DCL™

Data Conversion Laboratory Inc.

Automation Techniques to Create XML from Paper and Page Images

CIDM Best Practices - Santa Fe, NM

Table Talks & Rare Birds

September 13, 2016

Mark Gross, CEO

www.dclab.com

What does Data Conversion Laboratory (DCL) do?

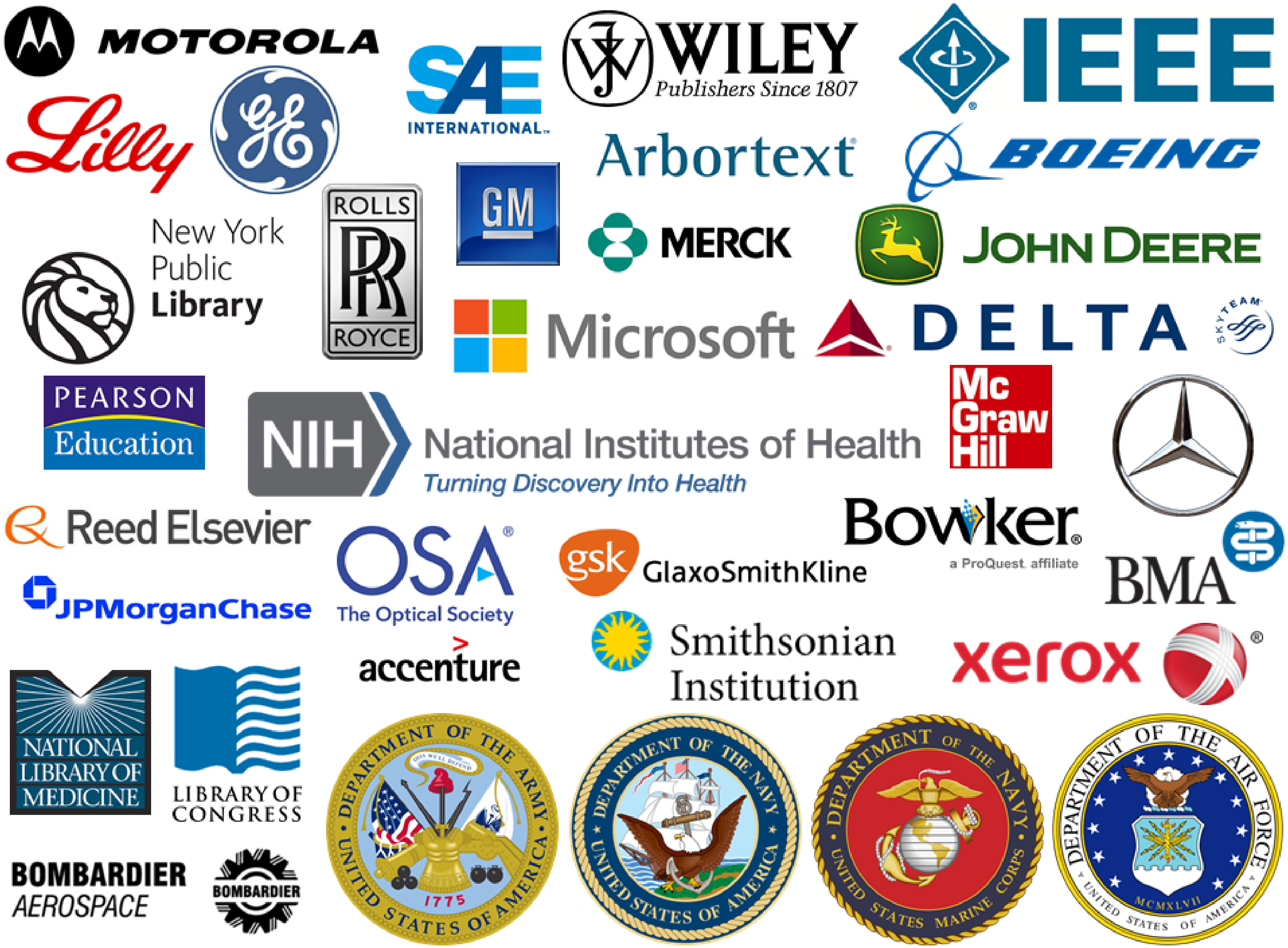
- DCL converts documents from all formats to enhanced digital formats such as XML, SGML and HTML for publishing, databases, eBooks, and distribution over the web.
- Services DCL provides:
    - Scan and digitize – we capture content from images, paper, microfilm and retrieve data with unique automated processes to greatly improve accuracy.
    - Convert – we convert data captured from any format, automatically and reliably, to XML, HTML, EPUB, and other formats needed to support new uses.
    - Enrich –we enhance data to make it more findable and usable making use of the latest data mining, metadata extraction, xml tagging and other enrichment techniques
    - Automate –we automate conversion processes to improve content reliability and reduce costs
    - Deliver Content– eBooks, data files, web-ready files, in the precise form you need, for whatever device your clients need.

