes. It is of vital importance to provide clear and detailed terms on this issue, among others, in preparing Requests for Proposals (RFPs) and procurement documents. In trying to assess the true cost of redaction, it is important to look beyond the numbers associated with procuring hardware or software. Accuracy rates differ among redaction technologies, and some software solutions cost more than others. The increased accuracy of automated software results in less manual review time and a reduction in labor costs, so the total cost of ownership for a given solution should be calculated to achieve the true costs to the Recorder.\(^{11}\)

\(^{11}\) A trained operator performing “heads down” redaction validation can review a single page image for SSNs in approximately 3 seconds using special purpose redaction validation software. Adding additional fields for validation, or having to correct for software mistakes, lengthens this time.
There are personnel costs involved with both manual and automated redaction approaches, so policymakers will want to take this into account. Furthermore, the state of the records (i.e., digital or paper-based) and the condition of the material that is in need of redaction must be examined. It is particularly necessary to ensure that the redaction technology being considered will work with legacy documents, especially if they have become weathered or damaged in storage, or converted from low resolution microfilm. Some redaction technologies include built-in image enhancement technology to handle such situations. Some public records are easier and less costly to work with than others. As a final thought on calculating costs, it is important to properly allocate resources to maintain quality customer service levels while undertaking a redaction project.

Ultimately, choosing the right technology does not simply depend on the technology that is the most effective. Officials must ask themselves a variety of questions before choosing an approach to redaction. They must determine what needs to be redacted, whether to locate and save other data items that might be redacted in the future, what document types are affected, what resources are available to implement redaction (e.g., budget, staffing levels), and how best to utilize those resources. One recording jurisdiction’s solution may require hiring additional staff to manually redact information, while another jurisdiction’s approach may involve automated redaction procedures with additional expenses for implementing and updating software.

When it comes to redaction, policymakers must balance the resources and the goals as part of the decision-making process. With a variety of technologies available (with differing levels of cost and sophistication), today’s policymakers must educate themselves about their options and how they fit with their needs. Policymakers should give close consideration to the outcomes they are expecting, and make those details very clear when issuing contracts for redaction projects. Budgets and staffing often play a large role in the technical approach that is ultimately adopted. The most effective method of redaction is not the type of technology, but rather the method that best suits the needs and resources of the Recorder.

**Risks Associated with Manual Redaction**

Manual redaction is traditionally labor intensive and costly, and can strain staff resources. The cost of redacting can easily run into thousands of dollars (or more) and the results may still be rife with human errors. Simply put: There are varying degrees of redaction capabilities and many risks associated with using basic redaction techniques for electronic documents. Merely placing a black “censor” bar on top of the text as described in Section 5 of this document may be subject to the discovery that the bar can inadvertently shift and reveal the text or that information thought to have been redacted can be “retrieved” by certain software products since the underlying text has not been removed from the document. The consequences of bad redaction are huge – akin to the consequences of misreading a prescription in healthcare.
Manual vs. Automated Solution Accuracy

The primary risk associated with any type of redaction is that sensitive data items could be missed. People make mistakes and it is impossible to expect them to visually identify all data items with 100% accuracy. Based on feedback from several counties currently redacting PII, there is roughly a 15% margin for error in a manual process.

Redaction legislation in most states is written to protect the Recorders as long as they make a good faith effort to redact, but it is a risk nonetheless. One option is to go through a second, manual review. However that becomes an expensive solution. A random second review of 1% of the image set may be more cost effective than completing a manual review of all documents.

A primary benefit of automated redaction solutions is the ability to eliminate large percentages of documents that do not contain PII from the manual review process while achieving high accuracy rates and saving the Recorder both time and money. Accuracy rates of automated solutions vary based on technology, image quality, type of documents, and fields to be redacted, regardless of whether the rate is calculated using document-, page-, or field-level accuracy.

“Redact on Demand,” “Just-In-Time” and “Request-Deferred” Redaction

There is an alternative if budgetary constraints prohibit redacting all historical images. There are options for “just-in-time” and “request-deferred” redaction. This is also known as “on-demand” redaction. On-demand redaction can best be described as 1) an individual requests a document; 2) it is retrieved from the repository; 3) the document is run through the redaction process; and, 4) the redacted document is presented to the requester.

While budget may be a consideration for implementing new technologies, Recorders must also consider:

- How many online requests are received for documents each year?
- How quickly is the Recorder required to provide the redacted image?
- Will the requests be processed in an automated environment or is it necessary to go through a human review?
- Is it necessary to OCR/ICR the documents for reasons other than redaction, such as auto indexing or conversions to searchable PDFs? Would those processes be negatively impacted by an on-demand redaction workflow?

Tie to Land Records Management System or Stand-Alone?

