



(19) **United States**  
(12) **Patent Application Publication**  
**GARG et al.**

(10) **Pub. No.: US 2008/0162603 A1**  
(43) **Pub. Date: Jul. 3, 2008**

(54) **DOCUMENT ARCHIVING SYSTEM**

**Related U.S. Application Data**

(75) Inventors: **Ashutosh GARG**, Sunnyvale, CA (US); **Mayur DATAR**, Santa Clara, CA (US)

(63) Continuation-in-part of application No. 11/617,537, filed on Dec. 28, 2006.

**Publication Classification**

Correspondence Address:  
**HARRITY SNYDER, LLP**  
11350 Random Hills Road, SUITE 600  
FAIRFAX, VA 22030

(51) **Int. Cl.**  
**G06F 12/00** (2006.01)  
(52) **U.S. Cl.** ..... **707/204; 707/E17.083**

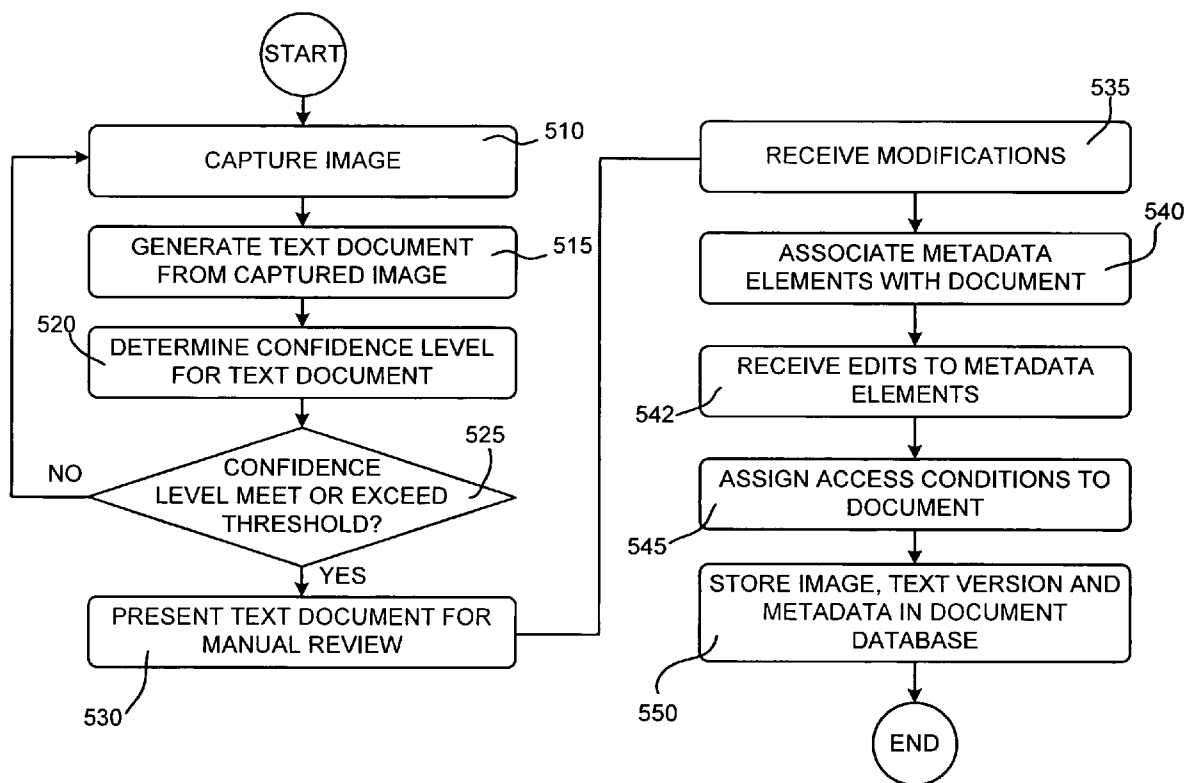
(57) **ABSTRACT**

A system generates a text document from a received document image. Metadata elements may be assigned to all or part of the text document by a user or by a template used to generate the text document. The text document and the associated metadata elements may be stored to facilitate subsequent searching and retrieval of the text document or the document image based on contents of the text document and/or its associated metadata elements.

(73) Assignee: **GOOGLE INC.**, Mountain View, CA (US)

