Lights-Out Automation

Using AI to Create Structured Data from Static Documents

a report from

DCL™
Data Conversion Laboratory Inc.

CGI

DCL

61-18 190th Street
Suite 205
Fresh Meadows, NY 11365

+1.800.321.2816

www.dclab.com

ABOUT DCL

Intelligent
data transformations

DCL (www.dataconversionlaboratory.com) provides data and content transformation services and solutions. Using the latest innovations in artificial intelligence, including machine learning and natural language processing, DCL helps businesses organize and structure data and content for modern technologies and platforms. With expertise across many industries including publishing, life sciences, government, manufacturing, technology and professional organizations, DCL uses its advanced technology and U.S.-based project management teams to solve the most complex conversion challenges securely, accurately and on time.

Your data:
transformed, validated, enriched

CONTENTS

04 AN OVERVIEW OF THE PATENT PROCESS

05 THE CHALLENGE

06 THE QUEST FOR A FULLY AUTOMATED PROCESS

07 AUTOMATED DATA REASSEMBLY

06 THE BUSINESS VALUE TO USPTO

10 SUMMARY

An Overview of the Patent Process

Many governmental and private organizations gather massive collections of content, including legal documents, filings, and contracts. Most such collections consist of images and image-based PDFs; they're not searchable or minable for the critical information that these organizations need to function. As data collections grow larger and are measured in terabytes, conventional conversion techniques—as efficient as they may be—are not economically feasible. The Holy Grail has always been a fully automated process without human intervention. This paper describes the implementation of such a system at the United States Patent and Trademark Office (USPTO). The system is processing millions of pages each month with turnaround measured in minutes.

A patent is a grant of property rights, very much like the deed to the property on which your home sits. If you go up to the courthouse and look up your deed, you'll find that it provides a description of boundaries to your property, so you know what is yours and what is your neighbor's. A patent does the same thing, but it applies to Intellectual Property, i.e. ideas. A patent includes a description that details the idea, then a set of claims. These claims circumscribe the boundary of protection of the idea.

Examiners read these descriptions and claims, evaluate them against several laws and rules, and make decisions on patentability. Examiners have been doing this since 1790 when Thomas Jefferson became the first patent examiner. In those days, patent applications were handwritten and then published in calligraphy. They were very pretty, but not many survived due to a fire in 1836. While today we work with "electronic paper," not much else has changed.

Due to sheer volume and complexity of the patent process, it needs to change; that is what this paper addresses.

Applicants (i.e., inventors) file applications for patents with the USPTO. These filings contain a lot of data, including who the inventors are, to whom the rights are assigned, when the invention was devised, the contact information and addresses of all involved parties, whether they have any related filings, and whether they have filed in another country. All this information is in addition to the actual content of the application, which includes the specifications, claims, drawings, and an abstract of the invention.

The USPTO has a mixed process that scrapes this data, both manual and electronic, depending on how it is filed, and loads it into USPTO databases. The USPTO then adds their own data including a classification, filing dates, statuses, security screenings, and assignment to a complex work unit and examiner for review. The application content itself, the real meat of the application, is converted from whatever format the USPTO receives into G4 compressed TIFF images and loaded into an image retrieval system.

Examiners use an in-house tool to review the patent application content and begin the process of examination. The examiners read the contents from a customized image

viewer, make notes (mainly on paper), perform searches against other patents or pertinent references, evaluate the contents for compliance with numerous rules and laws, and then create a communication document that is formally transmitted to applicants.

Once an examiner allows an application, it becomes a patent. At that time, the USPTO pays to have the contents converted to extremely high-quality text that is loaded into USPTO search systems for use by examiners and disseminated to other parties, such as Google Patents and Chemical Abstracts Service. This conversion takes place as an end result of the acceptance of a patent application, and that is the essence of the problem; conversion at this stage does not help the examiners to actually examine the applications.

The Challenge

The Technical Review Process Itself is Complex and Manually Intensive

There are many operations that patent examiners have to perform that would be simpler or more efficient if they had access to the data content of the application instead of only the image of the application that is available to them at this stage.

For example, examiners have to review whether any applicants have already received a patent for the invention of this particular application. Computers can easily compare two documents and show the user the similarities or differences. However, this is not possible when the data is in an image, so examiners must perform this comparison manually.

Also, claims are cumulative. A patent application usually contains many claims, and the way they depend on each other defines different boundaries for their protection. Computers could quickly map these dependencies for examiners, if the data was not locked in images.