Most redaction solutions can be integrated with a recording jurisdiction’s LRMS, used in a stand-alone environment, or operated as part of an enterprise solution of which land records is only one piece.
10 CALCULATING ACCURACY

In selecting a redaction process, you are effectively evaluating a solution based upon an expectation of results and the costs for achieving those results. Assessing the accuracy of automated redaction processing must include both the automated software costs, as well as manual review labor. If a manual only redaction process is elected, several layers of manual validation and their associated costs must be evaluated. The results of either redaction processing scenario are, of course, measured by the accuracy of the process or the technology elected, and the precision or repeatability of the accuracy, as measured over time. This section of the paper offers guidance for consideration in quantifying accuracy results using specific methodologies and benchmarks which are best practices within the industry. Using this information, the Recorder should be able to evaluate the various technology vendors and their solutions.

Accuracy benchmarks may even be performed for manual redaction processes, allowing one to determine the effective human error rate of performing a repetitive task that demands visual acuity skills and little to no distractions. For the purposes of this paper, automated redaction technology accuracy will guide our presentation of information.

When determining redaction accuracy methodology and establishing benchmarks, it is important that accuracy be measured across all documents, not just the subset of those identified by automated software for a manual review stage. Evaluation of random sample documents where no redactions were found is also warranted. Most vendors of automated redaction systems offer options for random samplings against a set of document images by percentages and document types. For example, a sample set might be 1% of all the images selected at random plus all lien documents. In calculating accuracy, do not just count the misses on documents suggested for review by an automated system.

It is also important that both under-redactions (i.e., misses) as well as over-redactions (i.e., excessive data redacted) are counted in the accuracy benchmarks. Although there may be less risk in over-redactions, over-redactions still require human resources to correct, and thus will increase the cost of the selected solution.

As a best practice, accuracy methodologies and benchmarks should be considered when using any automated redaction technology solution, and applied throughout the use of the system to guarantee that the redaction processes are producing expected results with known precision.

Following are details on various methodologies, the different measures of accuracy, and how each may be calculated accurately for large existing repositories, as well as the lower volume of daily "go forward" file processing.
Random and Targeted Sampling to Test Accuracy

A random sampling may not be statistically significant if less than 1% of the overall repository is used, depending on the size of the repository. There are multiple definitions and directions on the Internet about how to create a statistically significant random sampling. Also, it is true that the smaller the sample set, the less accurate the results. To conduct accuracy testing that is both measurable and precise for backfile redaction, the recommended practice is to create a statistically correct systematic sampling of documents from the existing repository. A sampling of documents can be created by random document selection using all document types across all the years in the repository. Recorders tend to know their repositories and should be able to target document types that are more likely to have PII. Adding more documents of these types to the sample helps provide a better analysis of accuracy.

This selection process yields a manageable subset of documents, or a mini-version of the repository, that is used for accuracy certification testing. Document sample sets can range from several hundred to hundreds of thousands of documents, with the actual quantity being based upon the size of the repository, the desired processing accuracy, and an assumed margin of error. Recorders must also take into account time and staff resources when establishing sampling sets. Some vendors provide software that performs random selections automatically or, if not, the vendor will usually assist in the process of preparing such sample sets.

The creation of a random and targeted initial sample set that will be provided to a vendor to establish a processing benchmark should be considered as part of the redaction vendor selection process. The initial sample set should be of sufficient size and should also be returned in a short time to ensure that the redaction vendor used its automated software. A large sample size and quick turnaround will eliminate the possibility of the use of enhanced validation that will not be possible in a true production setting.

In evaluating accuracy for “go forward” file processing, the best practice is to perform a separate controlled vendor accuracy benchmark. This benchmark is established by providing the vendor with documents the vendor has not been given previously, and then having the vendor process the documents in real time to demonstrate their process for “go forward” redaction. Unless redaction technology as a full service is purchased, it is the Recorder’s office that will be doing the manual validation of the documents output by the technology vendor’s solution. In this case, the least amount of validation for the highest level of accuracy that produces the lowest total cost is often overlooked in the selection process.

To prepare for as accurate an evaluation as possible, the document sample set should be specific to land records and include challenging documents, such as non-form-like documents, handwritten comments, marginal notations, data that requires redaction without machine-printed keywords, mislabeled data (e.g., SSM rather than SSN), and data requiring redaction without keywords or cues.

Statistical sample size calculations are outside the scope of this document; however many reference texts and Internet calculators exist.
Once sample set(s) have been selected, you are ready to proceed with processing your documents and scoring your results.

**Manual Validation**

To score the results after automated processing has been completed, a formal calculation of accuracy requires that each image be meticulously reviewed, modified as required, and certified as correct. This process can be handled either as a single-pass or a dual-pass review, with the second pass being used to certify the correctness of each document. During this process, as much time as may be required should be taken to review the image.

**Validation Passes**

Humans will make mistakes, whether in a single- or dual-pass review, or even if additional reviews beyond two are implemented. At some point, realization sets in that perfect validation, in all but highly controlled situations, is a unattainable. The job of manual inspection is to be as near perfect as possible in a production environment. Manual redaction validation is the ability to scan an image rapidly and validate that all data was properly redacted. Despite having the best human manually redact a document, redaction software will most likely find data that should have been redacted but that were missed in the only-by-human review process.