People **Technology** **Expertise**

Some Background on DCL

- Established in 1981, DCL has been providing publishing and XML-related services for 35 years; we are a woman-owned small business headquartered in New York City
- An industry leading, US-based company providing expertise to capture data in all it's varied forms, and convert, normalize and publish it to meet your information needs now and in the future
- A leader in automating large-scale and complex conversion projects; named a Top 100 Companies in the Digital Content Industry in 2015 by EContent Magazine – the fourth time we have so named
- Expert in managing large-scale, multi-vendor projects; with automated tracking and reporting of data throughout
- Sophisticated quality control workflow with both automated and human quality control steps to guarantee speed, accuracy, and economy.
- Wrote the data conversion chapters in *The XML Handbook* and the *Columbia guide to Digital Publishing;* Publisher of *DCL Newsletter* devoted to Conversion and Electronic Publishing topics

We Serve a Broad Client Base …



Automating Large-Scale Paper Management

Case Study – US Patent and Trademark Office (USPTO)

Customer Problem

- Patent Application documents were stored as images. There was no content search capability.
- Patent examiners could not perform their job efficiently.
- No viable approach to convert the large volume of incoming Patent applications with complex document types to searchable XML
  - ✓ > 5,000,000 applications per year with up to 2,500 per document
  - ✓ Over 500 different form/document types containing tables, charts, figures, drawings, foreign language characters, chemical formulae, computer code, etc.