(21) Appl. No.: **11/847,055**

(22) Filed: **Aug. 29, 2007**



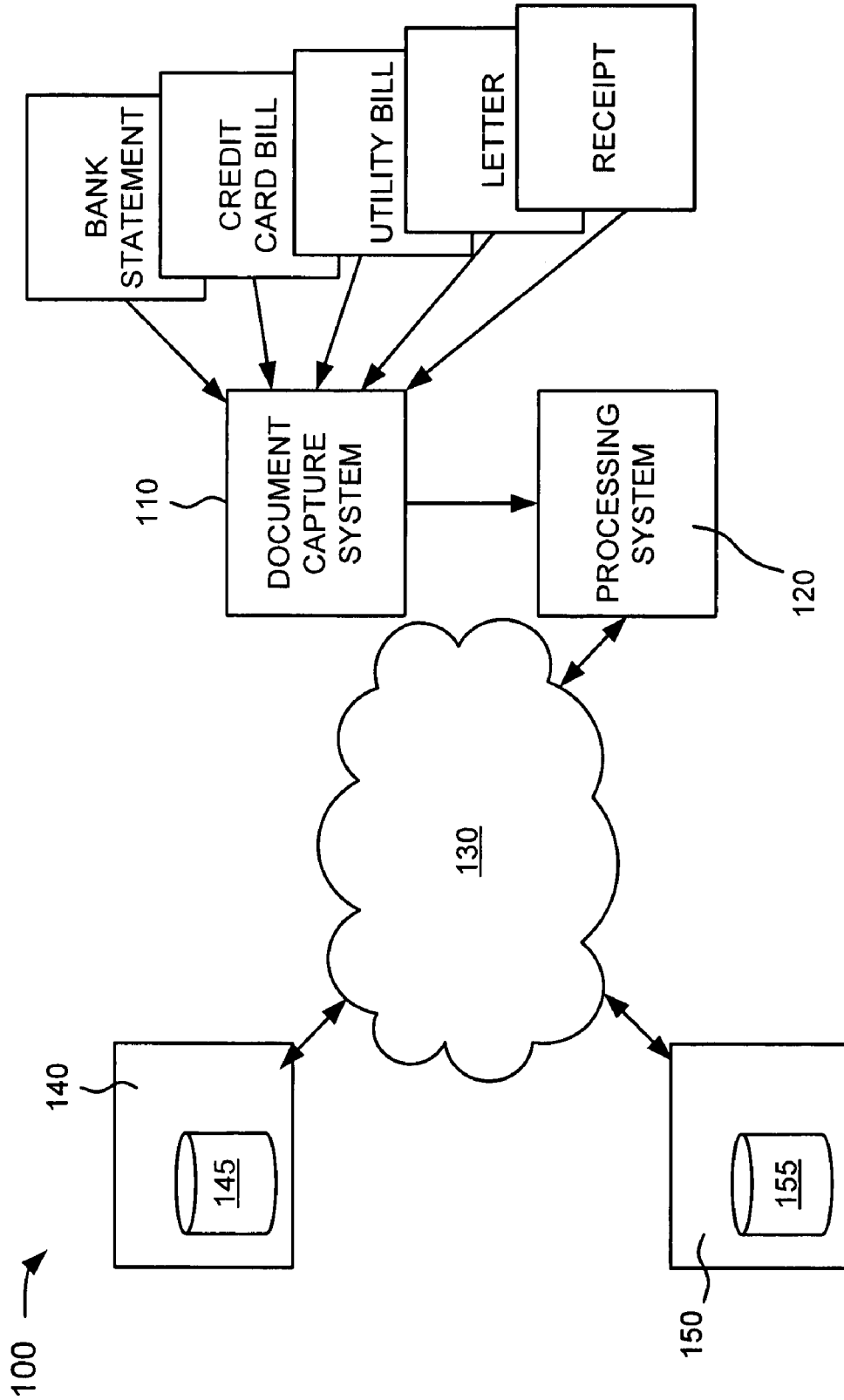


FIG. 1

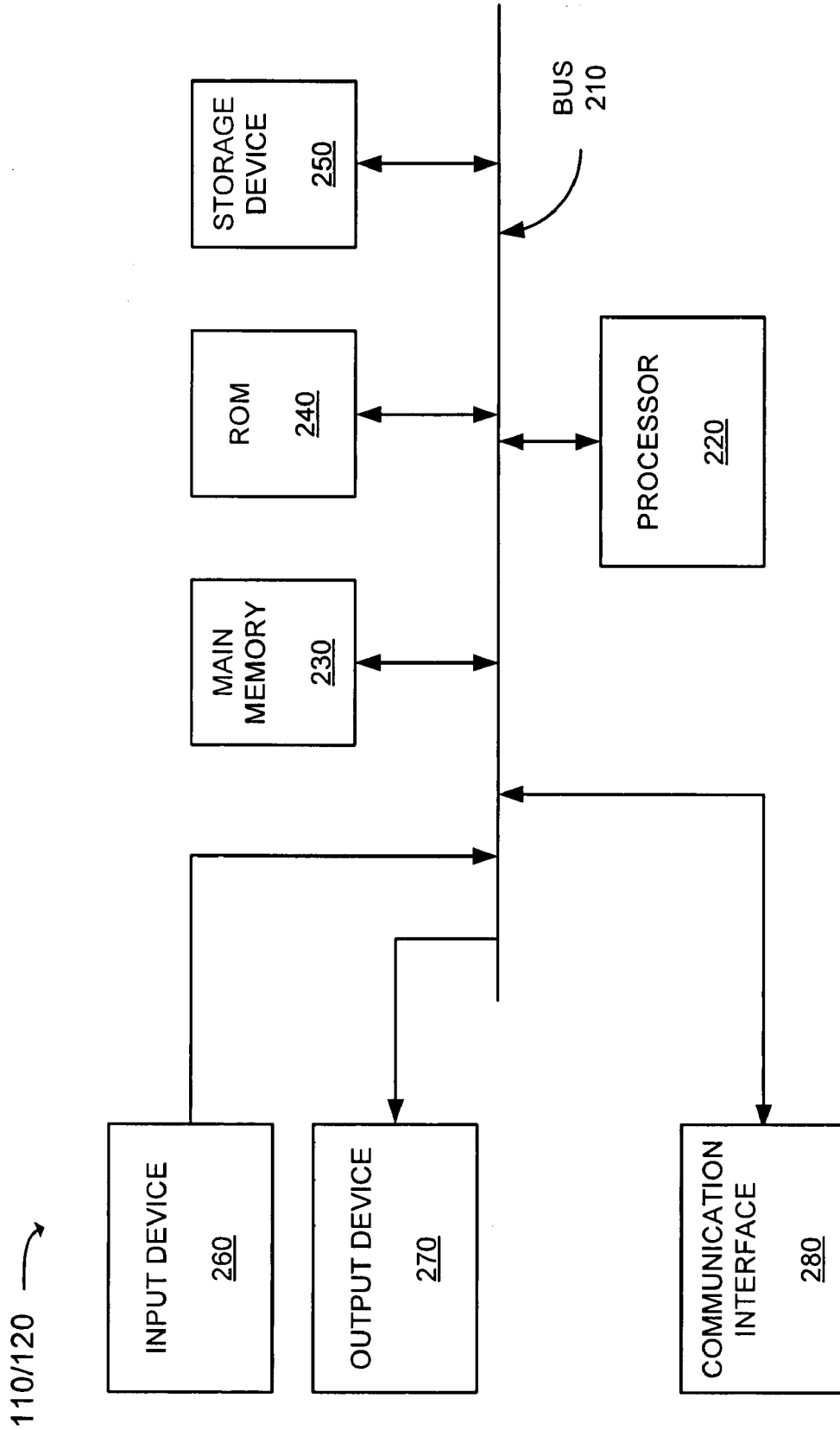


FIG. 2

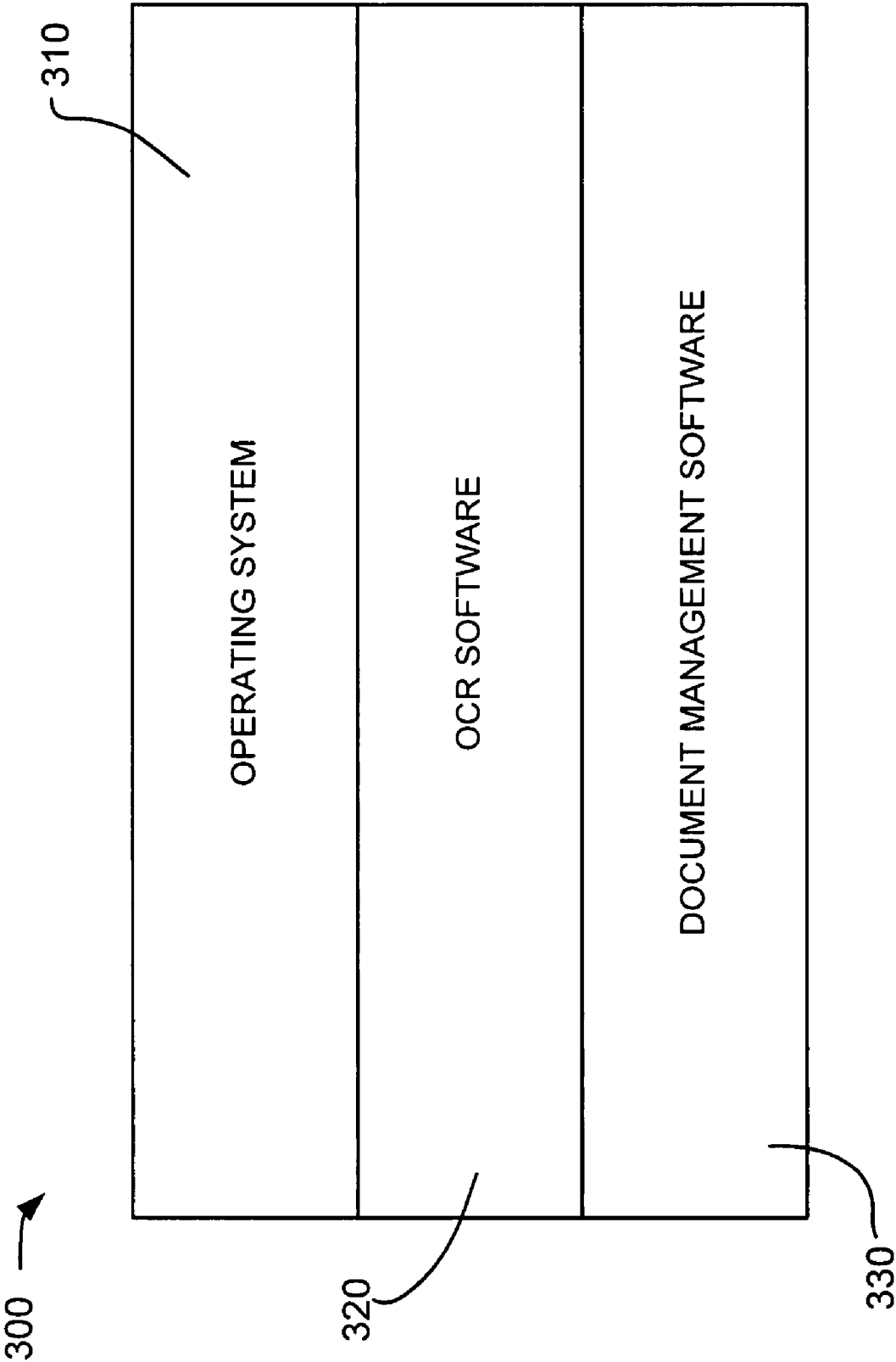


FIG. 3

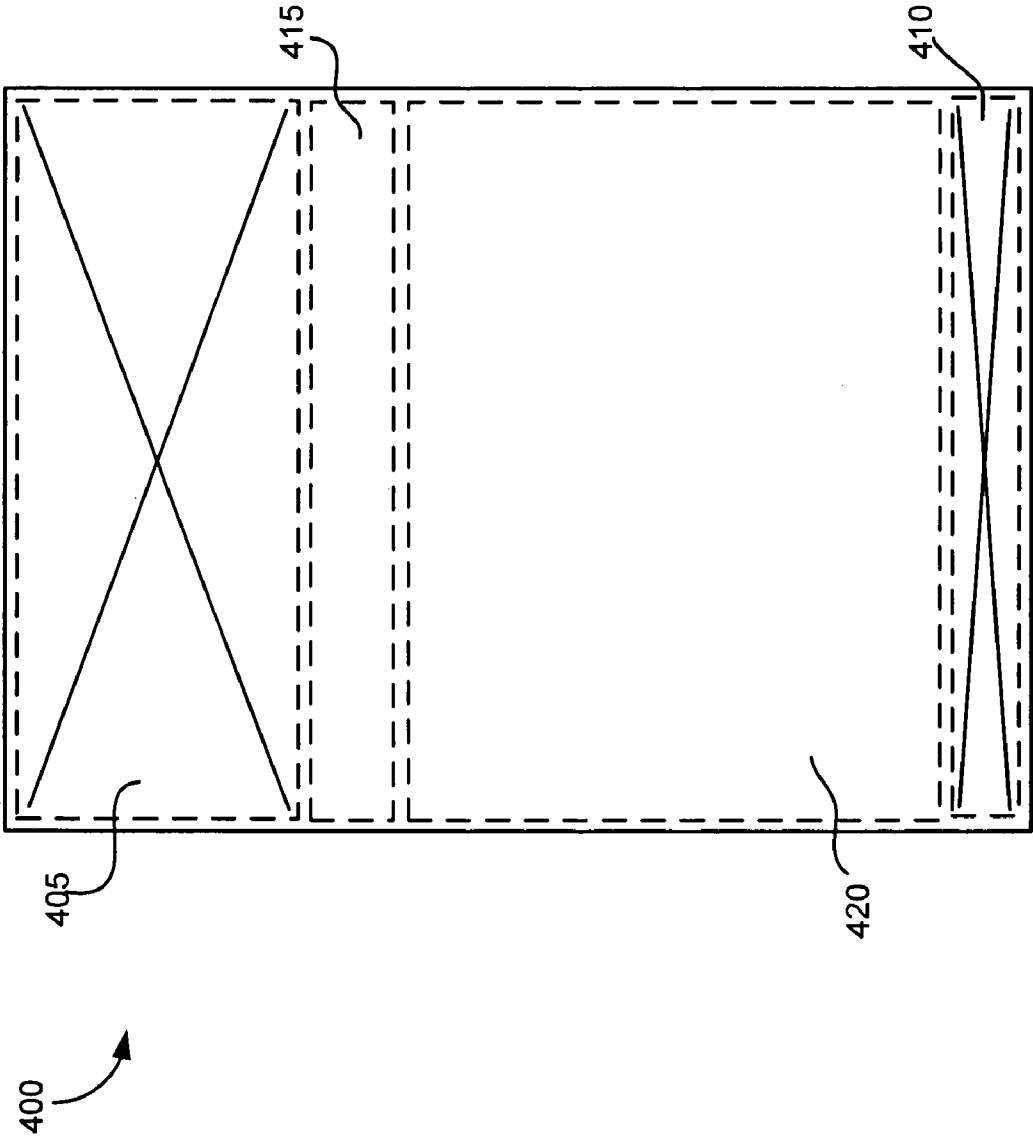


FIG. 4

FIG. 5

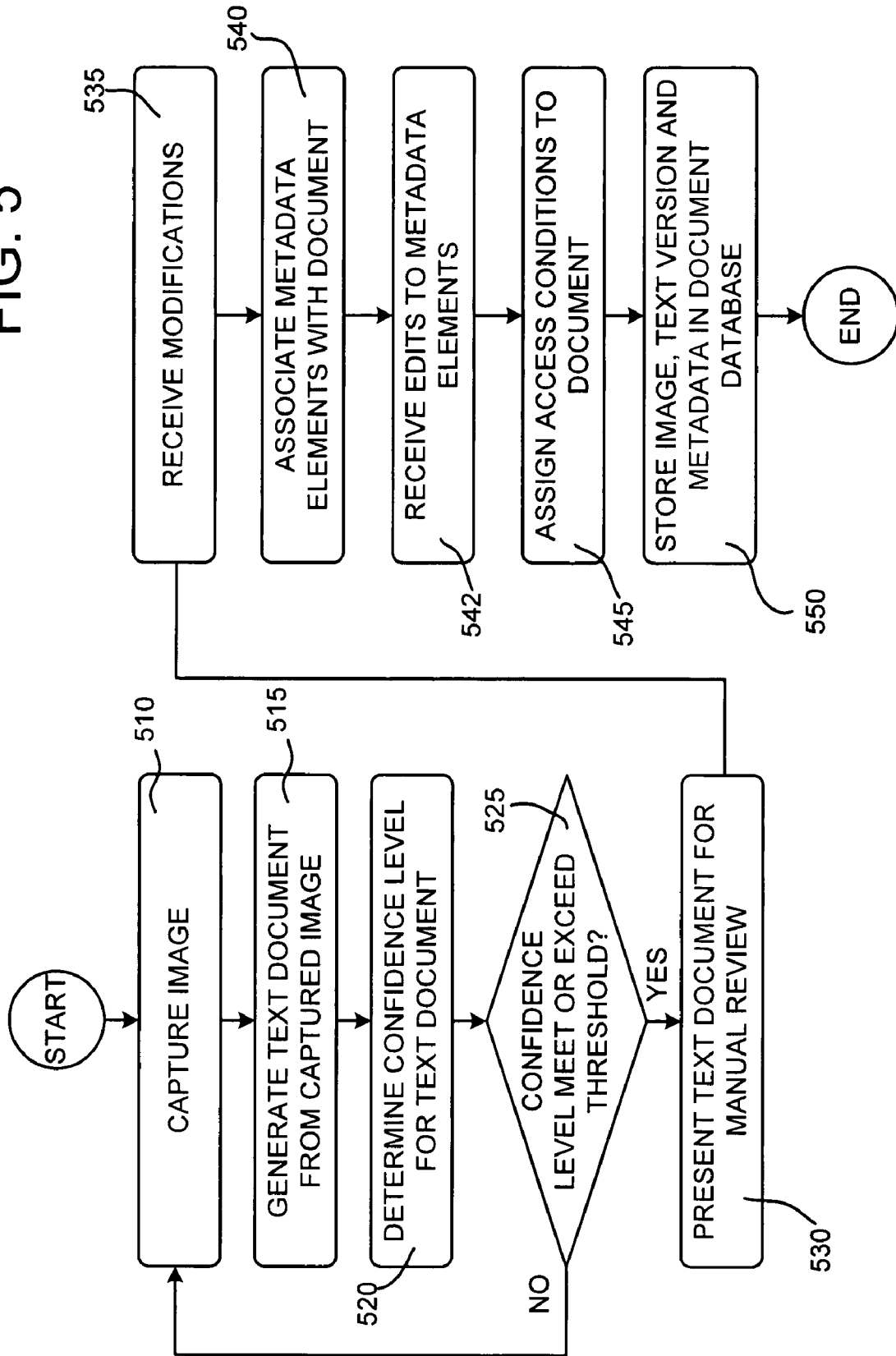
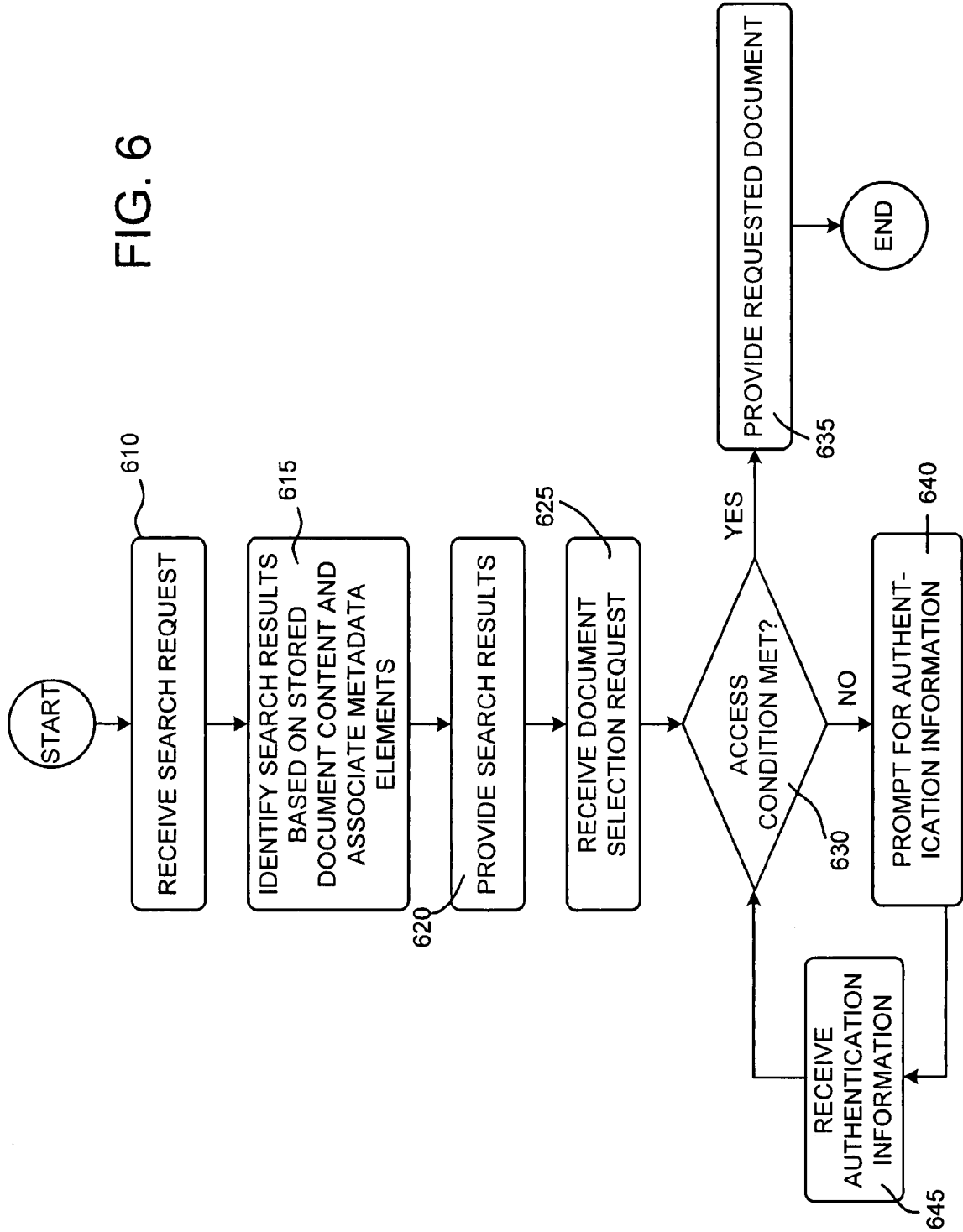


FIG. 6



**DOCUMENT ARCHIVING SYSTEM**

**RELATED APPLICATION**

[0001] This application is a continuation-in-part of U.S. patent application Ser. No. 11/617,537, filed Dec. 28, 2006, the entire content of which is incorporated by reference herein.

**BACKGROUND**

[0002] 1. Field of the Invention

[0003] Systems and methods described herein relate generally to information retrieval and, more particularly, to the archiving user information for subsequent searching and retrieval.

[0004] 2. Description of Related Art

[0005] Modern computer networks, and in particular, the Internet, have made large bodies of information widely and easily available. Internet search engines, for instance, index many millions of web documents that are linked to the Internet. A user connected to the Internet can enter a simple search query to quickly locate web documents relevant to the search query.