Scale of the Process

Transcribing 32 million pages manually is untenable. No one can afford that. Affording a portion of that is difficult to justify because merely transcription isn't functionally good enough. To be really useful, business elements within these documents should be tagged so they can be leveraged by machines in various processes.

Scale

- 2014 the USPTO received 578,802 utility applications
- USPTO maintains ~1.2 million active applications in the backlog
- Each application requires at least a specification, claim set, and abstract
  - Specification average size = ~18 pages
  - Claims set average size = ~5 pages
  - Abstract (by rule) = ~1 page
- Total pages processed in 2014 = ~32 million
  - Monthly average = ~2.6 million
  - 2015 average = 3.7 million pages/month

USPTO Requirements

The USPTO required a process that would

- Continuously receive incoming documents from USPTO in the form of G4 Compressed TIFF single page images
- Bypass OCR'ing embedded diagrams, or artifacts, but rather remove and convert them to SVG (Scalable Vector Graphics), reinserting them in the finished XML at the appropriate location.
- Perform Optical Character Recognition (OCR) on these images
- Identify characters where the OCR engine was uncertain of the accuracy of the conversion
- Tag important business elements in a custom XML vocabulary
- Return the completed documents to the USPTO within 4 hours of the document being received

Patent Documents Are NOT Simple

Patent documents include what the USPTO calls Complex Work Units (CWUs), a term for special images embedded in USPTO

content. CWUs are typically chemical and mathematical formulas, but can also include complex tables and hand-written signatures.

In addition to CWUs, patent documents can include underline, strikethrough, superscript, subscript, line numbers, headers, footers, and processing marks such as staple holes. Given the formality around patent documents, the USPTO worked with its partners to define what inputs the process could expect and what logic to create in order to deal with the various inputs.

The Quest for a Fully Automated Process

It was clear that the only way to deal with the volume that the USPTO was grappling with was an automated process. However, there were a number of barriers to deal with, not the least of them was that automated processing of this scale on such a wide variety of documents had apparently never been tried before; a search of the literature found no other such instances. The documents could be quite varied, and while there are guidelines and standard practices that patent attorneys use, they are still just guidelines and practices and are infrequently followed. The variety of potential formats was quite imposing.

There was also a psychological barrier. Document conversion as it is practiced today assumes a level of human review, and there's the fear factor of what may happen if no one looks at the documents. We had to become comfortable that processes could be robust enough to allow documents to be delivered without a human review pass. The process also had to be flexible enough not to fail in unexpected situations, but rather to identify the exceptions for later review. With the projected volume, if even every thousandth page stopped the process, we would be stopping every few minutes.

There was also a need to understand the tradeoffs, with the best of current technology; we understood that a fully automated system would not provide perfection. The question became whether it was worth getting less than absolute perfect results versus not extracting anything at all. In some cases perfection is so critical that this cannot be considered, but in the case of USPTO we were able to design appropriate controls and safeguards, to ensure integrity of the process.

One example of these tradeoffs is textual accuracy. While the pre-process described below improves OCR accuracy, OCR is simply not a perfect process. Even with the 99.5% accuracy we've been achieving, there may still be several erroneous characters on a page. While we would like perfection, for this application that accuracy was considered acceptable for searches and key functions, and then the image of the original would be retained with the XML and always be available for verification if needed. A further safeguard is that the OCR engine provides metrics on perceived accuracy which travels with the document, allowing a suspect document to be further analyzed as necessary. Other safeguards are discussed in the section on Automated Quality Analysis.

Process Overview

External to the process described in this paper, there is a facility that obtains the incoming documents, scans them to a standard format and provides them to the USPTO. It is these scanned documents which are delivered into the process we are describing.

In summary, documents entering the process are logged in, they go through preprocess to eliminate non-textual content, and they are OCR'd. The resulting text is converted to XML, recombined with the previously eliminated non-textual content, and sent back to the USPTO for loading into their system.

The devil would be in the details.

