CONCORD
TECHNOLOGIES

NEXTSTEP

INFORMATION INTELLIGENCE & DATA EXTRACTION

Table of Contents

Introduction and Scope

NEXTSTEP is a secure, enterprise-grade, cloud platform for converting image-based documents into text for the purpose of targeted data extraction. NEXTSTEP's HIPAA compliant capture platform is designed for organizations which need to rapidly and accurately extract data from high volumes of varying document types.

This white paper focuses on NEXTSTEP's document-conversion and data extraction capabilities.

Resilient OCR Image-to-Text Conversion

- Support for multiple image types and sources.
- No pre-sorting or batching images by type or template required.
- Automatic image pre-processing for maximum OCR accuracy.
- Multiple, dedicated OCR engines to separately handle document conversion and data extraction processing.

Dynamic Data Extraction Leveraging AI and Machine Learning

- Context and entity-driven data extraction for semi-structured documents, such as patient admission forms, records request, patient evaluation, imaging reports, and the like.
- Intelligent identification of healthcare document types enables NEXTSTEP to automatically apply the extraction profile that is most relevant for that document.
- Customizable and tunable data extraction engine based on a combination of natural language processing (NLP), rules and machine learning.

Simple UI for User-Driven Activities

- Suite of tools for users to review, process and route documents and data within any process including:
    - In-document and cross-repository text search
    - Field-level external database look-ups
    - Document re-classification
    - QA review queues
    - Document-to-field cut and paste
    - Document redaction