Solution

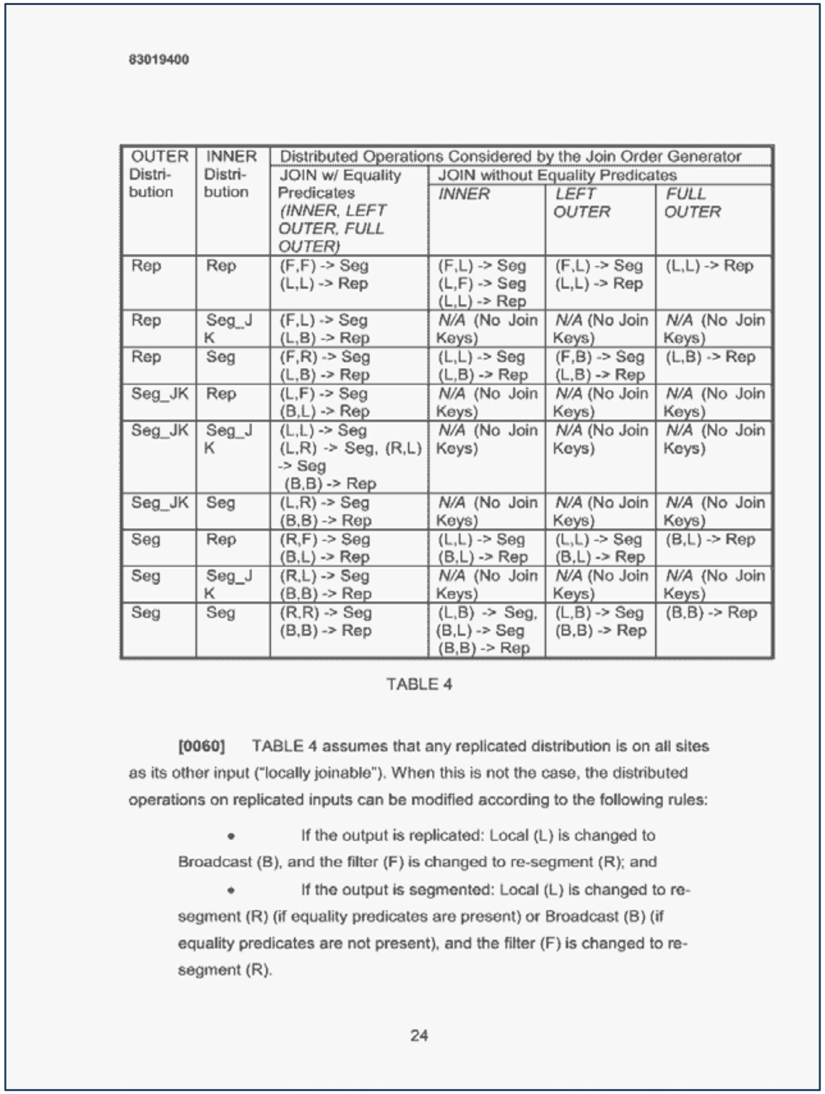
- Developed an automated conversion solution within 5 months (including agency security authorization to operate):
  - ✓ With accuracy > 99.6% and a 4 hour turnaround time
  - ✓ With current production rate of > 400,000 pages/week in a secure, fully automated, 24/7 production environment
  - ✓ With scalability to support the remaining 500+ document types and continually growing volumes

Challenges in the Quest for an Automated Process

- Software Robustness
  - Wide variety of document types and content
  - Document quality varies due to the wide audience of patent filers
- Process Flexibility and Scalability
  - Wide variances in volume – have done as many as 180,000 pages in a single day
  - How to keep a 5000 page document from clogging the works
- The Psychological Barrier – “Perfect” vs. “Good Enough”
- Identifying and Categorizing Documents that Fail (automated quality review)
  - Comprehensive metrics and reconciliation requirements
  - Confirmation messages to address all possible scenarios
- Continuous Feedback and Refinement