**Accuracy Computations**

After manual validation of the document sample sets from the entire repository, accuracy computations can be performed.

Accuracy may be computed on a document, page, or PII element/field basis. This section provides information on how to compute each. The differences among these accuracies are presented after explaining the individual calculation methods.

**First example:** one document, 4 pages long, with 5 PII fields requiring redaction, and all the redactions are on one page.

**Document Level Accuracy:** If even one of the 5 PII fields is under-redacted, or if an additional field is over-redacted, the document level accuracy is 0%.

Formula: $DLA = 100 - \left( \frac{ed}{td} \times 100 \right)$

Where

- $DLA$ is Document Level Accuracy (in percent);
- $ed$ is documents with errors; and
- $td$ is total number of processed documents

Math: $0 = 100 - \left( \frac{1}{1} \times 100 \right)$
Page Level Accuracy: If even one of the 5 PII fields is under-redacted, or if an additional field is over-redacted, the page level accuracy is 75%.

Formula: $PLA = 100 - ((ep/tp) \times 100)$

Where
- $PLA$ is Page Level Accuracy (in percent);
- $ep$ is pages with errors; and
- $tp$ is total number of processed pages

Math: $75 = 100 - ((1/4) \times 100)$

PII Element/Field Accuracy: If even one of the 5 PII fields is under-redacted, or if an additional field is over-redacted, the PII Element/Field level accuracy is 80%.

Formula: $FA = 100 - ((ef/tf) \times 100)$

Where
- $FA$ is Field Accuracy (in percent);
- $ef$ is fields with errors; and
- $tf$ is total number of processed fields

Math: $80 = 100 - ((1/5) \times 100)$

Second example: one document, 3 pages long, each page has 4 PII fields requiring redaction, and two of the fields from one of the pages was not redacted.

Document Level Accuracy: If even one of the 12 PII fields is under-redacted, or if an additional field is over-redacted, the document level accuracy is 0%.

Formula: $DLA = 100 - ((ed/td) \times 100)$

Where
- $DLA$ is Document Level Accuracy (in percent);
- $ed$ is documents with errors; and
- $td$ is total number of processed documents

Math: $0 = 100 - ((1/1) \times 100)$

Page Level Accuracy: One of the 3 pages is under-redacted, so the page level accuracy is 67%.

Formula: $PLA = 100 - ((ep/tp) \times 100)$

Where
- $PLA$ is Page Level Accuracy (in percent);
- $ep$ is pages with errors; and
- $tp$ is total number of processed pages

Math: $67 \approx 100 - ((1/3) \times 100)$
PII Element/Field Accuracy: Two of the 12 PII fields are under-redacted, so the PII Element/Field level accuracy is 83%.

Formula: \( FA = 100 - \left( \frac{ef}{tf} \times 100 \right) \)

Where
- \( FA \) is Field Accuracy (in percent);
- \( ef \) is fields with errors; and
- \( tf \) is total number of processed fields

Math: \( 83 \approx 100 - (2 ÷ 12) \times 100 \)

There is some difference of opinion in the industry as to which level of accuracy is more meaningful in judging the value of a redaction solution. One opinion is that calculating accuracy on a page level basis allows you also to set a level of user validation performance with accuracy (e.g., 6,000 pages a day and 3 pages with errors). Another opinion is that page level accuracy is not correct; if a page contains multiple fields that are incorrectly redacted, these incorrect redactions should be counted as multiple errors.

Simply put, field level accuracy provides the number of fields incorrectly redacted, while page level accuracy provides the number of pages incorrectly redacted, and document accuracy provides the number of documents incorrectly redacted. While one measure of accuracy may be beneficial in benchmarking vendor solutions on small document sets with known control data, another may be more meaningful in larger volume production processing where the data is not part of a controlled set. Neither is right or wrong.

As a best practice, select an accuracy measuring methodology you are comfortable in using and apply it evenly across all your benchmarks.

**Accuracy on Large Repositories**

A forewarning on accuracy computations using large sample sets: the manual accuracy calculation is itself problematic due to the sheer volume of data to manually inspect to determine what should have been redacted. In other words, you can only know how “wrong” you are by first knowing how “right” you were supposed to be. It is difficult to find the staff or the time to accurately inspect thousands, if not hundreds of thousands, of records. Therefore, a simpler (yet production proven) solution is described here to handle computation of accuracy on large image volumes.

As previously stated, the large repository accuracy problem is solved by using a statistically correct subset of documents, reducing the number of images for inspection to a manageable quantity. Multiple sample subsets can be used to benchmark a vendor’s initial accuracy, and secondary sample sets (that exclude documents previously selected) can be used to validate that redaction accuracy is indeed in line with the original statistical models. Common randomized statistical sampling allows a manageable yet representative set of documents from all document types from all years of data to be produced and
processed in the accuracy calculations, thus removing the problem of calculating the “how right” you should be number.