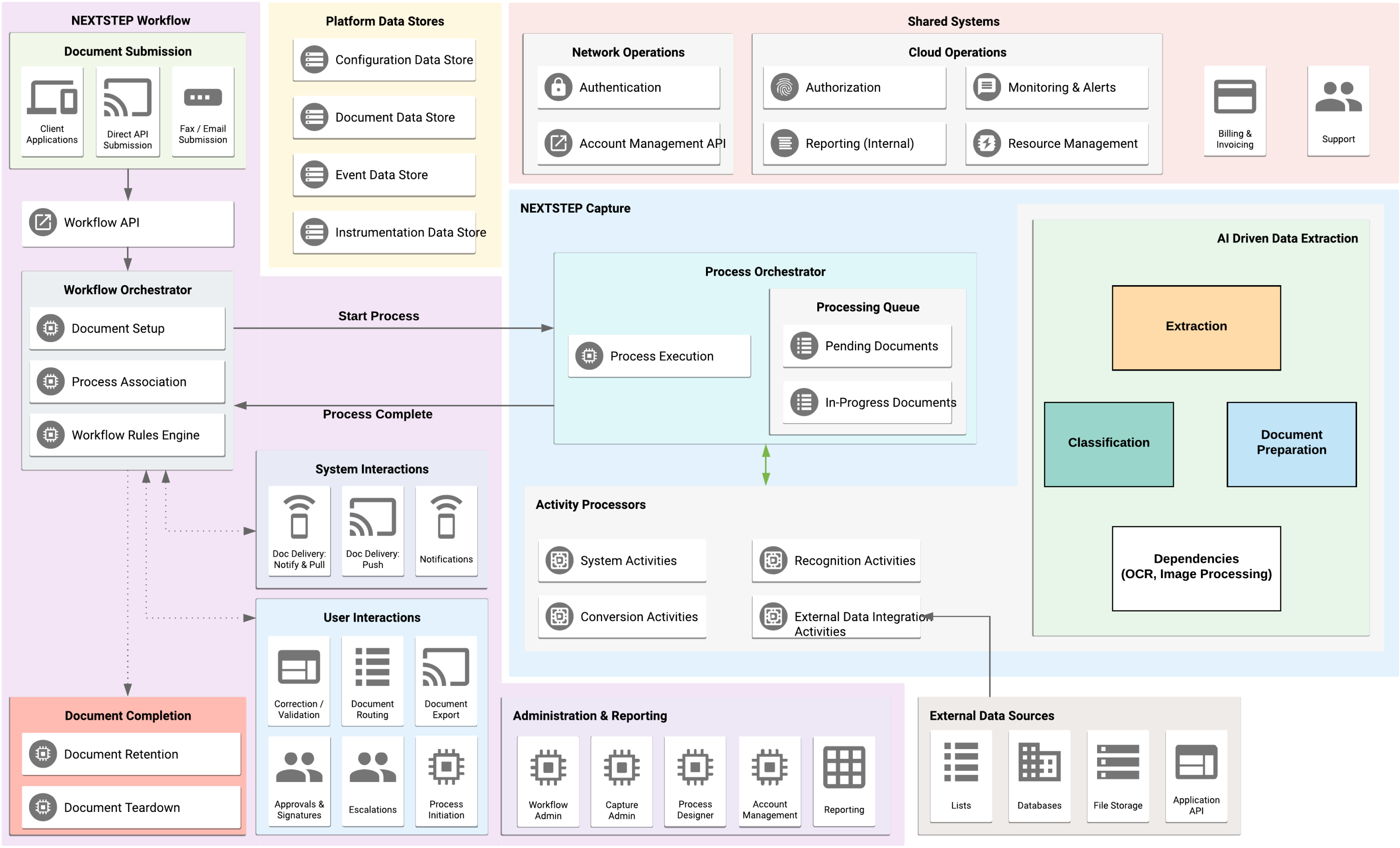
Platform Architecture

NEXTSTEP's highly-scalable design is optimized for maximum processing-bandwidth and platform reliability, critical for maintaining process velocity regardless of the volume of network traffic, or conversion demand.

The platform is deployed on Microsoft Azure, leveraging a micro-services architecture composed of smaller, independent modules.This loosely-coupled model enables Concord to manage and scale each platform component independently. Scaling occurs automatically, ensuring the platform always has more than sufficient bandwidth to process loads far-exceeding current demand. This approach ensures the platform is always ready to accommodate unforeseen spikes in content traffic.

Platform availability is achieved through component-level replication. This method offers a more granular approach to redundancy, enabling Concord to swiftly relocate any micro-services component to data centers in other geographical regions. In short, the NEXTSTEP platform continues operating in the event that an entire datacenter becomes unavailable.

NEXTSTEP Functional Architecture Diagram



The diagram above illustrates a logical flow through the system's functional components. Starting from the top-left corner, documents are submitted to the system, and progress through intake where they are then evaluated for workflow orchestration.

Workflow orchestration is not a physical component of the system, but represents the consolidated functionality that is applied to each document as it is initiated within NEXTSTEP. During this phase the document is assigned to the appropriate queue or personal inbox, and evaluated to determine if the document should be submitted for action by a defined capture process.

When a process is initiated against a document, the Process Orchestrator manages the execution of individual activities that are executed against the document. Each activity inspects the document and performs a configured action, updating the document metadata with results when applicable. Following completion of each activity, the document is handed back to the Process Orchestrator which determines the next activity to be run, or finalizes the process execution if all activities have been completed.

Capture Process

NEXTSTEP utilizes a configurable multi-stage process for image-to-text conversion and data extraction. This process is designed to optimize the quality and accuracy of extracted metadata, from any source, while enabling efficient document processing by configuring processing to include only those activities that are relevant to a particular business process.

In this way, NEXTSTEP Capture capabilities seamlessly increase the business value of every document processed. This is achieved by automatically providing relevant metadata from the document that can be utilized either within NEXTSTEP's user interface. Alternatively, the metadata can be delivered to a third party EHR or document management system within the enterprise.

Capture Process: Step-by-Step

1. **Document Intake** The submission of images to the NEXTSTEP platform, and creation of the logical NEXTSTEP document. This is a three-step process that is executed concurrently as the image pages are submitted to the platform:

    1. **Image Conversion:** Incoming files in PDF, PNG or JPEG format are converted to multipage TIFF images.
    2. **Image Pre-Processing:** The clean-up and optimization of images to normalize and improve the visual quality of the document.
    3. **Searchable PDF Creation:** Following pre-processing each page of the incoming image is converted to Searchable PDF. When all pages of the document are received, the PDF is merged to a single file and stored as a permanent part of the NEXTSTEP document.