[0006] In addition to publicly available documents, such as websites and other online documents, recent endeavors have been made to facilitate the indexing and storing of user documents, such as word processing documents, emails, music, etc. Applications such as Google Desktop Search, Copernic Desktop Search, and Apple Computer, Inc.'s Safari typically crawl designated portions of a user's local storage and maintain an index of searchable documents identified therein. Unfortunately, conventional document indexing tools do not provide for storage or efficient indexing of non-text based documents.

**SUMMARY**

[0007] According to one aspect, a method may include receiving a document image. The document image may be converted into a text document. At least one metadata element associated with the text document may be identified. The text document and the at least one metadata element may be stored for subsequent retrieval based on the at least one searchable metadata element.

[0008] According to another aspect a system may include means for receiving a document image; means for retrieving a template that includes instructions for converting portions of the document image into a text document; means for converting the portions of the document image into the text document based on the template; means for identifying at least one metadata element associated with the text document; means for associating the at least one metadata element with the text document; and means for storing the text document and the at least one searchable metadata element for subsequent retrieval based on the at least one searchable metadata element.

[0009] According to yet another aspect a system may include a document capture system to capture an image of a document and a processor system to identify text contained within the image; generate a text document based on the identified text; identify at least one metadata element associated with the text document; and transmit the text document and the at least one metadata element to a database for subsequent retrieval based on at least one of the identified text or the at least one metadata element.

[0010] According to still another aspect, a computer-readable medium containing computer-executable instructions may be provided. The computer-readable medium may include one or more instructions for receiving a document image; one or more instructions for identifying a template associated with the document image; one or more instructions for converting the document image into a text document based on the template; one or more instructions for identifying at least one metadata element associated with the text document, wherein at least one of the at least one metadata element is automatically designated by the template; and one or more instructions for storing the text document and the at least one metadata element for subsequent retrieval based on the at least one searchable metadata element.