Example of Large Repository
Statistically Reduced for Accuracy Testing

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Document Type 1</td>
<td>Subset 1 Doc Type 1</td>
<td>Subset 2 Doc Type 1</td>
</tr>
<tr>
<td>Document Type 2</td>
<td>Subset 1 Doc Type 2</td>
<td>Subset 2 Doc Type 2</td>
</tr>
<tr>
<td>Document Type 3</td>
<td>Subset 1 Doc Type 3</td>
<td>Subset 2 Doc Type 3</td>
</tr>
<tr>
<td>Document Type “N”</td>
<td>Subset 1 Doc Type N</td>
<td>Subset 2 Doc Type N</td>
</tr>
</tbody>
</table>

Statistical Subsets
11 LEGISLATIVE MANDATES

To address the growing concerns and challenges to policymakers to balance the public's right to information with an individual's right to privacy, PRIA's Records Access & Privacy Policy Committee has developed this paper and offers these best practices to policymakers and Recorders. It is specifically intended as a tool for legislators to be used before they craft legislation and for Recorders as they begin to address redaction.

Most redaction software contains features which allow users to comply with new legislative mandates that require additional PII to be redacted. Some software solutions can locate, categorize, and save additional PII data items upfront and turn the redaction zones/coordinates on if required in the future. Other solutions will require new rules to be written in the future (often supplied by the vendor at no additional cost), but such solutions would require the reprocessing of OCR data which does have storage and processing cost associated with it.
12 SOURCE DOCUMENT CONSIDERATIONS

Reasons for converting paper documents to digital images include:

1. **Establish redundancy** – Paper can decay, be lost or stolen, and is susceptible to fire, flood, insects, etc. Digital imaging enables a digital copy to be available.

2. **Ensure document security** – Losing documents may be costly to any organization, as it may cause a loss of productivity or an embarrassing situation. Having information in a secured environment prevents ad hoc deletions and can reduce risk.

3. **Enable access to content anywhere** – Providing secure image access over the Internet allows users to find the image they need whenever desired.

4. **Increase the readership of the document** – Easy access to images means that images or a link to images may be sent to any user anywhere in the world. No longer are users tethered to a physical document for information needed for job function.

5. **Reduce copying and shipping costs** – Copying and shipping paper documents via mail or other carrier is costly. Scanning reduces the need for a physical copy. If a physical copy is needed, it may be printed at the user’s location instead of being shipped to the user.

6. **Reduce both on- and off-site paper storage costs** – By reducing the amount of paper in any organization, the organization’s storage costs go down. While there are costs associated with storing and preserving images, these costs tend to be much lower than their paper equivalents, and the costs can be minimized with sound records management and retention policies.
13  REDACTION POLICY CONSIDERATIONS: ASKING THE RIGHT QUESTIONS

While there are specific questions that organizations must ask when adopting redaction policies, basic considerations should include:

Risk: The easiest way to minimize risk is to limit the amount of information that is collected, stored and shared.

Questions should include:

- Are there clear guidelines for what needs to be collected and what does not?
- Are there laws or policies currently in effect that protect the unlawful dissemination of sensitive information, and if so, will making a certain document accessible to the public conflict with these laws or policies?
- Are there ways other than redaction to ensure that personal information which is unnecessary will be omitted from retained records?
- Could the Recorder be held criminally liable for the publication of personal information?
- How is personal information being protected by reasonable security safeguards against risks such as loss, unauthorized access, unauthorized destruction, unauthorized use, unauthorized modification, or accidental disclosure?
- Is the information that is available subjecting citizens to security or privacy risks, and if so, does the degree of risk provide the government with a compelling interest to redact that information?
- What information must be collected, and do current practices minimize the amount of personal information that is collected?

Measurable Outcomes: Policymakers should pay close attention to how success rates for redaction will be measured.

- Are practices being employed to ensure that redaction policies are accurate, are effective, and are being enforced?
- How will success be defined and measured (e.g., error rates, reduction in risks to citizens, public satisfaction through perceived reduction in risks)?
- How will records that differ over time (due to image quality, document structure, fonts, etc.) be treated during the redaction process?
- What are the specific practices, and who is accountable for making sure they are carried out?
- What is the realistic success rate that redaction practices should be meeting?
- What is an acceptable error rate?
- What is the percentage of documents that will need to be manually reviewed or corrected to provide this error rate?
Costs: Removing information from legacy data or records, or even taking on the responsibility to ensure that certain personal information is not collected, can be expensive.

- Policymakers should develop an explicit understanding of these costs.
- Are cost offsets available and provided by the redaction process? For example, the use of the OCR data generated by the redaction process can also be used for automated data capture and automated indexing.
- If a Recorder’s database needs to be temporarily or permanently removed from the Internet, what is the cost for taking it offline and processing records requests internally?
- What is the cost, financial or otherwise, to consumers of these services?
- Will redaction of certain identifying information render certain functions useless (e.g., federal tax liens)?
- What reaction can be anticipated from the public as well as specialized users of land records, and how can it be effectively addressed?
- Who are the primary users of the system?
- Who will manage the work?
- Who will pay for these costs?

Education: It is vitally important that policymakers develop and conduct a comprehensive education and outreach plan.