2. **Barcode Processing:** The identification and extraction of barcode values from images. Identified values can optionally be parsed and assigned to fields on the document.
3. **Signature Detection:** Inspects a pre-defined region of the page to detect the presence of a signature. The result of inspection can optionally be assigned to a field on the document.

4. **Optical Character Recognition (OCR):** Performs extraction of text from all pages of the document image. Additional image pre-processing is executed prior to OCR to increase the efficiency and accuracy of the OCR process.

5. **Document Classification:** Automatically identifies the document type that should be associated with the document by leveraging a model that has learned from existing documents that are representative of each respective document type.

6. **Data Extraction:** Identifies and extracts target values from the text converted from document images.

7. **Document Return:** The return of extracted metadata and document files to a target application, system or service. This step is optional, and only utilized when long term storage or continued processing of documents should occur outside of the NEXTSTEP system. Document return is supported in two ways:

    1. **Document Delivery*:** Metadata and image files are pushed via HTTP to an endpoint provided by customer or third party application. Following successful delivery of document files, the document can optionally be processed by retention policies.

    2. **Document Retrieval:** Metadata and image files are published for eventual pickup by an external application. Once files are successfully picked up via the retrieval API the document can then be processed by retention policies.

8. **Document Retention:** Automated disposition of documents on an organization-defined schedule. Available retention actions allow scheduling of automatic archive, delete (soft-delete), and purge (permanent delete).

**Currently in development*

Document Intake

NEXTSTEP accepts image submission via a REST API, folder monitoring utility, scanning devices and direct integration with the Concord Cloud Fax service.

REST API

Input is of the format “multipart/form-data.” Detailed information on image submission via Concord’s web services API can be found in the NEXTSTEP API documentation. All other intake mechanisms described below rely on the REST API.

Folder Monitor

Images stored in a configured folder can be uploaded to NEXTSTEP via the Folder Monitor utility. Folders also act as the mechanism for uploading images created in, or exported from scanning devices, imaging applications, third party fax products and line-of-business applications. Multiple folders can be mapped to a single shared queue (discussed in the next section of the document).

The use of folders requires the installation of a small executable on a workstation and utilizes a Windows service to trigger the upload when new images are detected in the folder.

Concord Fax

While NEXTSTEP supports the import of faxes from any fax platform, importing directly via Concord Fax provides automatic import of all fax metadata and increased import speed by eliminating the requirement to upload fax images from end-user environment to the capture platform. Utilizing Concord’s Cloud Fax and NEXTSTEP platforms also centralizes many of the administrative, monitoring and reporting functions via a single web portal.

Supported Image Types

NEXTSTEP supports the upload of TIF, PDF, PNG and JPEG images. Should an unsupported image type be submitted, NEXTSTEP will return an error response via the API.

Barcode Recognition

NEXTSTEP’s barcode recognition engine is capable of rapidly decoding multiple barcodes on a single page, or across multiple pages. The barcode recognition engine will also read damaged, broken, and incorrect barcodes. All together NEXTSTEP can read over 30 different types of barcodes, making it suitable for virtually any customer situation.

Supported 2D Barcodes

- Aztec
- Data Matrix
- MicroPDF417
- PDF417
- QR Code

Supported 1D Barcodes

- Add-2
- Add-5
- Airline 2 of 5
- Australia Post 4-State Code
- BCD Matrix
- Codabar
- Code 128 (A,B,C)
- Code 2 of 5
- Code 32
- Code 39
- Code 39 Extended
- Code 93
- Code 93 Extended
- DataLogic 2 of 5
- EAN 128 (GS1, UCC)
- EAN-13
- EAN-8
- GS1 DataBar
- Industrial 2 of 5
- Intelligent Mail (OneCode)
- Interleaved 2 of 5
- Invert 2 of 5
- ITF-14 / SCC-14
- Matrix 2 of 5
- Patch Codes
- PostNet
- Royal Mail (RM4SCC)
- UCC 128
- UPC-A
- UPC-E
- UPU 4-State

