

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION
DOC-SS-PAPT-1100009**

THIS IS A REQUEST FOR INFORMATION (RFI) BEING RELEASED PURSUANT TO FEDERAL ACQUISITION REGULATION (FAR) PART 10: MARKET RESEARCH. This RFI is issued solely for informational, market research, and planning purposes only. It does not constitute a Request for Proposal (RFP) or a promise to issue an RFP in the future. This RFI does not commit the Government to contract for any supply or service whatsoever. Further, the United States Patent and Trademark Office (USPTO) is not at this time seeking proposals, and will not accept unsolicited proposals. Respondents are advised that the United States (U.S) Government will not pay for any information or administrative cost incurred in response to this RFI. All costs associated with responding to this RFI will be solely at the responding party's expense. Responses to the RFI will not be returned. Please be advised that all submissions become Government property and will not be returned. Not responding to this RFI does not preclude participation in any future RFP, if any is issued. Responses to this notice are not offers and cannot be accepted by the U.S Government to form a binding contract. It is the responsibility of the interested parties to monitor the Federal Business Opportunities (www.fbo.gov) site for additional information pertaining to this RFI.

1.0 RFI OBJECTIVE

Patents End to End (PE2E) is the United States Patent & Trademark Office's (USPTO) project to deliver a next-generation IT infrastructure supporting Patents business operations. This project will replace the nearly four dozen aging legacy systems used today with a single system that unifies electronic processing over the entire patent application lifecycle (hence "end-to-end"). Most relevant to this RFI, the new system will replace documents that are currently represented as scanned TIFF images with documents represented as structured text in XML.

This RFI seeks to obtain information from interested parties, including the vendor community, about potential solutions to the problem of converting legacy documents to XML format. The Office is exploring a wide range of possible solutions for several different but related needs.

The aim of this market research is to find solutions for this project at different stages. (1) initial processing with ad hoc collections with a very short time frame to make a September 2011 deadline to convert about 5,000 documents, (2) a path towards a fully automated system for converting legacy online documents into fully structured textual format, to begin approximately Sept 2011 and continue for approximately two years, and (3) when the new end-to-end system becomes operational, the intention will be for most information to enter the system as structured text or as text-backed documents, but there will continue to be a need for scanning and conversion of certain documents, so a solution that can accommodate this work going forward is also sought.

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION
DOC-SS-PAPT-1100009**

1.1 Problem Scope

The ideal solution has the ability to take as input a set of documents and automatically or semi-automatically produce a meaningful XML schema based on those documents, formulated to be similar to an existing XML schema designed for patents (called XML4IP). The ideal solution would then convert that set of documents and similar documents to text structured according to the schema while minimizing errors.

For the majority of the document types, it is desirable to have the rendering of the converted documents look as similar to their originals as possible. Lacking this, it is desirable for the locations of boundaries of structured elements (page boundaries, section boundaries, left or right column and line number) recorded and expressed in some manner in the representation of the document (whether or not the information would be rendered visibly).

Some documents contain non-Western characters and require translations from other languages into English.

Some documents contain mathematical formulas, tables, chemical formulas, and other textual or quasi-textual material that can be expressed via XML standards. In those cases, an ideal solution will convert those representations to the appropriate XML standard. Some documents contain figures or images. In the ideal solution, the figures will be recognized as such and any figure numbers, captions, or other identifying information will be associated with those figures. In some cases, documents will contain tables with a combination of figures and text; an ideal solution would use XML standards for representing such information.

For some specific types of documents, a useful substitute for producing a fully structured XML document might be a text summary that uses document markup cues such as titles, headings, bold face, and italics, to produce a short, meaningful text summary of the purpose and/or content of the document, to be displayed as a substitute for metadata and title of the document in a document list view.

An additional need is for technology to convert documents that appear in one XML standard (Redbook) to a new standard (XML4IP).

The potential solutions may vary along several dimensions:

- The proportion of the work that is done automatically versus manually,
- The accuracy of the solution,
- The granularity (detail) of the document structure produced.