Preprocessing Documents and Preparing for OCR

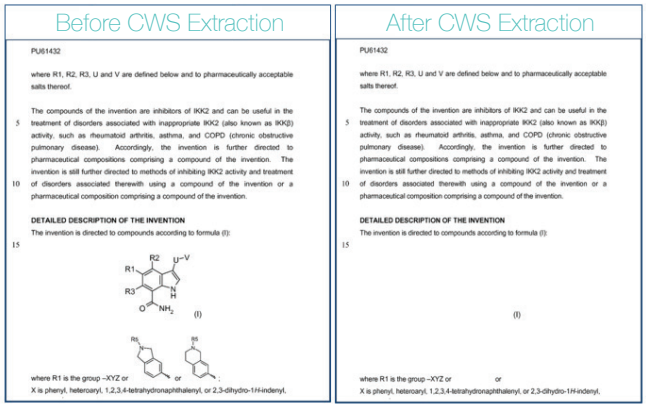
The key to an automated document conversion from images is the quality and accuracy of OCR. Using traditional OCR methods for patent documents, OCR accuracy would be degraded because of the CWUs that appear in the documents which interfere with the OCR process. Examples of CWUs are images, math, chemistry, charts, tables, etc.

Our solution was to use specially developed image analysis software to identify all such non-textual material. The CWUs were blocked off, identified as to type, size, and location, and removed digitally from the page image. The result was a page image with clear text and white space, along with separate images of each of the CWUs that had been eliminated and corresponding metadata. With pages preprocessed in this way, the OCR engine produced an accurate transcription for almost all documents.

Building a Robust Conversion Process

Building a process to handle any random document to non-trivial XML is very difficult. Knowing something about the documents however, allows for simplifying assumptions, and in most cases the reality is that we do know enough about the documents to allow for simplifying assumptions. While there is quite a bit of flexibility in how the USPTO application documents might be formatted, there are a number of attributes that appear frequently and allow us to analyze the page with quite a bit of precision. For example:

- **Line numbering and headings**—many legal documents have line numbering along the left edge, and frequently also headers and footers. These are not part of the text and if processed would wreak havoc with the conversion results. Our XML conversion software, in preprocess, was trained to remove these extraneous elements.
- **Metadata**—many documents have required elements which need to appear somewhere early in the document. While there is flexibility in how they appear, we built information into the process that by analyzing a combination of keywords, expected location, and expected format, we were able to find much of this metadata automatically.
- **Form data**—if all forms came precisely in the same format life would be easier, but they don't. There are many variations of forms that might be coming; however, there are usually identifiers somewhere on the page that allowed us to identify the form type, and allow us to mine for expected data.



We knew we would be coming across situations that we hadn't seen before. Our design goal was to make sure that even problem documents would get through without stopping the systems. Such documents were marked to indicate that they were problematic and required review, but they were not to stop the system.

Automated Data Reassembly

The XML conversion process produced valid XML to a specific USPTO DTD, but was still missing the previously removed CWUs. The final step was to recombine all the elements to create an XML document that incorporated all the CWUs as images linked appropriately into the XML in their proper locations. We are able to do this because in the preprocess step we had retained the extracted CWUs

along with their original location and sizing metadata. The OCR process we used allowed us to hold on to very detailed page geography information, and our process was able to identify the precise text location to which the image should be linked. The resulting XML document was fully tagged, with images linked to the proper location, allowing full on-the-fly document reassembly, along with all the other functionality described below.