Optical Character Recognition (OCR)

NEXTSTEP utilizes multiple discrete OCR engines for processing inbound documents. Each engine is optimized to execute specific conversion tasks. These engines are deployed independently but can be used in combination depending on customer conversion requirements. This approach enables NEXTSTEP to maximize processing velocity without sacrificing extraction accuracy, by selecting the best engine available for each task.

NEXTSTEP's PDF generation engine uses advanced OCR techniques that result in industry-leading recognition accuracy to deliver superior quality searchable PDF documents.

NEXTSTEP's Text Extraction engine leverages advanced image pre-processing prior to executing the OCR process to produce the highest quality text extraction available.

Where other platforms deliver OCR as a packaged service, the OCR capabilities in NEXTSTEP provide the foundation for advanced extraction and search feature sets that provide incremental value above and beyond the raw extracted text.

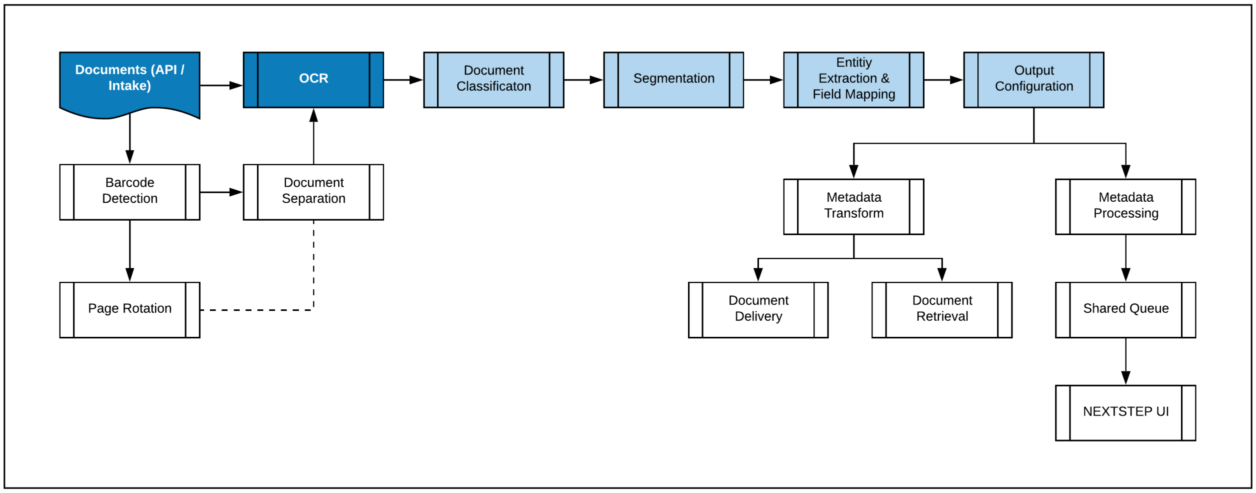
Intelligent Data Extraction

Intelligent data extraction is a process that automatically identifies and extracts key information and entities, also referred to as ‘relevant metadata’, from documents. Within healthcare forms, being able to extract patient metadata alone can be extremely valuable and lead to time and cost savings. Currently NEXTSTEP automatically extracts:

- Patient name,
- Date of Birth (DOB),
- Date of Encounter (DOE),
- Social Security Number (SSN),
- Medical Record Number (MRN),
- Patient related identifiers like patient ID, member ID, subscriber ID, account ID, etc., and
- National Provider ID (NPI)

Several approaches can be used for data or information extraction. Natural language processing (NLP) along with linguistic rules plays a key role when available data that machines can learn from is limited. The first iteration of NEXTSTEP Intelligent Data Extraction was primarily based on NLP, named entity recognition (NER), and linguistic rules. Within the past year, NEXTSTEP has embedded AI technologies, such as machine learning for internal classification, intelligent segmentation, and deep learning to enhance extraction accuracy.