[0011] According to yet another aspect, a method may include receiving a document image from a scanning device; performing optical character recognition on the document image to generate a text document based on the document image; identifying at least one metadata element associated with the text document; associating the at least one searchable metadata element with at least one portion of the text document; assigning an access level indication to the text document indicating authentication information required to retrieve the text document; and storing the text document, the at least one metadata element, and the access level indication for subsequent retrieval based on a content of the text document or the at least one metadata element.

**BRIEF DESCRIPTION OF THE DRAWINGS**

[0012] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate an embodiment of the invention and, together with the description, explain the invention. In the drawings,

[0013] FIG. 1 is a diagram of an exemplary system in which systems and methods consistent with the aspects described herein may be implemented;

[0014] FIG. 2 is a diagram of an exemplary client or server entity of FIG. 1;

[0015] FIG. 3 is a diagram of a portion of an exemplary computer-readable medium that may be used by a processing system of FIG. 1;

[0016] FIG. 4 is a diagram of an exemplary optical character recognition template;

[0017] FIG. 5 is a flowchart of exemplary processing for capturing, processing and managing documents; and

[0018] FIG. 6 is a flowchart of exemplary processing for responding to document search requests.

**DETAILED DESCRIPTION**

[0019] The following detailed description of the invention refers to the accompanying drawings. The same reference numbers in different drawings may identify the same or similar elements. Also, the following detailed description does not limit the invention.

**Overview**

[0020] More and more types of documents are becoming searchable via search engines. For example, some documents, such as personal documents, financial documents, receipts, correspondence, etc. may be scanned and their text recognized via optical character recognition (OCR). Consistent with implementations described herein, it may be ben-

eficial to enable archiving and searching of these documents in an efficient and simple manner.

[0021] Systems and methods consistent with embodiments described herein may facilitate capturing or retrieval of documents and assignment of relevant metadata information to the documents. The documents may be OCR'd or otherwise processed to generate a textual version of the captured document. The document and its associated metadata and text version may be stored in an online repository or server, such that the document information may be easily searchable or retrievable by a number of devices based on information included in the text version and the associated metadata.

#### Exemplary System

[0022] FIG. 1 is a diagram of an exemplary system 100 in which systems and methods consistent with the aspects described herein may be implemented. System 100 may include a document capture system 110, a processing system 120, a network 130, a document database server 140, and a template database server 150. In one embodiment, document capture system 110 may include a scanner or similar image capturing device configured to scan a page(s) of a document. Scanner may use conventional techniques for scanning or capturing documents. In another embodiment, document capture system 110 may be configured to retrieve and/or import digital documents that may or may not include computer-readable textual information, such as web page hyper-text markup language (html) documents, screen capture images (e.g., jpeg images, png images, bmp images, gif images, etc.), portable document format (pdf), or tag image file format (tiff) formatted documents, or the like. For example, document capture system 110 may be configured to retrieve an online bank statement from a bank web server (not shown) over network 130. Such an online bank statement may be initially retrieved in an image or non-textually-recognized electronic document format (e.g., pdf, tiff, jpeg, etc.).

[0023] In one implementation, document capture system 110 may be configured to automatically and/or periodically retrieve digital documents from a remote device, such as a web server. Using the above example, document capture system 110 may be configured to retrieve electronic bank statements each month. For example, a user may configure document capture system 110 to include additional information, such as a web site address, log in information, description of a document or document type, etc., to enable document capture system 110 to access a web site and retrieve the requested document.

[0024] A "document," as the term is used herein, is to be broadly interpreted to include any machine-readable and machine-storable work product, electronic media, print media, etc. A document may include, for example, information contained in print media (e.g., newspapers, magazines, books, encyclopedias, etc.), electronic newspapers, electronic books, electronic magazines, online encyclopedias, electronic media (e.g., image files, audio files, video files, web casts, podcasts, etc.), etc.

[0025] As described in more detail below, processing system 120 may be configured to perform OCR on documents captured or otherwise retrieved by document capture system 110 to recognize text associated with the document. Processing system 120 may include a client entity, where an entity may be defined as a device, such as a personal computer, a wireless telephone, a personal digital assistant (PDA), a laptop, or another type of computation or communication device,

a thread or process running on one of these devices, and/or an object executable by one of these devices. In other aspects, processing system 120 may include a server entity that gathers, processes, searches, and/or maintains documents. In such an aspect, a "thin client" device may be configured to interact with sever-based processing system 120, where processing of documents may be performed remotely to the client device.

[0026] In one implementation, OCR processing by processing system 120 may be performed on an entirety of each captured document, with no preconfigured metadata associated therewith. In an alternative implementation, OCR processing may be based on a template or preliminary configuration that may be either automatically selected by processing system 120 or selected and/or configured by a user. Templates may assign searchable metadata elements to sections of documents or may instruct processing system 120 to OCR only predetermined portions of documents.

[0027] Using the bank statement example from above, a bank provided OCR template may instruct processing system 120 as to what portions of the statement relate to what kinds of information. For example, a first portion of statement documents may include account information, while a second portion may include transaction information. The template may further indicate that only the transaction information portion of the statement should be OCR'd. By providing information about a document in advance of OCR or other processing of the document, information capturing may be performed more efficiently. In one exemplary implementation, templates may be stored or otherwise maintained on a template database 155 of template database server 150 and may be accessible via network 130. In another embodiment (not shown), template database server 150 and/or template database 155 may be local to processing system 120. Additional details relating to the above-described implementations are set forth in detail below.

[0028] Document database server 140 may include a document database 145 configured to store the OCR'd text associated with a document as well as any metadata assigned to or associated with the captured document. In one implementation, an electronic copy of the captured image document may be stored in document database 145 as well. Alternatively, document database 145 may be configured to store the captured image document along with its associated metadata. In another implementation, document database 145 may store an index associated with the stored text and/or image documents to assist in searching and retrieving information from document database 145. The index may include a reference to the stored text and/or image documents as well as the content of the document and any metadata elements associated therewith.

[0029] As shown, in one implementation, document database server 140 may be connected to processing system 120 via network 130. However, in alternative implementations, document database server 140 and/or document database 145 may be stored locally with respect to processing system 120.

[0030] Document database server 140 may store a document's textual information and metadata information within a database record of document database 145. In one implementation, the records of document database 145 may be arranged to form a relational database, although any suitable database structure may be implemented in accordance with aspects described herein.

[0031] Network 130 may include a local area network (LAN), a wide area network (WAN), a telephone network,

such as the Public Switched Telephone Network (PSTN) or a wireless cellular network, an intranet, the Internet, or a combination of networks. Processing system 120 and database servers 140 and 150 may connect to network 130 via wired and/or wireless connections.

#### Exemplary Processing System/Scanning System Architecture

[0032] FIG. 2 is an exemplary diagram of a client or server entity (hereinafter called “system 110/120”), which may correspond to one or more of document capture system 110, processing system 120, document database server 140, and/or template database server 150. In this implementation, system 110/120 may take the form of a computer. In another implementation, system 110/120 may include a set of cooperating computers. System 110/120 may include a bus 210, a processor 220, a main memory 230, a read only memory (ROM) 240, a storage device 250, an input device 260, an output device 270, and a communication interface 280. Bus 210 may include a path that permits communication among the elements of system 110/120.