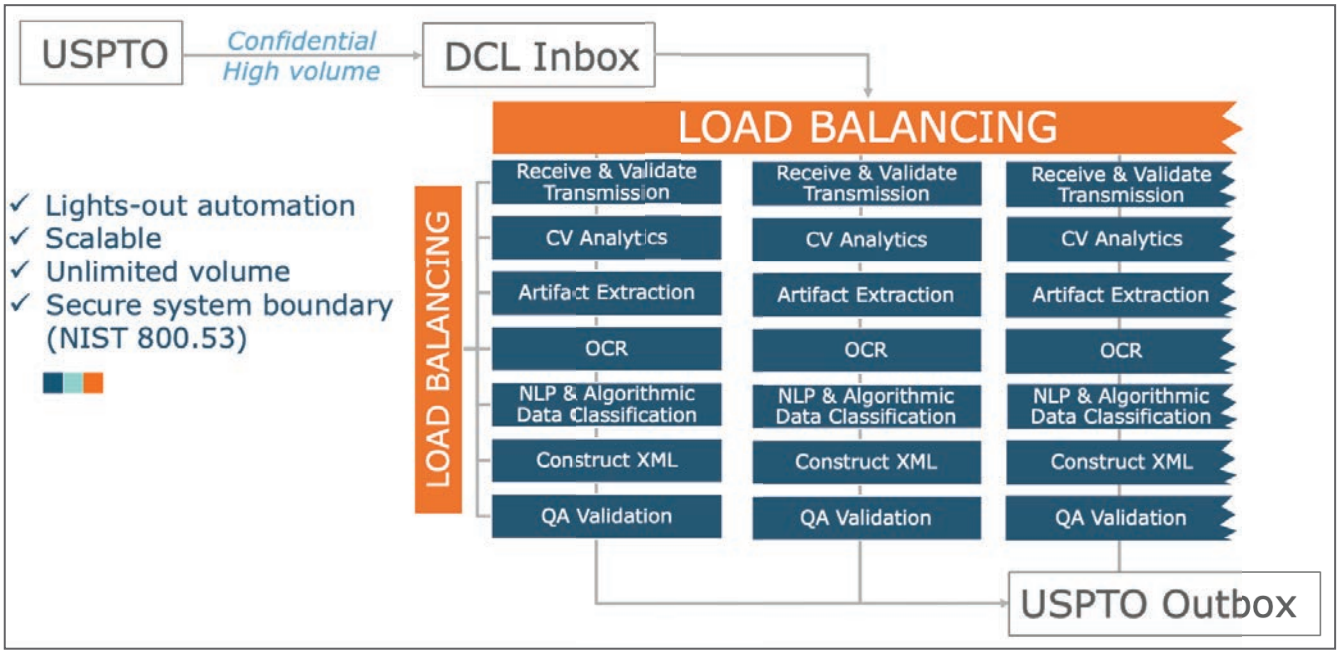
Automated Quality Analysis

The rule-of-thumb in the old days of paper application files was that 90% of filed documents will never be looked at; if only we could figure out in advance which documents they would be, we'd save a lot of file space. A similar logic applies when running a system that will process 98% of pages correctly—the quest to identify the small minority that will not be processed correctly becomes an obsession. Such was the case here. We included tools in the process that allowed us to find some of these needles in the haystack. The OCR software provided metrics on how well it thought it did, and we worked together to determine the appropriate threshold for re-review should that be necessary. Additionally, software identified unusual situations and potential issues, and recorded them into log files. The approach was for the software to complete the process of creating a valid, complete document, while identifying potential issues. We worked extensively to eliminate false positives, so as to reduce the number of documents to be reviewed to a minimum. The efforts to reduce the errors to a very small number, combined with the examiner's ability to easily review the original document when finding a questionable item, has proved to work well.

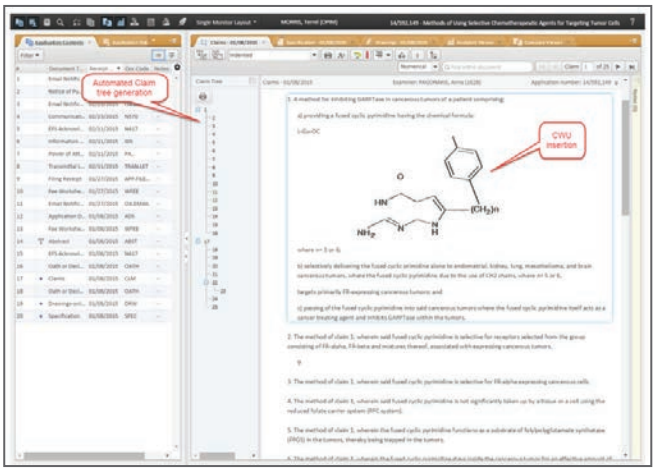
Building for Scalability – Parallel Computing Stack

Because of the extreme variability of the process we did not initially know the levels of computing capacity that would be required. We wanted adequate capacity to handle what might be coming, but at the same time we didn't want to overbuild. We also expected the need for fluctuation in processing capacity to allow for additional document types, seasonal variations in volume, and changes in the process over time. Our approach was a flexible parallel computing stack, running as an internal cloud within our secure environment.