AI in the capture processing pipeline



Machine Learning for internal identification of classes

Every document in healthcare contains some information pertaining to its specific document type. So, after a document is processed through OCR, the NEXTSTEP AI engine runs a machine learning model to identify the class of each page within a given document. The class of each page determines the extraction profile that is most appropriate to run for that page. For example: If page 1 of a document is patient evaluation, the extraction service will look for all relevant patient information including date of encounter (i.e., the date the patient was evaluated). If page 2 is a physician order, the extraction service knows to look for order date instead.

This healthcare-based page-wise classification capability is embedded within the AI architecture to provide more relevant and accurate extraction results for each document.

Intelligent Segmentation

A subset of the OCR component output, within the capture processing pipeline, includes a rudimentary form of segmentation. However, this segmentation is generated to support OCR processes like recognition and layout of characters, words, lines, etc. on a page and within a document. OCR segmentation does not provide any higher order indication of document structure: phrases within a sentence, word pairs with semantic meaning, word and line associations (header, footer, letter body, form table, etc.), and hierarchical document structure beyond the character.

Intelligent segmentation in NEXTSTEP expands the existing data features available to our extraction pipeline by addressing the gaps listed above. The OCR segmentation output is post-processed to restructure the document segments and provide meaningful sentences for prediction, so the resulting segments provide better and more aligned data for the platform's needs. Our segmentation approach enables the system to understand relationships between tags and values based on whether they appear in the same segment, parent segment, ancestor segment, etc. Some segments are based on NLP and observable text properties, such as the presence of a colon to designate a particular tag as a label for a value. Others form intelligence based on image-based properties like distance between 2 words. Enhanced segmentation techniques are used to understand different parts of the document, which makes for a much richer input for extraction.

Natural Language Processing (NLP)

Natural Language Processing (NLP) is the heart of our NEXTSTEP platform. Right from the initial step of data being fed into our system until the entities are being extracted, components of NLP are widely used.

In NEXTSTEP, after the OCR engine captures the image and extracts segments and characters, we use NLP to decipher them. To reconstruct the original page, we use a variety of NLP techniques in our tool kit, like merging based on characters or continuation, and splitting based on meaning or columns. We use a couple of popular techniques like POS tagging, NER and a multitude of techniques in order to annotate segments and sentences.

NLP also enables the platform to evaluate extraction rules using language context as the basis for decision making. This allows for rapid identification of target extraction values and, in certain scenarios, calculates nonexistent target values based on contextual information in the captured text. An example of a nonexistent target value is **"John Smith 31/05/94"**. NEXTSTEP can capture these values even if tags are not present.

The NLP engine breaks text up into individual parts of speech, enabling it to understand how each word is used within the text. This allows the engine to determine the context of the text so that it is able to differentiate between:

- Patient name: Theresa **May**
- Date of Birth: 09 **May** 1970
- "Patient **may** need further evaluation."

Named Entity Recognition (NER)

NER analyzes the text to find named entities such as name, date, time, location, etc. It does so by looking for text which follows patterns associated with those entities. Common patterns for name include:

- Valerie Smith: two words, both capitalized
- Valerie Henderson-Smith: three words, all capitalized, second and third word hyphenated
- Mrs. Valerie Smith: specific prefix followed by two words, both capitalized
- Valerie T Smith: two words, both capitalized, separated by single capitalized letter

Some examples of Named Entities:

- Names (John Smith, Bob Jones, Adele)
- Dates (Day, Month, Year)
- Locations (USA, London, 987 Main Street)
- Money ($1,000, 25p €50)
- Organizations (ABC Corp, XYZ Inc, 123 llc)
- Percentages (10%, 34.9%, 0.45%)
- Quantities (1, 100, 1,000,000)
- Time measurements (Minutes, Hours, Days)
- Application Specific (URL, Email Address, Social Security Number)

For entity extraction, we use a combination of robust NLP and Deep Learning approaches depending on whether the reconstructed page has segments or sentences or a combination of both. NEXTSTEP extracts fields from the both segments and sentences. NEXTSTEP's confidence level, combined with the priorities defined within the extraction pipeline, will dictate which values will be extracted.