[0033] Processor 220 may include a processor, microprocessor, or processing logic that may interpret and execute instructions. Main memory 230 may include a random access memory (RAM) or another type of dynamic storage device that may store information and instructions for execution by processor 220. ROM 240 may include a ROM device or another type of static storage device that may store static information and instructions for use by processor 220. Storage device 250 may include a magnetic and/or optical recording medium and its corresponding drive.

[0034] Input device 260 may include a mechanism that permits an operator to input information to system 110/120, such as a keyboard, a mouse, a pen, voice recognition and/or biometric mechanisms, etc. Output device 270 may include a mechanism that outputs information to the operator, including a display, a printer, a speaker, etc. Communication interface 280 may include any transceiver-like mechanism that enables system 110/120 to communicate with other devices and/or systems. For example, communication interface 280 may include mechanisms for communicating with another device or system via a network, such as network 130.

[0035] As will be described in detail below, system 110/120 may perform certain document processing-related operations. System 110/120 may perform these operations in response to processor 220 executing software instructions contained in a computer-readable medium, such as memory 230. A computer-readable medium may be defined as a physical or logical memory device and/or carrier wave.

[0036] The software instructions may be read into memory 230 from another computer-readable medium, such as data storage device 250, or from another device via communication interface 280. The software instructions contained in memory 230 may cause processor 220 to perform processes that will be described later. Alternatively, hardwired circuitry may be used in place of or in combination with software instructions to implement processes in various aspects of the invention. Thus, implementations of the invention are not limited to any specific combination of hardware circuitry and software.

#### Exemplary Computer-Readable Medium

[0037] FIG. 3 is a diagram of a portion of an exemplary computer-readable medium 300 that may be used by processing system 120. In one implementation, computer-readable medium 300 may correspond to memory 230 of a processing system 120. The portion of computer-readable medium 300 illustrated in FIG. 3 may include an operating system 310, OCR software 320, and document management software 330.

[0038] More specifically, operating system 310 may include operating system software, such as the Microsoft Windows®, Unix, or Linux operating systems. OCR software 320 may include or use software (e.g., drivers) for interfacing with document capture system 110 to initiate capturing of document images by document capture system 110. Additionally, OCR software 320 may include software for converting an image of a captured document to a text version. As described briefly above, OCR software 320 may use a template retrieved from template database server 150 to facilitate efficient recognition of the document and assignment of metadata elements thereto.

[0039] FIG. 4 is an exemplary diagram of an exemplary graphical depiction of an OCR template 400 relating to the bank statement example described above. As shown, template 400 may identify several non-OCR sections 405 and 410 relating to header and footer information, which may instruct processing system 120 to not perform OCR processing on portions of the captured document relating to the locations of these sections. An account section 415 may instruct processing system 120 to assign an “account information” metadata element to any text information identified in a portion of the captured document relating to the location of section 415. Similarly, a transaction section 420 may instruct processing system 120 to assign a “transactions” metadata element to any text information identified in a portion of the captured document relating to the location of section 420. By designating OCR processing and metadata assignment for documents processed using the template, recognition and metadata assignment may be performed more efficiently than through manual implementations.

[0040] In one implementation consistent with aspects described herein, OCR software 320 may determine an OCR confidence for a converted document that indicates or otherwise determines a likelihood that a document image has been accurately converted to a text version.

[0041] In one embodiment, OCR software may initiate a rescan or recapture of a document image when the OCR confidence is below a predetermined level. In one implementation, the rescan or recapture may be performed at an increased resolution. In still a further implementation, OCR confidence may be generated for each area identified in a template, with rescan or recapture only being performed when the OCR confidence for predetermined areas are below the predetermined level. Alternatively, OCR confidence thresholds for different areas of a document may be different, depending on a relative importance of the information contained therein. This eliminates unnecessary delays caused by rescanning or recapturing data from unimportant or less important areas, while maintaining highly accurate conversions for more important areas.

[0042] Returning to FIG. 3, document management software 330 may include software for enabling a manual review of a text version of a document(s) output by OCR software 320. Document management software 330 may provide for the correction or editing of the text version, as well as the

assignment of metadata elements to one or more portions of the text version. For example, continuing with the bank statement example described above, a statement date or date range and a bank or account name may be assigned to the document. Additionally, certain portions of the document may be assigned a “debit” metadata element, while additional portions of the document may be assigned a “credit” metadata element. Document management software **330** may provide for storage of the text version, its associated metadata elements, and/or its associated document image to document database server **140** for subsequent searching and retrieval. In one implementation, document management software **330** may include an image management application such as Google® Lighthouse™ or Picasa®.

**[0043]** Assignment of metadata elements to a searchable text version of a document may facilitate more efficient retrieval of information contained in the document, using a combination of document data as well as one or more metadata elements. For example, a document including a particular transaction may be more easily retrieved in response to a user search for a specific payee in the text version as well as a date within the document’s date range and a transaction type.

#### Exemplary Processing

**[0044]** FIG. **5** is a flowchart of exemplary processing for capturing, processing and managing documents. The processing of FIG. **5** may be performed by one or more software and/or hardware components within document capture system **110** or processing system **120**, or a combination thereof. In another implementation, the processing may be performed by one or more software and/or hardware components within another device or a group of devices separate from or including document capture system **110** and/or processing system **120**.

**[0045]** Processing may begin with the document capture system **110** capturing one or more images representing a document (act **510**). As described above, one implementation may use conventional scanning techniques to capture images of the pages of the document. Alternatively, document images may be retrieved or captured from an electronic source accessible either locally or from remote resources accessible via network **130**. For example, document images may be automatically and/or periodically retrieved from a remote device, such as a web server. In one example, document capture system **110** may be configured to periodically and automatically retrieve electronic bank statements each month from a web site associated with a user’s bank. To facilitate automatic retrieval of document images, document capture system **110** may receive document location and scheduling information indicating where a document image should be retrieved from and at what times/frequency.

**[0046]** Additionally, document capture system **110** may be configured to receive authentication information associated with the retrieved document, such as a username/password combination necessary to retrieve the document. Entities such as banks, brokerage houses, financial planners, attorneys, doctors, etc. may require such authentication information prior to release of requested or retrieved documents.

**[0047]** Once captured, OCR processing may be performed on the document images to generate a textual or searchable version of the document (act **515**). OCR processing may involve an analysis of an image for recognizable text and characteristics of the text (e.g., font, size, formatting, etc.)

included therein as well as information regarding where the text is located on the pages based on the images of the pages of the document.

**[0048]** In one implementation, OCR processing may be performed on an entirety of each document image. In another implementation, OCR processing may be performed on portions of the document images based on a template retrieved from template database server **150** or, alternatively, from local storage (e.g., data storage device **250**). For example, in one implementation, a bank may provide a template from a web site hosted on server **150**. In another example, a user may configure or save a template for subsequent use with similar types of documents. As described above, templates may indicate various areas in a type of document and may be used to establish or assign metadata elements to those areas or to the document as a whole. In another implementation consistent with aspects described herein, a template may instruct OCR processing to performing recognition to a certain confidence level.