Each of the processes described above runs on its own server. When a process completes, it is handed off to another server

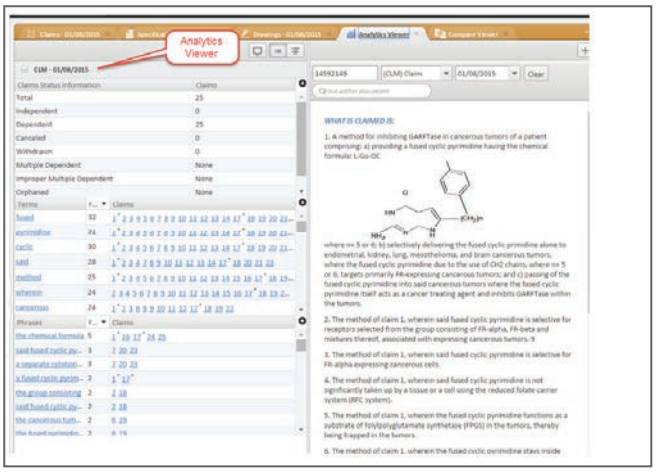


which is specialized in that next process, and so on, until the document is completely processed. Parallel processors are in place to handle volume and also eliminate the concern that a large document could hog a process leaving all other documents in queue behind it. We already had a robust Process Control Software system (DCLPCS) to track the location of each document in a process. We added load-balancing capability to invoke multiple instances of each of the processing engines, as needed, to support simultaneous processing of multiple applications. The various processes run at different speeds. For example, OCR might take more time than text conversion, requiring varying numbers of parallel processors for each process. The load-balancing software allows us to spin up additional processors based on volume and backlog. This approach has allowed for processing fluctuations in volume ranging from 500,000 pages per month to 2.5 million pages per month without strain, and it permits further expansion to ten times that volume.



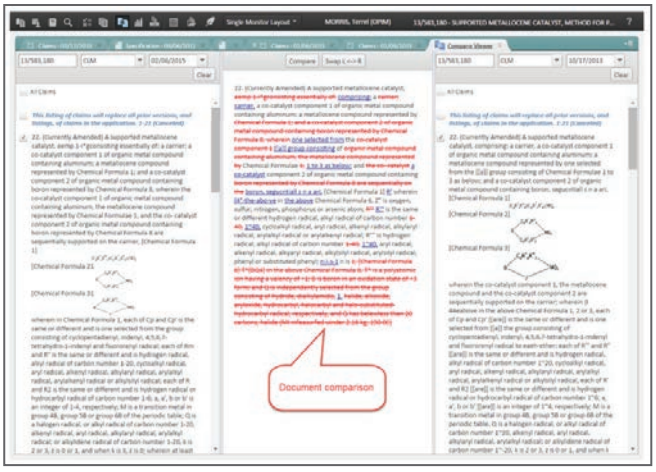
The Business Value to USPTO

We've described what we've done to produce the XML-based product we have today, which is processing millions of pages monthly. But it does not have value on its own. To justify this effort, you have to use the data. This XML was designed with the intended use of the data in mind. The following screenshots of the USPTO's new internal tools show this data being leveraged to provide functionality to examiners.



The Claim Tree

This is an application document, particularly a claim set. You can see that the XML is used to create a claim tree for an examiner. They can do this pretty easily when there are only a few claims, but for those cases with hundreds of claims, this can take a while, and it is critical to a good examination.



Also note that the body of the text from this document includes a chemical formula. The process extracted it, converted it to SVG,

and then reinserted it as properly placed in the XML. We render it for the examiner to consider during the examination. Traditional OCR would have obliterated this formula and forced the examiner to refer back to the original image.

We can also start evaluating the data of the document. We do an automated summary of the status of the claims, providing term and phrase identification, including frequency of use.

This last illustration shows that we leverage the XML to compare two documents and show the examiner the difference between them marked-up as if done by a track-changes system.

These figures only show the beginning of the functionality we envision to be powered by the data in our documents.

Summary

While government and industry have been working on getting to a paperless environment for at least four decades, and have accomplished the goal for many internal processes, when it comes to moving information among organizations, some form of paper or image presentation is still the norm. This was the battle that USPTO has been facing, as have many governmental agencies and commercial entities. This paper presents an approach that USPTO, working with DCL and CGI, is successfully using to process millions of pages that arrive as images, and converting them to rich XML which is used extensively to automate downstream processes at USPTO. While the target DTD is specific to USPTO, the approach and processes are expandable to solve similar problems for many organizations that work with large volumes of incoming documents.

DCL™

Data Conversion Laboratory Inc.

Address	61-18 190th Street Suite 205 Fresh Meadows, NY 11365
Telephone	+1 800.321.2816
Web	www.dataconversionlaboratory.com
Twitter	@dclaboratory
LinkedIn	linkedin.com/company/dclab