- Does the education and outreach plan emphasize why the personal information is required at the time of collection and how it will be disclosed (online or otherwise)?
- Does the education and outreach plan stress to bulk filers and other stakeholders the importance of not collecting unnecessary or extraneous information and how the law and its penalties apply?
- Does the education and outreach plan convey the importance of online government services?
- Is there a dialogue aimed to ensure that citizens comprehend what information about them is actually available in public records accessible online and the steps they can take to have their protected personal information redacted or removed from the public documents?
- Is there a general policy that indicates a commitment to educating the public and other key stakeholders about redaction policies and practices?

Storage Components Issues:

- Does the redaction solution require additional permanent storage of OCR/ICR data or any output processing files?
- Does your existing system provide for storage of redacted and non-redacted documents or will you be duplicating your entire image repository?
- Is any of your data being stored offsite by the vendor to allow for processing?
Additional considerations for redaction policymaking:

- Add penalties for document preparers who fail to remove PII from documents intended for recording.
- Pay close attention to agreements with third party vendors, such as redaction software vendors, who will have access to sensitive data and contractually obligate them to safeguard all data.
- Recommend any measures taken or considered to be permitted by law.
14 CONCLUSION

It is important that key stakeholders understand that government services and available data are not contributing to the erosion of privacy or personal security. As busy legislatures grapple with redaction laws, they are wise to avoid costly and ineffective blanket solutions that may not totally appease a worried public. Lawmakers should also be mindful that legislation does not adversely impact constructive notice elements in their jurisdictions so as to impair the chain of title for real property. Some information is widely available online, and embarking on an expensive redaction effort may not actually result in protecting individuals.

As a best practice, legislators should consult with PRIA and other like organizations such as NACRC www.nacrc.org, IACREOT www.iacreot.com, and ALTA® www.alta.org during bill development.

Bearing in mind the potential for identity theft, information security is of paramount concern. Redaction is the most popular answer to masking data in the public record from public view. Redaction is the process of removing sensitive data from the publicly viewable government record. Redaction allows sensitive, private, and confidential information, which is exempt by law from disclosure, to be concealed in a manner that does not alter the meaning of the record. To comply with current statutes and laws that provide open access to public records without exposing individuals’ PII, officials are responding with the utilization of both automated and manual redaction tools.

As a best practice, PRIA recommends the use of automated redaction tools for digital images.

Automated redaction software can help counties make their land records available in a secure manner. Advanced technology can be harnessed to save labor costs and eliminate a significant percentage of tedious data entry tasks.

PRIA’s Records Access & Privacy Policy Committee has collaborated with industry partners and stakeholders to set forth best practices for policymakers and Recorders to address redaction. This paper is specifically intended as a resource tool for legislators before they craft legislation and for Recorders as they begin to address redaction, the availability of PII contained in land records documents, and the technical solutions of redaction.
## APPENDIX A: BEST PRACTICES

<table>
<thead>
<tr>
<th><strong>State Statutes</strong></th>
<th>The foundational best practice for custodians of land records is to follow their specific state statutes regarding their duty to redact personally identifiable information and to provide public access to the records.</th>
<th>Pages 1, 7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consult with Subject Matter Experts</strong></td>
<td>As a best practice, legislators should consult their Recorders, as well as PRIA and other like organizations such as the National Association of County Recorders, Election Officials, and Clerks (NACRC), the International Association of Clerks, Recorders, Election Officials, &amp; Treasurers (IACREOT) and ALTA. These organizations and individuals provide a broad, systemic view of the purpose of public records and would be directly affected by redaction legislation.</td>
<td>Pages 2, 42</td>
</tr>
<tr>
<td><strong>Automated Software</strong></td>
<td>PRIA recommends the use of automated redaction software for covering up data from digital images.</td>
<td>Pages 2, 42</td>
</tr>
<tr>
<td><strong>Redacted Version Access</strong></td>
<td>PRIA recommends allowing full open access to the redacted version of a document, while maintaining the un-redacted copy as the master/legal document.</td>
<td>Pages 3, 24</td>
</tr>
<tr>
<td><strong>Document Marking</strong></td>
<td>Avoid marking on documents in such a way that inhibits the effectiveness of OCR software on that document.</td>
<td>Page 3</td>
</tr>
<tr>
<td><strong>300 DPI/TIFF Group 4</strong></td>
<td>When OCR is a part of the redaction process, scanning documents to produce images at 300 DPI Tagged Image File Format (TIFF) Group 4 is the desired file format for best character recognition accuracy (as of the time of publishing this paper).</td>
<td>Pages 3, 28</td>
</tr>
<tr>
<td><strong>Benchmark Testing</strong></td>
<td>Conduct accuracy benchmarks prior to purchasing an automated redaction solution and apply the same benchmarks throughout the use of the system to ensure the redaction processes are producing expected results.</td>
<td>Pages 4, 31</td>
</tr>
<tr>
<td><strong>Go Forward Benchmark Testing</strong></td>
<td>For evaluating accuracy in &quot;go forward&quot; file processing, perform a separate onsite vendor accuracy benchmark, providing the vendor with documents that have not been supplied previously.</td>
<td>Page 4</td>
</tr>
<tr>
<td><strong>Property Address</strong></td>
<td>Property addresses should NOT be redacted from land records.</td>
<td>Page 8</td>
</tr>
<tr>
<td><strong>Master or Authoritative Documents</strong></td>
<td>Authoritative or master/legal documents should NOT be redacted. Redaction should be applied only to the publically accessible copies of the documents.</td>
<td>Page 9</td>
</tr>
<tr>
<td><strong>Signatures</strong></td>
<td>Signatures should NOT be redacted from land records.</td>
<td>Page 9</td>
</tr>
<tr>
<td>Automated Redaction Tools</td>
<td>PRIA recommends the use of automated redaction tools for digital images.</td>
<td>Page 10</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Statistically Correct Sample Sets</td>
<td>Create statistically correct sample document sets for your existing repository.</td>
<td>Page 33</td>
</tr>
<tr>
<td>Accuracy Measuring Methodology</td>
<td>Select an accuracy measuring methodology you are comfortable using and apply it evenly across all your benchmarks.</td>
<td>Page 35</td>
</tr>
</tbody>
</table>
# APPENDIX B: GLOSSARY OF TERMS