Linguistic and Positional (LiPo) Engine

In instances where data is limited and deep learning is not an option for extraction, the engine takes advantage of the linguistic rules and positional intelligence. The engine supports the implementation of sophisticated rules which directly deal with NLP, NER and Anchor-tagging technologies. The LiPo engine has been designed to add new fields and create and process rules dynamically. It also provides the flexibility to easily edit the rules for fields manually and fine tune the extraction results. LiPo engine has evolved over a period and ensures that the extraction happens real time.

Anchor Tags analyze the position of words in relation to the document or other words in the document.

- A patient’s date of birth is located within six horizontal spaces of the word “DOB”
- A customer’s account number is located directly below the words “Customer Acc #”
- The patient’s medical record number is located X and Y distance from the top left of the document.

Deep Learning: The Heart of Artificial Intelligence

Artificial Intelligence is when the system mimics and automates human behavior and intelligence. In NEXTSTEP, our objective is to make our platform look at the documents, extract key information and present it the way humans would. To serve this, we have taken advantage of customized Deep Learning models which helps us identify the hidden patterns in large volumes of documents. These models are designed to improve performance and accuracy over time as they continuously learn and grow from more documents.

NEXTSTEP’s Deep Learning technology increases the accuracy and efficiency of the capture process without compromising on extraction velocity. This is achieved through continuous improvement in the NLP and NER engines’ abilities to identify target data values.

Deep Learning’s ability to increase the accuracy and efficiency of the capture process is highly dependent on the size of the data set it learns from. The larger the data set, the more effective Deep Learning will become. The size of the data set required to effectively “teach” the Deep Learning technology will vary significantly depending on the complexity of the source document text and the data values being targeted for extraction.

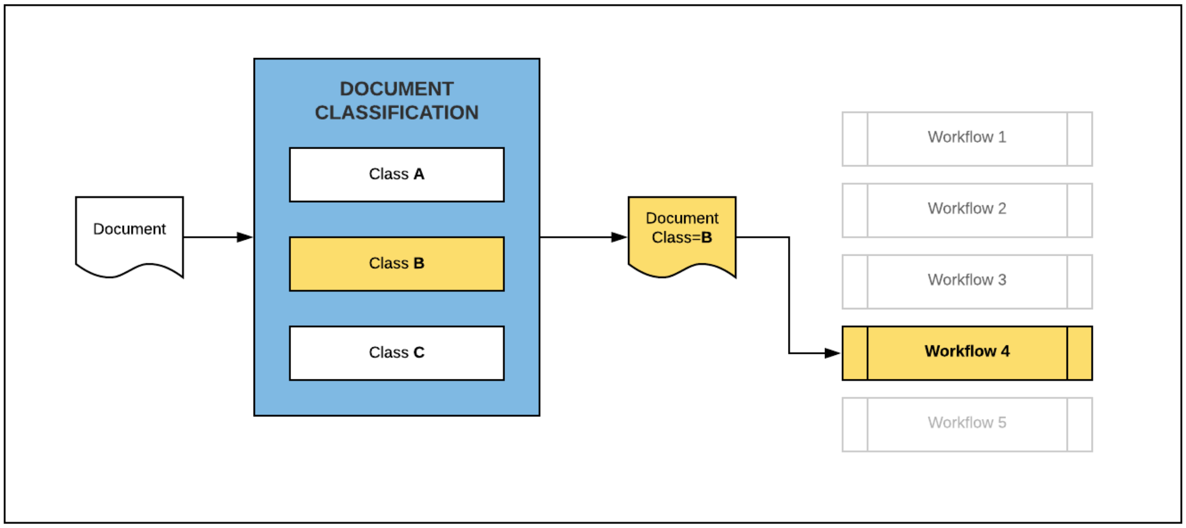
Document Classification

Document classification has long-been the primary method for assigning inbound documents to specific teams, departments or work processes. Within a healthcare provider setting, document classes might include patient visit summary, release of information request, patient referral etc. On the payer side, classes may be set-up for various types of claim forms such as the UB-O4 and CMS-1500.