This variation may be influenced by several factors pertaining to the documents being analyzed, including but not limited to:

- The amount of structure inherent in the document,

REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION
DOC-SS-PAPT-1100009

- The amount of non-text information in the document (figures, tables, formulae, etc),
- The variation of structure for a given document type,
- The quality of the scan for those documents that are stored as scans,
- The language(s) the document is written in.
- Recognition and handling of extraneous information, such as headers and footers.

1.2 Sample Documents for Conversion

Attachment A contains a sample of legacy documents. Only the first five pages from each document are included, so some documents are truncated. Associated with each type is a code called a “doc code.” The first portion of each file name indicates the doc code. In some cases the doc code represents a wide range of document types, in other cases, documents are misclassified to a doc code, and in still other cases, multiple document types are concatenated into one document and associated with one code although the materials correspond to more than one code. The ideal solution would recognize misidentified doc codes as well as cases in which multiple disparate documents are associated with one code.

The following paragraphs explain the materials associated with the sample doc codes as well as what the ideal solution would extract.

CLM: These files contain patent claims, including amended claims. Ideal processing would identify the claim numbers, the relationships between dependent and independent claims, and would recognize which are original claims, new claims, amended (changed) claims, and deleted claims. (Dependent claims mention the independent claims upon which they depend.) Underlining or italic font are sometimes, but not always, used to indicate new claims, and strikethrough or square brackets are sometimes used to indicate deleted text. Footers can appear at the bottom of the page and should not be confused with the claims listing.

TRNA: This information is sometimes entered as a form (applicants fill out a pdf form and then upload the pdf to the system which then converts it to a tiff file; applicants may also mail or fax the form). However, the applicant is also allowed to submit a free-form document that expresses the relevant information. The purpose of this form is to request a patent reexamination although the doc code is sometimes assigned to other kinds of transmittals. The correct version of the form is shown in sample document TRNA_10-30-2002_90006430.pdf. The ideal solution in this case would extract the relevant information from the form data, identifying the attribute/value pairs, and would also extract the relevant information from free form documents as shown in the example TRNA_01-10-2007_90005710.pdf.

REM: This is a free-form document, usually written in letter form, from the patent application to the office providing arguments or remarks made when amending a patent application. The ideal solution would identify and extract the application number(s),

REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION
DOC-SS-PAPT-1100009

reexamination control number(s), attorney name, date of the letter, and would identify claim numbers whether or not included in tabular format.

IDS: This form contains a list of references to US Patents, foreign patents, and what is called the “non patent literature,” or publications and documents other than patents. There are a wide range of types of documents in this category, but most are published research articles in journals and conference proceedings, technical manuals, press articles, etc., as well as patent related documents such as office actions and translations of international patents. The ideal solution would recognize the citations in this form, and even more importantly, would determine which of the NPL documents that are filed in the case these references refer to. For instance, if a line in the IDS file refers to a journal article called “*Heat Transfer Advances*” *Jrnl Heat Transf., 2001*, the ideal solution would analyze the NPL documents (which are currently present in the case as TIFF scanned files) and link the citation to the document. An even better solution would use the metadata from the journal article to fill out missing portions of the citation, converting the example citation above to something of the form of *Jones et al. “Heat Transfer Advances,” Journal of Heat Transfer, 34 (1), pp 2-30, 2001.* Attachment D, sample_npl_citations.xls shows the NPL citations for about 5000 recently issues patents (about half of published patents do not contain NPL citations).

SPEC: In the sample data, these documents refer to changes in the original specification portion of the patent application. The ideal solution would recognize important components that are noted in the XML4IP standard (see below), such as figure numbers, paragraph numbers, and references to locations within the original patent specification.

RXR.NF: This is a non-final office action and includes both forms and a prose description of the reasons for the office actions. References to claim numbers, figure numbers, document references and locations within document references are all information that should be recognized and extracted.

RXNOCP: This doc code can indicate a formal notice of a court decision and is often used to label important court documents such as Markman orders as seen in RXNOCP_06-11-2003_90006492.pdf. In these cases, the order type should be recognized and the relevant patent numbers extracted as well as the court case number(s), the plaintiffs in the case, and the court in which the proceedings are taking place.