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td>The number, street name, and other information that describes where a building is or where somebody lives.</td>
</tr>
<tr>
<td>Annotation</td>
<td>In terms of redaction, the notes/comment box is used to cover up sensitive information within documents. Typically, the annotation is not permanent and can be removed to access sensitive information unless it is burned-in the document.</td>
</tr>
<tr>
<td>Backfile Conversion</td>
<td>The process of converting large volumes of documents usually accumulated over time. Generally, backfile conversion consists of digitizing files or documents using high-speed scanning methods.</td>
</tr>
<tr>
<td>Bank or Financial Account Numbers</td>
<td>A bank account number is issued by a financial institution to an account holder at the time an account is opened and is unique to the individual account. A bank uses this account number to identify individual holders of accounts. Bank account numbers are usually eight or more digits.</td>
</tr>
<tr>
<td>Batch Processing</td>
<td>A method of executing a series of items via computer all at one time without human intervention. A method of grouping a set of documents by a common factor.</td>
</tr>
<tr>
<td>Best Practice</td>
<td>The term “best practice” means a guideline or specification that is advisory in nature and compliance with which is strongly recommended, however, compliance is not mandatory for anyone.</td>
</tr>
<tr>
<td>Bi-tonal image</td>
<td>An image consisting only of a foreground color and a background color.</td>
</tr>
<tr>
<td>Bitmap Image (BMP)</td>
<td>Bitmap Images (BMP images) are commonly used by Microsoft® Windows® programs, and the Windows® operating system itself. BMP is the only graphics format where compression actually enlarges the file.</td>
</tr>
<tr>
<td>Credit and Debit Card Numbers</td>
<td>The numbers found on credit cards have a certain internal structure, and share a common numbering scheme. Credit card numbers are a special case of ISO 7812 numbers. An ISO 7812 number contains a single-digit major industry identifier (MII), a six-digit issuer identifier number (IIN), an account number, and a single digit checksum using the Luhn algorithm. The MII is usually considered to be part of the IIN.</td>
</tr>
<tr>
<td>Data</td>
<td>The term “data” includes, but is not limited to, information in a database, information about an operating system (OS), operational policies and procedures, system design, organization policies and procedures, and system status.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Day-forward</strong></td>
<td>The term “day-forward” redaction means the redaction of PII from land records that are created or placed into the Recorder's database on and after the implementation of redaction by the Recorder.</td>
</tr>
<tr>
<td><strong>Date of Birth</strong></td>
<td>The month, day, and year someone was born.</td>
</tr>
<tr>
<td><strong>Dots per Inch (DPI)</strong></td>
<td>A measure of the resolution of a printer is called Dots per Inch (DPI). The measure properly refers to the dots of ink or toner used by an image setter, laser printer, or other printing device to print text and graphics. In general, the more dots per inch, the better and sharper the image. While DPI is not officially a measure of image resolution, it is frequently used that way.</td>
</tr>
<tr>
<td><strong>Driver's License Number</strong></td>
<td>Driver's license numbers are numbers assigned to drivers by their state's driving licensing agency. The numbers are essentially a way to identify and separate an individual's information from all the other drivers in the state.</td>
</tr>
<tr>
<td><strong>Email address</strong></td>
<td>An email address identifies an email box to which email messages are delivered. An example format of an email address is <a href="mailto:lewis@example.net">lewis@example.net</a> which is read as “lewis at example dot net.”</td>
</tr>
<tr>
<td><strong>Encryption</strong></td>
<td>The term encryption refers to the use of a software key in a computer process which converts data to something incomprehensible, so that the data can be re-converted only by an authorized recipient holding the matching key.</td>
</tr>
<tr>
<td><strong>False Positive</strong></td>
<td>A result that is erroneously positive when a situation is normal. An example might be a nine digit Zip+4 identified by redaction software as an SSN.</td>
</tr>
<tr>
<td><strong>Federal Employer Identification Number (FEIN)</strong></td>
<td>A Federal Employer Identification Number (FEIN) is a number issued by the federal government which is used to identify taxpayers that are required to file various business tax returns. The FEIN is also known as a Federal Tax Identification Number.</td>
</tr>
<tr>
<td><strong>GIF</strong></td>
<td>The Graphics Interchange Format (GIF), a bit-mapped graphics file format that is used extensively on the Internet. The standard supports only 255 colors per frame, so it requires “lossy” quantization for full-color images (dithering); color precision can be improved through the use of multiple frames. GIF images can use lossless LZW (Lempel-Ziv-Welch) compression.</td>
</tr>
<tr>
<td><strong>Hybrid Processing</strong></td>
<td>The ability to redact some documents in a fully-automated environment without human intervention, and pass other documents to human verifiers for review before redacting.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Intelligent Character Recognition (ICR)</strong></td>
<td>Intelligent Character Recognition (ICR) is an advanced optical character recognition (OCR) system that allows fonts and different styles of handwriting to be learned by a computer during processing to improve accuracy and recognition levels.</td>
</tr>
<tr>
<td><strong>JPEG</strong></td>
<td>The Joint Photographic Experts Group (JPEG) standard is used extensively for photos and other continuous tone images on the web. The standard uses “lossy” compression by trying to equalize eight by eight pixels; the quality can vary greatly depending on the compression settings.</td>
</tr>
<tr>
<td><strong>JPG2000</strong></td>
<td>JPEG2000 is the new wavelet based file format from the Joint Photographic Experts Group (JPEG) consortium and has to be considered, so far, the best raster file format available in computer science. JPEG2000 also includes a lossless compression option.</td>
</tr>
<tr>
<td><strong>Lossy</strong></td>