USPTO Sample Documents: Specifications

Contains table layouts and complex math



83019400

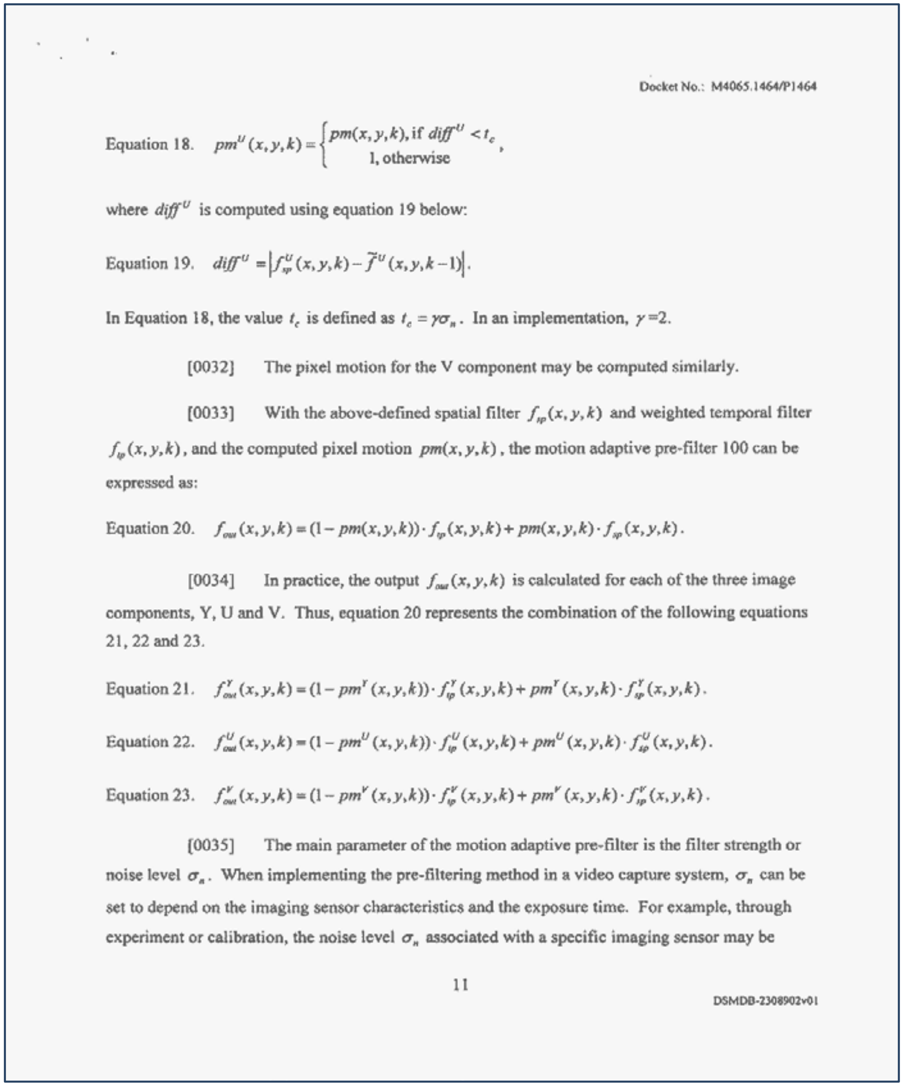
OUTER Distribution	INNER Distribution	Distributed Operations Considered by the Join Order Generator			
		JOIN w/ Equality Predicates *(INNER, LEFT OUTER, FULL OUTER)*	JOIN without Equality Predicates		
			INNER	*LEFT OUTER*	*FULL OUTER*
Rep	Rep	(F,F) -> Seg (L,L) -> Rep	(F,L) -> Seg (L,F) -> Seg (L,L) -> Rep	(F,L) -> Seg (L,L) -> Rep	(L,L) -> Rep
Rep	Seg_JK	(F,L) -> Seg (L,B) -> Rep	*N/A* (No Join Keys)	*N/A* (No Join Keys)	*N/A* (No Join Keys)
Rep	Seg	(F,R) -> Seg (L,B) -> Rep	(L,L) -> Seg (L,B) -> Rep	(F,B) -> Seg (L,B) -> Rep	(L,B) -> Rep
Seg_JK	Rep	(L,F) -> Seg (B,L) -> Rep	*N/A* (No Join Keys)	*N/A* (No Join Keys)	*N/A* (No Join Keys)
Seg_JK	Seg_JK	(L,L) -> Seg (L,R) -> Seg, (R,L) -> Seg (B,B) -> Rep	*N/A* (No Join Keys)	*N/A* (No Join Keys)	*N/A* (No Join Keys)
Seg_JK	Seg	(L,R) -> Seg (B,B) -> Rep	*N/A* (No Join Keys)	*N/A* (No Join Keys)	*N/A* (No Join Keys)
Seg	Rep	(R,F) -> Seg (B,L) -> Rep	(L,L) -> Seg (B,L) -> Rep	(L,L) -> Seg (B,L) -> Rep	(B,L) -> Rep
Seg	Seg_JK	(R,L) -> Seg (B,B) -> Rep	*N/A* (No Join Keys)	*N/A* (No Join Keys)	*N/A* (No Join Keys)
Seg	Seg	(R,R) -> Seg (B,B) -> Rep	(L,B) -> Seg, (B,L) -> Seg (B,B) -> Rep	(L,B) -> Seg (B,B) -> Rep	(B,B) -> Rep

TABLE 4

[0060] TABLE 4 assumes that any replicated distribution is on all sites as its other input ("locally joinable"). When this is not the case, the distributed operations on replicated inputs can be modified according to the following rules:

- If the output is replicated: Local (L) is changed to Broadcast (B), and the filter (F) is changed to re-segment (R); and
- If the output is segmented: Local (L) is changed to re-segment (R) (if equality predicates are present) or Broadcast (B) (if equality predicates are not present), and the filter (F) is changed to re-segment (R).

