

Secretary of State Redaction Symposium 2007

What to ask when buying a redaction solution? (Version 2.1)

Henry Sal hsal@csisoft.com

If I were buying a redaction solution, what questions would I want answered?

- 1) How exactly does the solution identify what is to be redacted?
- 2) Is the technology: a) rule based; b) self learning; c) learns by example; d) is a hybrid?
- 3) If learning technology (self learning or learn by example):
 - a) How many documents are required to be processed and validated for it to learn what to do correctly on a new document type or additional redaction field (several hundred or several hundred thousand)?
 - b) How is incorrect learning prevented?
 - c) Do I have the capability for an administrative user to teach it what to do onsite?
- 4) If rules based, can my staff specify what to redact so we may accommodate changing requirements without requiring additional vendor services or is it a “closed” proprietary system?
- 5) Upon completion of processing does the solution vendor require keeping a copy of my private documents? If so, why do they require this and does the law allow me to provide such to them?
- 6) How does the solution handle non expected data that is valid information to redact? Specific examples are of embedded SSN fields? Composite SSN fields? Masquerading SSN fields? If I find one of these on page 1 of a multipage document, does the solution accommodate their redaction on subsequent pages?

Embedded fields are data valid for redaction that is embedded within another numeric field (i.e. nnnSSNnnn with SSN being a valid Social security number and the others characters directly attached to the field being anything).

Composite fields are data valid for redaction having non expected characters either at the start or end of the string (i.e. 999-99-9999-a with the 9's being a properly formed SSN).

Masquerading fields are data valid for redaction being completely mislabeled and of improper format but once again being valid for redaction. (i.e. ID 99-9999999 with this actually being a valid SSN for an individual in an improper format and mislabeled).

- 7) How am I scoring my accuracy? For example, will 3 missed redactions on a single image count as 3 separate errors or just 1? How am I going to score redaction of fields that should not be redacted (over redactions)?
- 8) Do I have to invent my own score keeping or does the vendor provide accuracy scoring within their solution? If so, how is this performed and does it make sense? (Note: Well known double blind key data entry accuracy formulas have little to do with redaction validation; you are not performing data keying operations so the 99.5% rule for double key entry does not apply).
- 9) Does the vendor have a solution that “future proofs” my investment by allowing me to specify additional fields to process now, redacting a subset now, but then allow me to redact the other fields in the future without additional processing costs? If so, how do they propose doing this and are the fields for future use validated now as well?
- 10) How does the solution handle large volumes?
- 11) Is the solution fault tolerant? If so, how?
- 12) How does the solution handle poor image quality? Does it just process raw OCR data or include “fuzzy” logic?
- 13) Who else is using the vendor’s solution?
- 14) Does the vendor have experience in my industry and with my document types / formats?
- 15) Is the vendor reselling other peoples technology, and if so, what value do they add to the process?
- 16) Are subcontractors being employed in the solution and if so for what pieces?
- 17) What is the vendor doing in R&D to improve any investment I make in their technology? How are they recognized in the industry?
- 18) Whose character recognition engines is the vendor using; do the engines provide handprint recognition for redaction or in do I have to look at every image to make sure I redact hand printed information. (OCE RecoStar Pro; Abby fine reader, or Scansoft (now Nuance) are three of the known ones, no one writes their own engines).
- 19) Does the vendor provide an engine that will detect and/or redact cursive script?
- 20) Does the vendor’s solution have the ability to locate and/or redact handprint absent machine print keywords?

- 21) Will the redaction engines find account numbers on Checks that are MICR fonts?
- 22) Does image clean up processing exist and if so what clean ups are performed?
- 23) Can the system provide me co-ordinates of redacted fields in case I can't store a redacted image in my system? Does the system provide alternative mechanisms for removing private information?
- 24) Can the software identify and redact images in both landscape modes or do I have to perform image preparation to locate and rotate the images before processing?
- 25) As an addition to the landscape orientation question, can an image contain vertical text as well as normal horizontal text and the software locates and redacts the vertical text?
- 26) What is the exact process the vendor will follow to provide me with redacted images?
- 27) If the vendor is processing offsite, how is integrity of the images provided for during their processing now that I have my image in two locations provided for?
- 28) Can I see a demonstration of the solution onsite processing my exact documents in real time? I don't want to send my documents away to be processed and just see the results. I want to see accuracy out of the box, on my documents, for all fields that I intend to redact; account numbers, DOB's, crummy images, etc... (The solutions accuracy without human intervention is important as back filing corrections are handled by vendor staff, but forward filing corrections requires your own manpower).
- 29) What happens when the software misses?
- 30) What is the vendor's process for refinement of accuracy?
- 31) Is accuracy automatically measured or only addressed when I identify mistakes? If automatic, how?
- 32) What % of images does the vendor require to be validated to achieve their stated accuracy?
- 33) Are software tools provided for me to validate the results?
- 34) Do I own the processing database when processing is complete? If so, in what format is it provided?
- 35) What document types is the software capable of processing? (TIFF, GIF, JPEG, PDF/Image, PDF/Text?)
- 36) For PDF/Text documents, how is the text redacted from the text layer of the final document?
- 37) Do I have information on microfilm that I need redacted?

- 38) Does the vendor's solution accommodate alternative media of microfilm and paper? If so, how?
- 39) Are any document / image / meta data consistency checks performed in the processing?
- 40) What happens if the accuracy is not met?
- 41) How does the vendor's solution interface with my existing system(s)?
- 42) Does the vendor provide a solution for new documents?
- 43) What interfaces does the vendor provide "out of the box"?
- 44) Can I streamline my operations (increase speed and consistency of data capture) using extraction of data from my documents?
- 45) Can I use the vendor's product to locate and extract data as well?
- 46) Is extraction a separate product at an additional cost?
- 47) Does extraction require a separate pass of image processing and double the resources or is it handled at same time as redaction?
- 48) Does the solution allow me to validate redactions and extractions in a single user step using a single user interface or is the redaction validation separate from extraction validation and require additional manpower?
- 49) How long does processing take per image?
- 50) What additional equipment will I need?
- 51) What is the cost of the software?
- 52) What is my projected cost in additional human resources to perform any forward file validation processing?