

DOCUMENT CONVERSION AND XML GENERATION SOLUTION RFI Q&A
March 28, 2012

No.	Section No.	Sub-Section/ Page No.	Question	Answer
1.	General	General	The Request for Information (RFI) provides offerors a specific framework (OCR tool identified, formats for XML, time requirements, etc.) for what the USPTO needs. Is USPTO looking to identify companies that have the capabilities to develop a solution exactly as specified in the RFI or is USPTO interested in alternative concepts or ideas?	Yes, the USPTO would like to identify companies that can at least deliver the exact solution specified. However, the USPTO is not closed to competitive, immediately implementable alternatives.
2.	General	General	Does USPTO have a preferred automated method to convert the text data from PrimeOCR into indexed XML?	PrimeOCR is the preferred method.
3.	General	General	Is the scope of this RFI to provide conversions for incoming applications or converting the existing patent data? Also, if it's incoming applications, what format is the data being received initially?	This scope of the RFI is for the conversion of incoming documents, not backfile. Initial receipt of data will be TIFF format.
4.	Section 1.1	Page 4	Has or is USPTO currently using Prime OCR version 5.3 to perform .tiff to text conversion? If so, can USPTO provide data as to the current performance (accuracy in percentage, speed in number of OCRed Image pages per hour) of Prime OCR version 5.3? Will the USPTO also be willing to provide the per page unit cost?	The USPTO has used PrimeOCR to provide the conversion specified in the RFI. Prime Recognition provides performance data for its product but a single CPU core processes a page in about 3 seconds with the specified configuration. Due to typographical error, "version 5.3" was incorrect. The correct version is 5.1. The USPTO will not provide per page unit cost. There is an economy of scale involved.
5.	General	General	Is USPTO looking for a point forward solution, or are they wanting us to convert current text based images to XML?	The USPTO is looking for a point-forward solution.

DOCUMENT CONVERSION AND XML GENERATION SOLUTION RFI Q&A
March 28, 2012

No.	Section No.	Sub-Section/ Page No.	Question	Answer
6.	Section 1	Page 1	They mention three document types, SPEC, CLM, ABST. Are the forms associated within each document type identical, or do they vary?	They each use the same schema but with different set of optional tags. The conversion utility uses the document type to define the algorithm used for tagging the OCR output.
7.	General	General	Will you send us the patent applications in scanned printed tiffs? Also, I would like to know if some of the applications will have some handwritten texts. As handwritten text needs to be captured manually, OCR always doesn't ensure accurate results.	It is anticipated an implemented solution will include transmission of TIFF images corresponding to pages from a document. There will be no manual handling of data with this solution. The USPTO utility will process handwriting automatically as an artifact.
8.	General	General	Can we get the DTD/XSD used in the conversion?	Will be available if the Office proceeds to an RFQ.
9.	General	General	Are there any sample scanned tiff/tiff image(s) that you can show us.	Sample images can be obtained from the USPTO Public Pair website.
10.	Section 1.1 (Feature 1.2)	Page 2	Given Feature 1.2. requirement, what is the maximum number of pages we should expect to process in two hours?	The USPTO receives up to about 5000 filings a day. Each filing averages about 10 pages. The largest filing has been as much as 17,000 pages. The smallest a single page.
11.	Section 1.1 (Feature 1.5.2)	Page 3	In Feature 1.5.2., with what accuracy is the PTO's utility identifying artifacts (i.e., how frequently does the utility fail to correctly identify artifacts, and how frequently does it yield false hits)?	PTO's utility exhibits an accuracy of above 99% in identifying artifacts in the small scale tests performed so far on very complex documents. Large scale testing is planned for later this spring.
12.	Section 1.1 (Feature 1.5.2)	Page 3	Can the PTO utility described in Feature 1.5.2 be provided for evaluation purposes?	Will be available if the Office proceeds to an RFQ.
13.	Section 1.1 (Feature 1.6)	Page 4	In Feature 1.6., what is the PTO's expectation for OCR accuracy using Version 5.3 of PrimeOCR,	Prime Recognition provides accuracy information with respect to its product. Actual

DOCUMENT CONVERSION AND XML GENERATION SOLUTION RFI Q&A
March 28, 2012

No.	Section No.	Sub-Section/ Page No.	Question	Answer
			given the specified settings? What is the average number of recognition errors per page?	accuracy is dependent on the input. See the response to Question #4.
14.			As a follow on to the previous question, how many documents has the PTO processed to arrive at this expectation?	The USPTO has transformed 1.8 million pages of claim, specification and abstract documents successfully and is currently transforming another 56 million more pages.
15.	Section 1.1 (Feature 1.7)	Page 4	In Feature 1.7., what is the PTO XML format and what are the tagging requirements?	The PTO-XML format referred to is very similar to XML4IP and ST.36. The PTO would provide appropriate DTDs after an RFQ.
16.	General	General	PrimeOCR Version 5.3 appears to be unavailable as a stand-alone commercial product, but rather is available only as integrated with EMC's Captiva InputAccel Capture Software. However, Version 5.1 is available commercially and according to Prime is substantially the same as Version 5.3. Does the USPTO require the version integrated with EMC's Captiva InputAccel Capture Software, or is use of Version 5.1 an acceptable alternative?	The USPTO apologizes for the typographical error in the RFI. Version 5.1 was intended. See response to Question #4.