24



Docket No.: M4065.1464/P1464

Equation 18. $pm^U(x,y,k) = \begin{cases} pm(x,y,k), \text{if } diff^U < t_c \\ 1, \text{otherwise} \end{cases},$

where $diff^U$ is computed using equation 19 below:

Equation 19. $diff^U = \left| f_{sp}^U(x,y,k) - \tilde{f}^U(x,y,k-1) \right|.$

In Equation 18, the value t_c is defined as $t_c = \gamma\sigma_n$. In an implementation, γ =2.

[0032] The pixel motion for the V component may be computed similarly.

[0033] With the above-defined spatial filter $f_{sp}(x,y,k)$ and weighted temporal filter $f_{tp}(x,y,k)$, and the computed pixel motion $pm(x,y,k)$, the motion adaptive pre-filter 100 can be expressed as:

Equation 20. $f_{out}(x,y,k) = (1 - pm(x,y,k)) \cdot f_{tp}(x,y,k) + pm(x,y,k) \cdot f_{sp}(x,y,k).$

[0034] In practice, the output $f_{out}(x,y,k)$ is calculated for each of the three image components, Y, U and V. Thus, equation 20 represents the combination of the following equations 21, 22 and 23.

Equation 21. $f_{out}^Y(x,y,k) = (1 - pm^Y(x,y,k)) \cdot f_{tp}^Y(x,y,k) + pm^Y(x,y,k) \cdot f_{sp}^Y(x,y,k).$

Equation 22. $f_{out}^U(x,y,k) = (1 - pm^U(x,y,k)) \cdot f_{tp}^U(x,y,k) + pm^U(x,y,k) \cdot f_{sp}^U(x,y,k).$

Equation 23. $f_{out}^V(x,y,k) = (1 - pm^V(x,y,k)) \cdot f_{tp}^V(x,y,k) + pm^V(x,y,k) \cdot f_{sp}^V(x,y,k).$

[0035] The main parameter of the motion adaptive pre-filter is the filter strength or noise level σ_n. When implementing the pre-filtering method in a video capture system, σ_n can be set to depend on the imaging sensor characteristics and the exposure time. For example, through experiment or calibration, the noise level σ_n associated with a specific imaging sensor may be