<td>“Lossy” file compression results in lost data and quality from the original version. Lossy compression is typically associated with image files, such as JPEG, but such compression can also be found in audio files, like MP3s or AACs.</td>
</tr>
<tr>
<td><strong>Manual Redaction</strong></td>
<td>Recording offices may redact information manually by creating a copy of the document and manually crossing out or covering the information to be redacted. The redacted copy will then be added to the public record.</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>Data about data. By describing the contents and context of data files, the usefulness of the original data/files is greatly increased.</td>
</tr>
<tr>
<td><strong>Microfiche</strong></td>
<td>A microfilm card or sheet used in some information storage systems; consists of a film format about 4 by 6 inches (10 by 15 centimeters) containing micro-images of information, with a title heading large enough to be read by the unaided eye.</td>
</tr>
<tr>
<td><strong>Microfilm</strong></td>
<td>Greatly reduced film of records, books, newspapers, engineering drawings, reports, and manuscripts; copies are made on fine-grain film of 16, 35, 70, and 105-millimeter width, permitting easy storage and handling.</td>
</tr>
<tr>
<td><strong>Optical Character Recognition (OCR)</strong></td>
<td>Optical Character Recognition (OCR) is the mechanical or electronic translation of images of handwritten or typewritten text (usually captured by a scanner) into machine-editable text.</td>
</tr>
<tr>
<td><strong>PDF</strong></td>
<td>Portable Document Format (PDF) is an open standard for document exchange. This file format, created by Adobe Systems in 1993, is used for representing documents in a manner independent of application software, hardware, and operating systems.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>PDF/A</strong></td>
<td>PDF/A is an ISO-standardized version of the Portable Document Format (PDF) standard. PDF/A is specialized for the digital preservation of electronic documents.</td>
</tr>
<tr>
<td><strong>PDF/E</strong></td>
<td>PDF/E is a subset of the Portable Document Format (PDF) standard designed to be an open and neutral exchange format for engineering and technical documentation</td>
</tr>
<tr>
<td>** Personally Identifiable Information (PII)**</td>
<td>Personally Identifiable Information (PII) is any information relating to an identified or identifiable individual who is the subject of the information (e.g., social security number, date of birth, mother's maiden name, address).</td>
</tr>
<tr>
<td><strong>Photostat</strong></td>
<td>A type of high contrast photographic negative or positive in the form of paper. Also referred to as “Stat.”</td>
</tr>
<tr>
<td><strong>PNG</strong></td>
<td>Portable Network Graphics (PNG) is a 24 bit color depth image format with lossless compression. Originally designed to replace the use of GIF on the web.</td>
</tr>
<tr>
<td><strong>Public Record</strong></td>
<td>The federal Freedom of Information Act requires almost every unclassified document possessed by a federal agency to be available to the public for inspection and copying. States, in turn, have each passed similar legislation requiring virtually all levels of state and local government to make their records available to the state’s citizens for inspection and copying. While the records may pertain to the government’s financial information and operations, there are a few exemptions that restrict the release of certain records. Additionally, land records have also been subject to the common law right of access, since the purpose of establishing and maintaining land records is to provide constructive notice of the contents of the record to the public.</td>
</tr>
<tr>
<td><strong>Quality Control</strong></td>
<td>The term “quality control” means the assessment of product compliance with stated requirements. Quality control analysis should be performed by personnel who are independent from production personnel.</td>
</tr>
<tr>
<td><strong>Redaction</strong></td>
<td>The act of covering over or otherwise removing from the public record or public view any sensitive, private, or confidential information.</td>
</tr>
<tr>
<td><strong>Resolution</strong></td>
<td>The maximum number of pixels that can be displayed or scanned, expressed as (number of horizontal pixels) x (number of vertical pixels), e.g., 1024x768. See also Dots per Inch (DPI).</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Social Security Number (SSN)</strong></td>
<td>In the United States, a Social Security number (SSN) is a nine-digit number issued to U.S. citizens, permanent residents, and temporary (working) residents under section 205(c)(2) of the Social Security Act, codified at 42 U.S.C. § 405(c)(2). The number is issued to an individual by the Social Security Administration, an independent agency of the United States government. The SSN's primary purpose is to track individuals for taxation purposes.</td>
</tr>
<tr>
<td><strong>Source Document</strong></td>
<td>The original medium containing the basic data to be used by a data-processing system, from which the data are converted into a form which can be read into a computer. Also known as original document.</td>
</tr>
<tr>
<td><strong>Subject Matter Expert</strong></td>
<td>An individual who is an authority in a particular area or topic.</td>
</tr>
<tr>
<td><strong>Taxpayer Identification Number (TIN)</strong></td>
<td>A Taxpayer Identification Number (TIN) is an identification number used by the Internal Revenue Service (IRS) in the administration of tax laws. It is issued either by the Social Security Administration (SSA) or by the IRS. A Social Security number (SSN) is issued by the SSA whereas all other TINs are issued by the IRS.</td>
</tr>
<tr>
<td><strong>Tagged Image File Format (TIFF)</strong></td>
<td>Tagged Image File Format (TIFF) is used extensively for traditional print graphics. Both “lossy” and lossless compression is available, but many programs only support a subset of the available options.</td>
</tr>
<tr>
<td><strong>Uniform Commercial Code (UCC)</strong></td>
<td>The Uniform Commercial Code (UCC) is a comprehensive model statute addressing most aspects of commercial law. Currently, the UCC has been enacted in 49 states, the District of Columbia, and the U.S. Virgin Islands. Louisiana has only partially adopted the UCC.</td>
</tr>
<tr>
<td><strong>User</strong></td>
<td>The term “user” means an individual or group who has access to an information system or its data.</td>
</tr>
</tbody>
</table>
APPENDIX C: RELATED DOCUMENTS