**[0049]** Once a text version of a document has been generated, a confidence level for the conversion may be determined (act **520**). It may then be determined whether the confidence level meets or exceeds a predetermined threshold level indicative of an accurate conversion (act **525**). If the predetermined threshold has not been met (act **525**—NO), the process may return to act **510** for recapture at a same or enhanced resolution. However, if the predetermined threshold has been met (act **525**—YES), the generated text version may be presented to a user for manual review and/or editing (act **530**). Any modifications, additions, or deletions to the text version may be received (act **535**). An OCR process, no matter how precise, may result in errors included in the resulting text version. Such errors may be caused by a variety of reasons, such as poor quality image capturing of the image document, artifacts present in the image document (such as, staples, holes, handwritten annotations, dirt, etc.), improperly aligned image documents, word breaks, etc. Document capture system **110** and/or processing system **120** may provide an interface for manually inspecting and editing the generated text document, to correct noted errors or inconsistencies. By providing for a manual review of the generated text version, users may efficiently correct OCR errors and may remove information from the text version that is considered sensitive or confidential. In one implementation, manually edited or corrected information may provide a feedback to document capture system **110** to facilitate subsequently enhanced OCR accuracy.

**[0050]** One or more metadata elements may be associated with or assigned to the text version of the document to facilitate enhanced searching and/or retrieval of the text version (act **540**). As described above, metadata elements may include information not explicitly present within the text of the document, but representative of the document content or selected portions of the document content. The metadata elements may be associated with either the entire text document, or to designated portions of the text document. Representative metadata elements may include content type designations, such as names, categories, date ranges, etc., customized “tags” created by a user, etc. The metadata elements may be “associated” with the text document by storing the metadata elements in document database **145** along with the text document and optionally the image document. In one implementation, the metadata elements may be embedded within or stored along with the text document and/or the image docu-

ment to facilitate subsequent searching based on the associated metadata elements as well as the converted text. Alternatively, the metadata elements associated with the text document may be stored within an index in document database 145 along with the stored text and/or image documents to assist in searching and retrieving information.

[0051] For example, using the bank statement example initially presented above, metadata elements such as “bank statement”, a document date or date range, account nickname, etc. may be assigned to the text version of the document. Additionally, metadata elements may be assigned to selected portions of the text version of the document. For example, individual credit transactions or groups of credit transactions may be assigned a “credits” metadata element, while debit transactions in the bank statement may be assigned a “debits” metadata element. In this way, information relating to the OCR’d content may be associated with the text document and used to subsequently identify and retrieve the text document.

[0052] In one implementation consistent with aspects described herein, association of metadata elements may be performed automatically upon application of a template to the image document. For example, document capture system 110 may be configured to process a selected document image based on a selected template. The selected template may instruct document capture system 110 to OCR only selected portions of the image document to a desired confidence level and assign specific metadata elements to designated portions of the resulting text document.

[0053] Once desired metadata elements have been assigned or, if automatically assigned upon application of a template, the metadata elements may be reviewed and changes to the metadata elements may be received (act 542). For example, an interface may be provided for enabling users to view any metadata elements associated with the text document. The interface may further enable users to manually edit or remove previously associated metadata elements.

[0054] One or more access conditions may be assigned to the document (act 545). In one exemplary implementation, captured documents may be assigned access conditions based on various criteria, such as document type, content, or an identity of a user initiating the capture. For example, receipts, letters, community newsletters, etc. may be assigned a first access level, while financial statements, legal correspondence, and other sensitive information may be assigned a secondary or more private access level. Alternatively, access conditions may be assigned to individual documents, upon user instruction. Searching and/or retrieval of documents assigned to the secondary access level may be available only upon receipt of authentication information indicating that the user is authorized to view content associated with the secondary access level. Suitable authentication information may include a username/password combination, a personal identification number (PIN) or code, or biometric information (e.g., voice print, finger print, retinal pattern, palm print, etc.)

[0055] In one implementation, the access level assigned to the documents may be based on the manner in which the document was captured. For example, documents obtained from predetermined sources (such as particular web sites, remote locations, etc.) may be assigned to the secure secondary access level, while all other documents may be initially assigned to a general, or less secure access level. In yet another implementation, documents obtained from secure password protected resources may be assigned an access level requiring similar password entry as the source location. For

example, searching, retrieval, or viewing of statements retrieved from a bank may require submission of authentication credentials similar to those required to retrieve the information directly from the bank. In another implementation, secure information may be searched prior to receipt of the authentication credentials; however, any identified information may not be displayed to the user prior to receipt of suitable credentials.

[0056] In another implementation, the assigned access conditions may indicate one or more individuals with whom the document may be shared or made available. For example, a first user may initiate document capture and may indicate that a second user (or group of users) may be provided access to the document text and/or document image. Additionally, the first user may indicate that the content of the captured document may be integrated within an index associated with both the first user and the second user (or group of users). For example, a husband and wife may make certain documents available to each other for subsequent identification, searching, and/or retrieval. Identification of users may be made by indicating information associated with the users, such as email addresses, usernames, etc.

[0057] Upon assignment of access conditions, the text version and its associated metadata elements may be stored in document database 145 on document database server 140 (act 550). Depending on the access conditions assigned in act 545, the stored information may be made available in response to subsequently received search queries. By providing for optimized image conversion and metadata association, stored documents may be retrieved based not only on information included within the document itself, but also based on metadata elements associated and stored with the document. For example, a store receipt may be captured and converted to text. Additionally, a metadata element, such as “John birthday gift 2006” may be associated with one or more items listed on the receipt. Subsequent searches that return this document may be based on any combination of textual information present on the receipt, such as the store name, the date, the item description, as well as any text including within a metadata element associated with the document, such as “birthday gift”, “john”, “gift”, etc.

[0058] Document searches may be performed on all documents associated with a user or group of users; however, display of a document returned in the search may only be made upon verification of a user’s access condition. Alternatively, searches may be performed only on documents to which the user has been provided access, with additional search results being available upon entry of suitable authentication information.

[0059] In one exemplary implementation, document database server 140 may be a web server configured to maintain an online storage environment for the user’s OCR’s documents. In other implementations, users may also store the captured images in document database 145, thereby enabling subsequent retrieval of the actual image document along with its text version.

Exemplary Search Processing

[0060] FIG. 6 is a flowchart of exemplary processing for responding to a document search request. The processing of FIG. 6 may be performed by one or more software and/or hardware components within document database server 140. In another implementation, the processing may be performed by one or more software and/or hardware components within

another device or a group of devices separate from or including document database server **140**, such as a web server or search engine.

**[0061]** Processing may begin with the document database server **140** receiving a document search request including one or more search terms (act **610**). In one implementation, the document search request may be received via a search engine interface, although any suitable interface may be applicable, such as an operating system or local search interface. Document database server **140** may identify search results in document database **145** that match the received search terms (act **615**). As described above, the search may be performed based on any combination of text content and metadata elements associated with the text documents stored in database **145**. Additionally, the search may be performed based on an index including the text document contents and the associated metadata elements.

**[0062]** A listing of search results may be provided (act **620**). A document selection request may be received indicating that the user wishes to view or download a selected document from the search results listing (act **625**). It may be determined whether access conditions associated with the selected document have been met (act **630**). For example, certain documents maintained in document database **145** may be restricted to authorized users. Upon selection of such a document, it may be determined whether the user requesting the document is an authorized user.

**[0063]** If the document is unprotected or if the user is an authorized user meeting the access condition associated with the selected document, the requested document may be provided to the user (act **630**). The provided document may include either the text document, or the image document, depending on user preferences or whether image documents have been saved along with the text document. Otherwise, server **140** may prompt the user to input authentication information sufficient to meet the access condition associated with the selected document (act **635**). Authentication information may be received from the user (act **635**) and the process may return to act **630** where it is again determined whether the access conditions associated with the selected document have been met.

