

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION –
SS-PAPT12-00078**

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) will be the United States Patent & Trademark Office's (USPTO) next-generation IT infrastructure, supporting Patents business operations. The PE2E system will replace the nearly four (4) dozen aging legacy systems used today with a single system that unifies electronic processing over the entire patent application lifecycle (hence "end-to-end").

The USPTO seeks a solution that will convert documents that are currently represented as scanned TIFF images to documents represented as structured text in XML format.

As the PE2E system will operate on XML representations of all patent application documents from initial filing to publication, this RFI seeks to obtain information from interested parties, including the vendor community, about potential solutions to the problem of converting image-based patent application documents to XML format. The USPTO is exploring a wide range of possible solutions for several different but related needs relating to data conversion of patent application documents.

This RFI seeks to obtain information from interested parties, including the vendor community, about potential opportunities to find solutions for this project at three stages. (1) Initial processing with ad hoc collections to convert about 500 documents, while ramping up conversion capabilities to produce a 3-day turnaround time and validate the feasibility of item #2, which follows. (2) A path towards an automated or semi-automated system for converting all document types that are in image format to fully structured textual format. (3) When PE2E becomes operational, the intention will be for most documents submitted to the USPTO to be structured text or as text-backed documents; however, there will continue to be a need for scanning and conversion of certain documents and thus, the solution should be scalable to addresses this future need.

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION –
SS-PAPT12-00078**

1.1 Problem Scope

Document Boundary

There are approximately 800 different possible document types that can be used during the patent application process; each document type has a corresponding “doc code.” Certain document types such as the claims, specification, and abstract are included in essentially every patent application, whereas other documents types are less commonly used.

Submissions from applicants often include different document types concatenated together or one document type divided across multiple files. Consequently, documents are often assigned an incorrect doc code. These issues complicate the indexing of the image file wrapper and impair the ability of examiners to quickly find relevant information.

The solution will provide the ability to classify documents into their appropriate document types. The solution will also review previously assigned doc codes and correct or reassign doc codes as necessary. Moreover, the solution will resolve document boundary issues that include at least the following scenarios:

- Multiple document types are concatenated into one file—these documents types will need to be separated according to appropriate page boundaries, properly named, and properly classified into a document type(s) and assigned a corresponding doc code(s).
- One document type is divided across multiple files—pages from each document type will need to be rejoined into one file with proper naming, classification, and doc code.

Data Conversion

As the PE2E system will operate on XML representations of all patent application documents from initial filing to publication, the solution will provide quick and accurate delivery of text data from patent application documents to the PE2E system.

The solution has the ability to take, as input, image-based documents and automatically or semi-automatically produce XML content based on USPTO-defined schemas (see attached). Specifically, the solution will perform automated optical character recognition (OCR) and schema processing to convert image-based documents to structured text for all document types.

Each document type will require a corresponding level (e.g., high, medium, light) of editorial review and granularity (e.g., generic, granular) of tagging. Additionally, each document type will have an associated XML schema that will direct the data conversion requirements.

For document types that require the highest level of review and highest granularity of tagging, the solution will automatically or semi-automatically review OCR output to correct OCR errors and review XML content to ensure schema compliance and accuracy of tagging.

For the majority of document types, the solution will provide rendering of the converted documents that looks as similar to the source documents as possible. Lacking this, it is desirable for the locations of boundaries of structured elements (page boundaries, section boundaries, left

REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION –
SS-PAPT12-00078

or right column and line number) recorded and expressed in some manner in the representation of the converted document.

Additionally, the solution will provide the ability to search and/or highlight text in the image-based source document that corresponds to searched and/or highlighted text in the converted XML version of the document (e.g., word geometry).

Certain documents contain mathematical formulas, tables, chemical formulas, and other quasi-textual content. The solution will capture this information from the image-based source document and preserve its content and context in the converted XML version of the document.

Certain documents contain figures or drawings. The solution will capture figures as cropped images with no more than one figure per image. Additionally, the solution will capture text components of a figure, such as figure numbers, captions, or other identifying information, and tag each text component as link targets within the image file. In some cases, documents will contain tables with a combination of figures and text; the solution would use XML standards for representing such information.

Certain documents comprise USPTO-provided and applicant-submitted forms. The solution will capture and convert relevant text information for all forms. In certain forms, such as the Information Disclosure Statement (IDS), the solution will capture all citations included on each page of the IDS and provide a linking mechanism for each captured citation to enable functionality for automatically retrieving the cited document from USPTO and approved external systems. In addition, certain foreign documents that are cited in the IDS will require translation to English.

The potential solutions may vary:

- The proportion of the work that is done automatically versus manually,
- The accuracy (OCR, tagging) 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,
- 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 scanned documents,
- The language(s) of the document, and
- Recognition and handling of extraneous information, such as headers and footers.

1.2 Sample Documents for Conversion

Each document type has an associated “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

REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION –
SS-PAPT12-00078

types are concatenated into one document and associated with one code although the materials correspond to more than one code. The solution would recognize misidentified doc codes, as well as multiple disparate documents associated with one code, and resolve any document boundary issues.

The following are examples of doc codes included in Attachment 1 with an explanation of the content associated with each doc code, as well as data conversion expectations for each doc code.