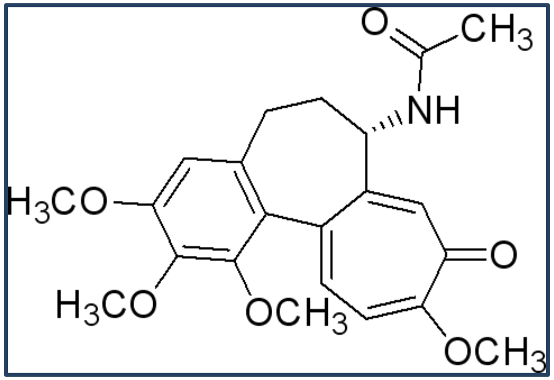
11

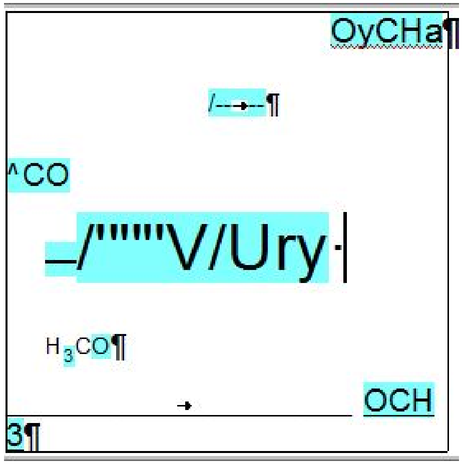
DSMDB-2308902v01

Why Off-the-Shelf OCR Tools Don’t Work

Sample Document Text

OCR Output





OyCHa¶

/---→---¶

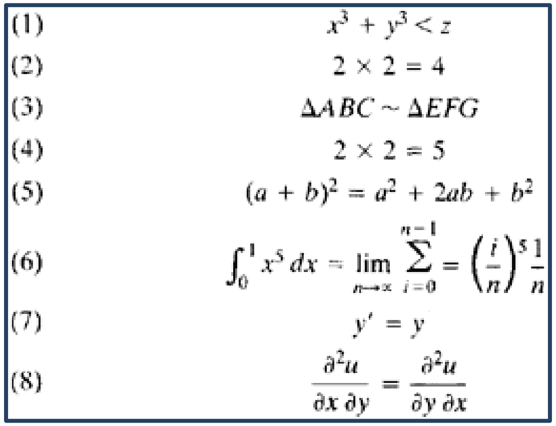
^CO

_/""""V/Ury·

H_3CO¶

→ OCH

3¶



(1) $$x^3 + y^3 < z$$

(2) $$2 \times 2 = 4$$

(3) $$\Delta ABC \sim \Delta EFG$$

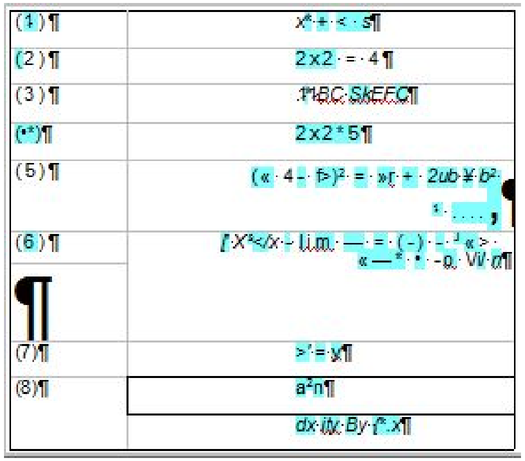
(4) $$2 \times 2 = 5$$

(5) $$(a + b)^2 = a^2 + 2ab + b^2$$

(6) $$\int_0^1 x^5\,dx = \lim_{n\to\infty} \sum_{i=0}^{n-1} = \left(\frac{i}{n}\right)^5 \frac{1}{n}$$