**[0064]** In an additional implementation consistent with aspects described herein, document database server **140** may receive access authentication information from the user prior to receiving a search request. In this implementation, access conditions meeting a subset of documents in the document database may already have been received. As described briefly above, document searches in this implementation (e.g., act **615**) may be performed on a subset of documents corresponding to a user's current access privileges. In such an implementation, the provided listing of search results (e.g., act **620**) may include only documents to which the user is currently authorized to view. In such an implementation, processing may proceed directly from act **625** to act **635** for provision of the requested document.

#### CONCLUSION

**[0065]** Systems and methods described herein may automatically identify metadata associated with a document and may create an association between the metadata and the image and/or text version of the document, making both the document content and its associated metadata available for searching and/or other processing.

**[0066]** The foregoing description provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and

variations are possible in light of the above teachings or may be acquired from practice of the invention.

**[0067]** For example, while series of acts have been described with regard to FIG. **5**, the order of the acts may be modified in other implementations consistent with the principles of the invention. Further, non-dependent acts may be performed in parallel.

**[0068]** It will be apparent that aspects of the invention, as described above, may be implemented in many different forms of software, firmware, and hardware in the implementations illustrated in the figures. The actual software code or specialized control hardware used to implement aspects consistent with the principles of the invention is not limiting of the present invention. Thus, the operation and behavior of the aspects were described without reference to the specific software code—it being understood that one would be able to design software and control hardware to implement the aspects based on the description herein. No element, act, or instruction used in the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article “a” is intended to include one or more items. Where only one item is intended, the term “one” or similar language is used. Further, the phrase “based on” is intended to mean “based, at least in part, on” unless explicitly stated otherwise.

What is claimed is:

1. A method, comprising:
  - receiving a document image;
  - converting the document image into a text document;
  - identifying at least one metadata element associated with the text document; and
  - storing the text document and the at least one searchable metadata element for subsequent retrieval based on a content of the text document or the at least one metadata element.
2. The method of claim **1**, wherein receiving the document image comprises capturing the document image with an optical scanner device.
3. The method of claim **1**, wherein receiving the document image comprises receiving an electronic version of the document image.
4. The method of claim **1**, wherein receiving the document image comprises automatically retrieving an electronic version of the document image from a remote device.
5. The method of claim **1**, wherein converting the document image into the text document comprises:
  - performing optical character recognition on the document image to recognize the text of the document; and
  - generating the text document to include the recognized text of the document.
6. The method of claim **1**, further comprising:
  - retrieving a template that includes instructions for converting portions of the document image into the text document; and
  - converting the document image into the text document based on the template.
7. The method of claim **6**, wherein retrieving the template comprises retrieving the template from a template database accessible from a remote device.
8. The method of claim **6**, wherein retrieving the template is performed automatically upon receipt of the document image.
9. The method of claim **6**, wherein the template comprises information designating portions of the document image to be converted.

**10.** The method of claim **6**, wherein the template comprises information designating a minimum conversion confidence level for at least a portion of the text document.

**11.** The method of claim **10**, further comprising:  
determining a conversion confidence level for the portion of the text document;  
determining whether the conversion confidence level meets or exceeds the minimum conversion confidence level designated in the template; and  
reconverting at least a portion of the document image corresponding to the portion of the text document when it is determined that the conversion confidence level does not meet or exceed the minimum conversion confidence level.

**12.** The method of claim **11**, wherein reconverting at least a portion of the document image comprises converting the portion of the document image at an increased resolution.

**13.** The method of claim **1**, further comprising:  
retrieving a template including instructions for assigning the at least one metadata element to at least one portion of the text document corresponding to at least one portion of the document image; and  
associating the at least one metadata element to the at least one portion of the text document based on the template.

**14.** The method of claim **1**, wherein associating the at least one metadata element with the text document comprises:  
embedding the at least one metadata element within the text document.

**15.** The method of claim **1**, further comprising:  
generating an index that references the text document and includes content of the document and the at least one metadata element associated with the text document; and  
storing the index for subsequent retrieval based on the content of the text document or the at least one metadata element.

**16.** The method of claim **1**, further comprising:  
receiving instructions to modify the text document;  
modifying the text document in response to the received instructions to generate a modified text document; and  
storing the modified text document and the at least one metadata element for subsequent retrieval based on the at least one searchable metadata element.

**17.** The method of claim **16**, wherein the instructions include instructions to correct or remove at least a portion of the text document.

**18.** The method of claim **1**, comprising:  
determining a confidence level indicative of an accuracy of the text document relative to the document image; and  
recapturing the document image when it is determined that the confidence level is below a predetermined threshold.

**19.** The method of claim **18**, wherein recapturing the document image comprises recapturing the document image at an increased resolution.

**20.** The method of claim **1**, comprising:  
assigning an access level to the text document prior to storing the text document.

**21.** The method of claim **20**, wherein assigning the access level, further comprises:  
identifying the access level associated with the text document; and  
identifying authentication information required for the access level.

**22.** The method of claim **21**, wherein the authentication information includes one of a username/password combina-

tion, a personal identification number, a code, or biometric information.

**23.** The method of claim **20**, wherein the access level is based on a manner in which the document image was received.

**24.** The method of claim **20**, wherein the assigning the access level further comprises:  
receiving a designation of one or more users that are granted access to the text document.

**25.** The method of claim **24**, further comprising:  
storing the text document and the at least one metadata element for subsequent retrieval by the designated one or more users based on a content of the text document or the at least one metadata element.

**26.** A system, comprising:  
means for receiving a document image;  
means for retrieving a template that includes instructions for converting portions of the document image into a text document;  
means for converting the portions of the document image into the text document based on the template;  
means for identifying at least one metadata element associated with the text document; and  
means for indexing the text document and the at least one metadata element to facilitate subsequent retrieval of the text document or the document image.

**27.** A system, comprising:  
a document capture system to capture an image of a document; and  
a processor system to:  
identify text contained within the image;  
generate a text document based on the identified text;  
identify at least one metadata element associated with the text document; and  
transmit the text document and the at least one metadata element to a database for subsequent retrieval of the text document or the document image based on at least one of the identified text or the at least one metadata element.

**28.** The system of claim **27**, wherein the document capture system comprises an optical scanner.

**29.** The system of claim **27**, wherein the processor system is further configured to:  
automatically assign at least one initial metadata element to the text document based on a template.

**30.** The system of claim **29**, wherein the at least one initial metadata element is associated with an entirety of the text document.

**31.** The system of claim **29**, wherein the at least one initial metadata element is associated with a portion of the text document identified in the template.

**32.** The system of claim **27**, wherein the processor system is further configured to:  
assign an access level to the text document,  
wherein the access level identifies authentication information to be received prior to subsequent display of the text or image document.

**33.** The system of claim **32**, wherein the access level is associated with a group of users.

**34.** A computer-readable medium containing computer-executable instructions, comprising:  
one or more instructions for receiving a document image;  
one or more instructions for identifying a template associated with the document image;  
one or more instructions for converting the document image into a text document based on the template;

one or more instructions for identifying at least one metadata element associated with the text document, wherein at least one of the at least one metadata element is automatically designated by the template; and one or more instructions for storing the text document and the at least one metadata element for subsequent retrieval of the text document or the document image based on the at least one searchable metadata element.

**35.** A method, comprising:

receiving an image of a document from a scanning device; performing optical character recognition on the document image to generate a text document based on the document image;

identifying at least one metadata element associated with the text document;

associating the at least one searchable metadata element with at least one portion of the text document;

assigning an access level indication to the text document indicating authentication information required to retrieve the text document; and

storing the text document the at least one metadata element, and the access level indication for subsequent retrieval of the document based on a content of the text document or the at least one metadata element.

\* \* \* \* \*