CLM: This doc code represents claims, including amended claims. Processing would identify the claim numbers, the relationships between dependent and independent claims, and would recognize original claims, newly added claims, amended claims, and deleted claims. Dependent claims mention the independent claims upon which they depend. Underlining or italic font are often used to indicate new claims, and strikethrough or square brackets are often used to indicate deleted text. Footers can appear at the bottom of the page and should not be confused with the claims listing. High editorial review and granular tagging is expected to be performed for claims.

SPEC: This doc code represents the specification. The specification is the written description of the invention. The solution will recognize important components that are noted in the XML4IP standard (see below), such as figure numbers, paragraph numbers, and references to locations (column and line number) within specification. High editorial review and granular tagging is expected to be performed for the specification.

IDS: This doc code represents the Information Disclosure Statement, which is a form that contains a list of cited references including US patents, foreign patents (FOR), and non-patent literature (NPL). NPL includes publications and anything that is not a patent (US or another country). There is a wide range of NPL documents, but most are published research articles in journals, conference proceedings, technical manuals, press articles, etc., as well as patent related documents such as office actions. The solution would recognize the citations on the IDS form and create a relationship between the documented citation and the NPL, FOR or US reference for instance, if a line in the IDS refers to a journal article cited as "*Heat Transfer Advances*" *Jrnl Heat Transf., 2001*, the solution would analyze the NPL documents and link the citation to the document. The solution would use the metadata from the journal article to fill out missing portions of the citation, converting the example citation above to e.g. *Jones et al. "Heat Transfer Advances," Journal of Heat Transfer, 34 (1), pp 2-30, 2001.*

DRW: This doc code represents patent drawings filed by the applicant. Patent drawings should be captured as cropped images, with no more than one figure per image. Where there are drawing pages that contain multiple figures, each figure shall be captured in a separate image file. Text components of a figure, such as a caption, figure number, and any reference numbers or letters, shall be captured as text and tagged as link targets within the image file. References to figures or drawings in the text of the specification shall be tagged as XML resource links to their corresponding targets in the image files.

TRNA: This doc code represents a transmittal of new application indicating the contents of the submission, including any accompanying fees.

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION –
SS-PAPT12-00078**

REM: This doc code represents remarks filed by applicant. Remarks are a reply, by the applicant, to an office action that must distinctly and specifically point out the supposed errors in the examiner's action and must reply to every ground of objection and rejection in the prior Office action in order to entitle the applicant to reconsideration or further examination. Remarks are typically filed as a free-form document, usually written in letter form, by the applicant and submitted to the Office to provide arguments or remarks made when responding to an office action. The solution would identify and extract any of the following: application number(s), reexamination control number(s), attorney name, and date of the letter. The solution would also identify claim numbers, which may be represented in tabular format.

Additional sample documents can be obtained from the USPTO's Public Pair website: <http://portal.uspto.gov/external/portal/pair>. Additional samples can be found in bulk at: <http://www.google.com/googlebooks/uspto-patents-pair.html>

These datasets do not contain non-patent literature (for Copyright purposes)

1.3 Sample XML Documents

Attachment 2 contains sample documents in the XML format currently used for publishing patents, called Redbook (ST.36), which is based on DTDs. Attachment 3 contains an XML schema representation for XML4IP (ST.96) which is the format that the Redbook documents are to be translated into for use with the PE2E system. 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

In addition to the information requested below, (see instructions below, §2.2), your response to this RFI, should 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, accuracy of, and speed of document conversion step
2. Fidelity of document layout markup
3. Degree of manual work vs. automation in each solution
4. Accuracy for different languages
5. Accuracy of converting non-textual information such as those listed below, or 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)
6. Accuracy of conversion of documents from Redbook to XML4IP

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION –
SS-PAPT12-00078**

2.0 ACQUISITION STRATEGY:

The USPTO acquisition strategy alternatives are still under development. The acquisition strategy will be partially dependent upon the technical capabilities and 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:

Vendors who wish to submit questions concerning the RFI may do so by e-mail to vanne.tugbang@uspto.gov by **Tuesday, September 4, 2012, 2:00 p.m.** Eastern Standard Time. All questions and answers will be made available in a FAQ format on the Federal Business Opportunities site (www.fbo.gov) within 1 week before the due date of the RFI responses. The written responses may include other alternatives and solutions for USPTO consideration.

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
Vanne.tugbang@uspto.gov

EMAIL:

Respondents must submit their responses no later than **Tuesday, October 2, 2012, 10:00 a.m.** Eastern, in white paper/thesis format on 8.5"x11" and not more than 25 pages in length. The total page count of **25** pages does not include the cover letter, table of contents and acronym list. The total page count does include all other information provided (i.e., graphs, technical, etc.). Font size should be 12 point, Times New Roman, single sided, and prepared in Microsoft Word. Fold out charts for tables or graphics are allowed with a limited size of 11"x17" and each foldout counts as a single page and also counts towards the total page limit. Charts and graphics must be in Microsoft Excel, Microsoft Visio, or PDF format. 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 and SIN, if applicable.

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION –
SS-PAPT12-00078**

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) 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.

Corporate Experience/Past Performance: Provide a brief description of your corporate experience in same or similar services and/or product solution(s) to both commercial and government organizations; a minimum of three (3) references is requested for same or similar services and/or any product solution(s).

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 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. 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 **Tuesday, October 2, 2012, 10:00 a.m.** Eastern Standard Time.

2.4 RFI Response Contact:

**REQUEST FOR INFORMATION (RFI)
FOR USPTO's PATENT DOCUMENT CONVERSION AND XML GENERATION SOLUTION –
SS-PAPT12-00078**

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

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.