RXLITSR: These documents contain literature searchers against a commercial database, looking for court document that have appeared and pertain to a particular patent application or granted patent. The ideal solution would extract the date of the search and identify each of the retrieved documents, for further processing by a module that would identify which are the most important.

RXAF/DR: Various legal forms pertaining to patent reexamination. Similar to RXNOCP in terms of the kind of information to be recognized and extracted.

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION
DOC-SS-PAPT-1100009**

Additional sample documents can be obtained manually on the USPTO's Public Pair website: <http://portal.uspto.gov/external/portal/pair> . To see representative documents, access Public PAIR at the url above, fill out the captcha, and then search on a patent application number. For the purposes of this example, search for application number 90006317. Next, click on the tab labeled "Image File Wrapper". The documents shown in this tab are an example of the type to be processed. Additional samples can be found in bulk at this location: <http://www.google.com/googlebooks/uspto-patents-pair.html>

These datasets contain references to files labeled with the term "non-patent literature" (NPL). These documents are not viewable from within the bulk data download nor public pair, but these documents are also of interest for this project. As mentioned above, the ideal solution would be able to extract out the citation metadata from a published document, or for those documents that are not formal publications, would pull out brief summary information that characterizes the contents of the document.

1.3 Sample XML Documents

Attachment B contains sample documents in the XML format currently used for publishing patents, called Redbook (ST 36), which is based on DTDs. Attachment C contains an XML schema representation for XML4IP (ST 96) which is the format that the Redbook documents are to be translated into. For more documentation on the Redbook standard, see: http://www.uspto.gov/products/cis/updates/patents_xml.jsp & <http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/redbook/rb2004/rb2004.html> Redbook documents can be downloaded in bulk from here: <http://www.google.com/googlebooks/uspto-patents-grants-text.html>

1.4 Questions for Response

Respondents are requested to address as many of the following questions as possible for some or all of the document types described in the previous section:

1. Method for and accuracy of schema creation step
2. Method for and accuracy of and speed of document conversion step
3. Fidelity of document layout markup
4. Degree of manual work vs. automation in each solution
5. Accuracy for different languages
6. Accuracy of converting non-textual information such as those listed below, or else of recognizing the occurrence of these and converting to inline images:
 - a. Mathematical formulae
 - b. Chemical formulae
 - c. Tables (including captions)
 - d. Figures (creating captions, linking references to figures)
7. Accuracy of conversion of documents from Redbook to XML4IP

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION
DOC-SS-PAPT-1100009**

8. Accuracy and informativeness of short summaries produced in lieu of full XML conversion, based on markup as well as text content.

For each of the above items, rough order of magnitude (ROM) pricing is requested for both a short-term, less automated step for the September deadline as well as for a longer term, more fully automated approach. The goal of the longer term approach is to develop a fully automated system that can convert all of the public and internal documentation that is currently available online with high accuracy. If this cannot be automated fully, then a long term strategy that automates as much as possible combined with a manual approach is sought. The duration of this part will depend on how fast the conversion of legacy documents can be accomplished, subject to cost constraints. Additionally, some conversion will likely be required on an ongoing basis to handle documents that enter the system in the future.

2.0 ACQUISITION STRATEGY

The USPTO acquisition strategy alternatives are still under development. The acquisition strategy will be partially dependent upon the solutions offered as a result of this RFI.

The respondent is encouraged to identify and provide any unique solutions that will result in effective/efficient operations.

2.1 Inquiries

Those who wish to submit questions concerning the RFI may do so by e-mail to the Contracting Officer, VAAnne.Tugbang@uspto.gov by Wednesday, May 18, 2011, 5:00 p.m. Eastern Standard Time (EST). Please include the RFI number in the subject line. All questions and answers will be made available via a modification to the RFI posted on the Federal Business Opportunities website (www.fbo.gov) no later than 1 week before the due date of the RFI responses. Those parties that are interested in this project will be requested to provide written responses which discuss their technical solutions and the feasibility of their approach. The written responses may include other alternatives and solutions for USPTO consideration. Written responses will not be returned and become the property of the USPTO.