For many organizations, the process of classifying documents is an entirely manual process reliant on physical documents and folders. For others, the classification is managed using a technology solution, but requires users to physically review the document on-screen in order to determine its classification. In contrast, NEXTSTEP's document classification technology is automated.

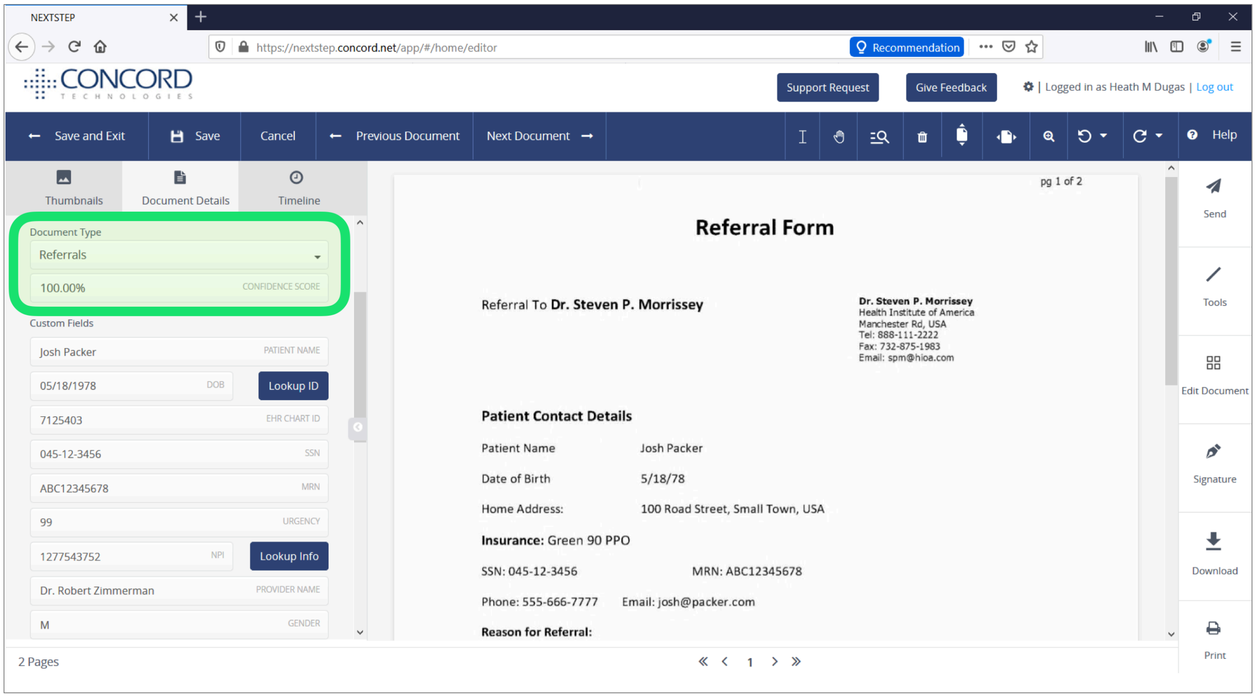
NEXTSTEP's Automated classification removes much of the user-burden from the process. Instead, NEXTSTEP deploys a variety of AI technologies to determine the class based on specific patterns or characteristics within the document. One of the primary AI technologies in document classification is machine learning.

Machine learning begins by building a learning model to identify each document class. Once the model is built, NEXTSTEP determines the class of each document by comparing the documents with the classes in the model.



As NEXTSTEP processes more documents, document classes can be retrained. Retraining rapidly increases NEXTSTEP's ability in determining the correct class. In cases where a document is mis-classified, users can simply re-classify.