While this paper focuses on redaction issues related to land records, it should be acknowledged that there are additional and different issues related to redacting other collections of documents, many of which are connected to property ownership.

Redaction of documents filed in legal (court) proceedings may include the names of minor children, identifying information for public officials (e.g., police officers), and confidential intellectual property. The rationale and techniques available for the redaction of such legal documents extend beyond those for land records. The following list of public records outside the property records arena is provided only to offer insight to the variety of sensitive information that might be contained therein.

- **Birth certificates** are often on file in the county in which the birth occurred and at the Office of Vital Statistics within each state. Birth records usually contain the name of the child, date and time of birth, the city of birth, and the hospital in which the child was born, the parents' names, the attending physician’s name, and various signatures. Confidentiality policies differ by county and state.

- **Court records**, unless the record involves a juvenile or incapacitated person, court records are usually considered to be public records. Superior, circuit, district, municipal, and small claims court records are most often kept in the Court Clerk’s office. The Court Clerk maintains an index of all cases under its jurisdiction. Cases may be indexed by case number, alphabetical order by the names of the parties involved, by attorney name, or a variety of other identifying characteristics. The physical or digital case files can most often be retrieved under the name of either the plaintiff or the defendant. The files contain the initial complaint, the defendant’s answer, and motions filed in the case. Case files may also contain evidence or exhibits that were used in court. Court records are increasingly available on the Internet.

- **Death certificates** are also public documents. They are usually kept on file in the county in which the death occurred at the County Recorder’s/Clerk’s office. Death certificates often contain the name of the person who died, where the death occurred, the date of death and the person’s Social Security number.

- **Divorce records** are also generally considered to be public documents and are usually part of court files. They are most often filed at the Court Clerk’s office of the county in which the divorce was granted. The files may contain financial information such as bank accounts, debts, and assets of the parties.

- **Marriage certificates** are usually filed in the County Recorder’s/Clerk’s office where the marriage application was filed, as well as in the Office of Vital Statistics within each state. An index is usually available to the public, either online or in the appropriate office. The index contains usually contains the bride’s and groom’s
names, the county where the application was filed, the date of the marriage, and may also contain birth dates, number of prior marriages, and other PII.

- **Voting records** are typically kept at the County Clerk’s or Registrar of Voters office and at the Secretary of State offices. Registered voter records might contain address, phone number (even unlisted numbers) and sometimes Social Security numbers.