2.2 Instructions for RFI Responses

Responses must be submitted electronically to the e-mail address below:

U.S. Patent and Trademark Office
Office of Procurement
ATTN: V' Anne Tugbang, Contracting Officer

EMAIL: VAAnne.Tugbang@uspto.gov

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION
DOC-SS-PAPT-1100009**

Respondents must submit their responses via email in pdf format no later than **June 1, 2011, 5 p.m. EST.** The response should be no more than 15 pages in length and no larger than 5 megabytes and should use font size 12 or larger.

Proprietary information submitted in response to this RFI will be protected from unauthorized disclosure as required by the Federal Acquisition Regulation (FAR). All proprietary markings should be clearly delineated. The respondent shall identify where data is restricted by proprietary or other rights and mark it accordingly.

The format for the RFI responses is described below:

The cover page shall contain (1) Company name, (2) Primary Point of Contact, (3) Phone Number and Email Address, (4) Cage Code, (5) NAICS Code, (6) Business Size, and (7) Federal Supply Schedule (FSS) Contract Number, if applicable.

Introduction: Provide a brief description of existing capability to perform the requirements or provide proposed Statement of Work language for the services and/or any proposed solution. In the event your company chooses to provide information subject to inclusion in a future RFP Statement of Work (SOW), clearly identify those portions and provide any appropriate authorizations for release of that portion of information within any subsequent RFP SOW issued by the USPTO, exclusive of any proprietary markings.

Technical Capability: The respondent's technical ability shall describe the services and/or any product solution(s) or dataset for the areas described in Paragraph 1.0 of this RFI. The responses should include an overall description of the proposed services and/or any product solution(s) and provide technical data and a demonstrated ability for those areas identified. The descriptions should include schedule information for delivery of services and/or product(s); and the technical rationale for providing these to the USPTO. Interested parties should provide information on their ability to use existing assets or procure, customize/configure, maintain and/or provide technical support for the resources needed for the proposed services and/or product(s). Interested parties should also describe technical benefits of their proposed services and/or product solution(s) in terms of existing technologies or resources, improvements/enhancements, cost efficiencies of their specific approach, and any other support capabilities that provide service and/or product excellence or uniqueness.

Organization Experience/Past Performance: Provide a brief description of your organization's experience in same or similar services and/or product solution(s) to both commercial and government organizations; and optionally up to three references for same or similar services and/or any product solution(s) should be provided.

Not responding to the RFI does not preclude participation in any future RFP. If a solicitation is released, it will be issued via the Federal Business Opportunities website (www.fbo.gov). It is the responsibility of the potential offerors to monitor this website for any information that may pertain to this RFI or a future RFP. The information

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION
DOC-SS-PAPT-1100009**

provided in this RFI and any future changes to the RFI are subject to change and are not binding on the USPTO.

Participation in this effort is strictly voluntary. All costs associated with responding to this RFI will be solely at the interested respondent's expense. Respondents are advised that the United States (U.S) Government will not pay for any information or administrative cost incurred in response to this RFI. The objective of this RFI is to assess vendor capabilities and interest. Review of the responses to the RFI will focus on the offeror's technical capability to provide a quality solution, corporate experience/past performance for same or similar activity with commercial activities or government agencies, and responsiveness to the RFI.

2.3 RFI Response Due Date

Please submit information via e-mail to V'Anne Tugbang, Contracting Officer, at VAnne.Tugbang@uspto.gov no later than June 1, 2011 5:00 p.m. Eastern time.

2.4 RFI Response Contact

Respondents to this RFI shall designate a primary and one alternate point of contact within the company (Name, Address, Email, and Telephone Number).

2.5 Clarification of RFI Responses

To fully comprehend the information contained within a response to this RFI, there may be a need to seek further clarification from those respondent(s) identified as capable. This clarification may be requested in the form of brief verbal communication by telephone; written communication; electronic communication; or a request for a presentation of the response to a specific USPTO group or groups. The USPTO reserves the right to seek additional information from those vendors identified with unique solutions that are determined to be beneficial to the USPTO.