By automating the classification and mapping document classes to specific workflow processes, NEXTSTEP significantly reduces the time and effort required to get documents to those who need to work with them.



Document Type: Referrals

100.00% CONFIDENCE SCORE

Custom Fields

PATIENT NAME	Josh Packer
DOB	05/18/1978
EHR CHART ID	7125403
SSN	045-12-3456
MRN	ABC12345678
URGENCY	99
NPI	1277543752
PROVIDER NAME	Dr. Robert Zimmerman
GENDER	M

Referral Form

Referral To **Dr. Steven P. Morrissey**

Dr. Steven P. Morrissey
Health Institute of America
Manchester Rd, USA
Tel: 888-111-2222
Fax: 732-875-1983
Email: spm@hioa.com

Patient Contact Details

Patient Name	Josh Packer
Date of Birth	5/18/78
Home Address:	100 Road Street, Small Town, USA

Insurance: Green 90 PPO

SSN: 045-12-3456 MRN: ABC12345678

Phone: 555-666-7777 Email: josh@packer.com

Reason for Referral:

Document Class (Displayed as "Document Type" this interface) shows NEXTSTEP classifying document as a patient referral.

Document Return

Document Delivery (HTTP PUSH)

Document Delivery executes an HTTP POST operation to a defined endpoint to effect delivery of document files created during NEXTSTEP processing. Each document is posted to the delivery endpoint independently. The files available and included in the delivery message can be configured to include only files that are relevant to a particular business process. Available output file types are shared by both the delivery and retrieval modules, and are documented below.

Document Retrieval (HTTP PULL)

Document Retrieval provides a pull-based mechanism for the integration of document files into external systems. Files created by NEXTSTEP are optionally published for retrieval upon completion of all configured processes, at which point they become visible to the retrieval API. The files available for retrieval via the API can be configured to include only files that are relevant to a particular business process, and the available output file types are shared by both the delivery and retrieval modules, as documented below.

Other Delivery Mechanisms

In addition to the API based delivery mechanisms above, NEXTSTEP also supports sending documents via email or fax. The current or original document image can be sent using either the email or fax delivery channels. Email additionally supports sending of PDF files (searchable or image-only) as well as any available metadata files.

Document Output

Plain Text

Provides a plain text representation of the document, with minimal layout formatting.

Searchable PDF

A hybrid file type that combines searchable text layered behind the original document image to retain the original appearance of the document as printed, while enabling the ability to search for text within the document.

Image-Only PDF

PDF files without searchable text can also be rendered from either the original or current document file. This format is useful when searchable content is not required and file size is a concern.

Original Document

The original source document submitted to the NEXTSTEP system, in TIFF format.

Current Document

The current document, with any modifications or annotations applied, rendered in TIFF format.

Metadata Files

Metadata created and managed by NEXTSTEP can be rendered into a single metadata file, formatted as either XML or JSON. Metadata files also support transformation to customer specified formats using pre-render transformation.

Document Retention

Document Retention in NEXTSTEP is configured as a policy by organization. Retention allows the automatic disposition of documents to be applied in accordance with time periods defined by a retention policy.

Retention actions are executed when a document passes an aging threshold that triggers a retention event. For example, a retention policy may specify:

Archive Document: 30 days (from date of of creation)

Delete Document*: 10 days after archival date

Document Purge: 90 days after deletion

* *Delete is considered a "soft-delete" similar to moving an item into the trash.*

Each of the retention periods and actions (listed above) are configurable, and default to off (no retention applied).

About Concord

At Concord Technologies, our primary mission is to simplify the way that organizations interact with their crucial documents, with a focus on those organizations in compliance-oriented industries. For over twenty years, we've been enabling businesses to simply send, receive, process and manage crucial documents using our secure, compliance-optimized cloud network.

Today, we have over a hundred thousand users in the enterprise and healthcare industries who rely on Concord every day. Concord was created to serve businesses in need of 24x7 on-demand, secure, compliant cloud fax and document management services.