(7) $$y' = y$$

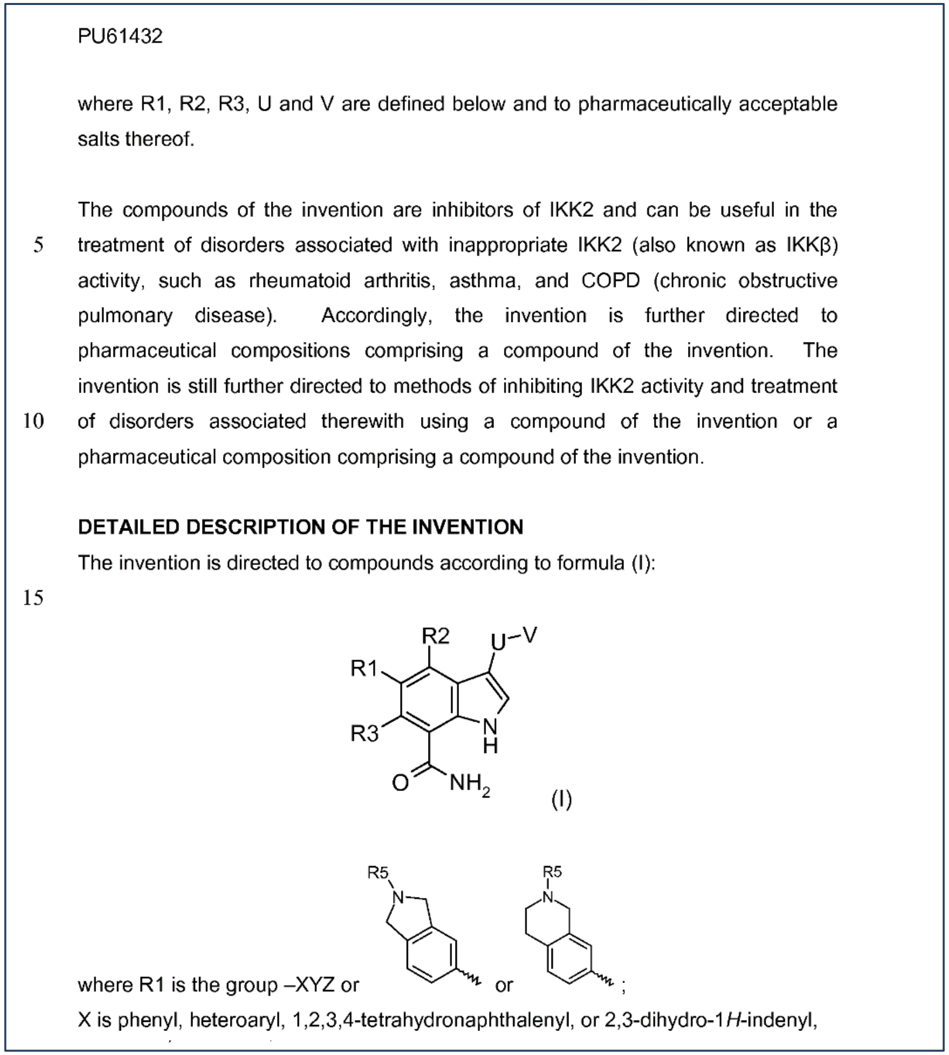
(8) $$\frac{\partial^2 u}{\partial x\,\partial y} = \frac{\partial^2 u}{\partial y\,\partial x}$$



(1)¶	x^6·+·<·s¶
(2)¶	2x2·=·4¶
(3)¶	.1ⁿBC·SkEFC¶
(•*)¶	2x2*5¶
(5)¶	(«·4-·f>)^2·=·»[·+·2ub¥b^2· ⁱ·....; ¶
(6)¶	ƒ·X^5</x·-l.i.m.·—·=·(-)·-·^{J}«>· «—*·*·-0.·Vil·α¶
¶	
(7)¶	>'=y¶
(8)¶	a^{2}n¶
	dx·ity·By·f*.x¶

USPTO – How the Solution Works

Sample Page



PU61432

where R1, R2, R3, U and V are defined below and to pharmaceutically acceptable salts thereof.

The compounds of the invention are inhibitors of IKK2 and can be useful in the treatment of disorders associated with inappropriate IKK2 (also known as IKKβ) activity, such as rheumatoid arthritis, asthma, and COPD (chronic obstructive pulmonary disease). Accordingly, the invention is further directed to pharmaceutical compositions comprising a compound of the invention. The invention is still further directed to methods of inhibiting IKK2 activity and treatment of disorders associated therewith using a compound of the invention or a pharmaceutical composition comprising a compound of the invention.

DETAILED DESCRIPTION OF THE INVENTION

The invention is directed to compounds according to formula (I):

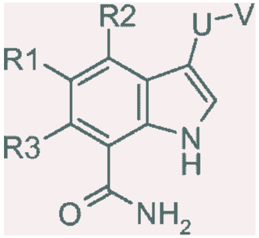
(I)

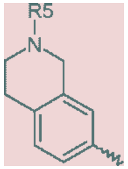
where R1 is the group –XYZ or or ;

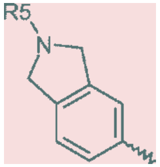
X is phenyl, heteroaryl, 1,2,3,4-tetrahydronaphthalenyl, or 2,3-dihydro-1*H*-indenyl,

Conversion Step 1– Artifact Removal

- Our solution extracts anything which is not text or can create problems for OCR conversion i.e. tables, extensive math, subscript, superscript.
- It creates a separate file for each image that is extracted, with its coordinates on the original page, to help us recreate the document.

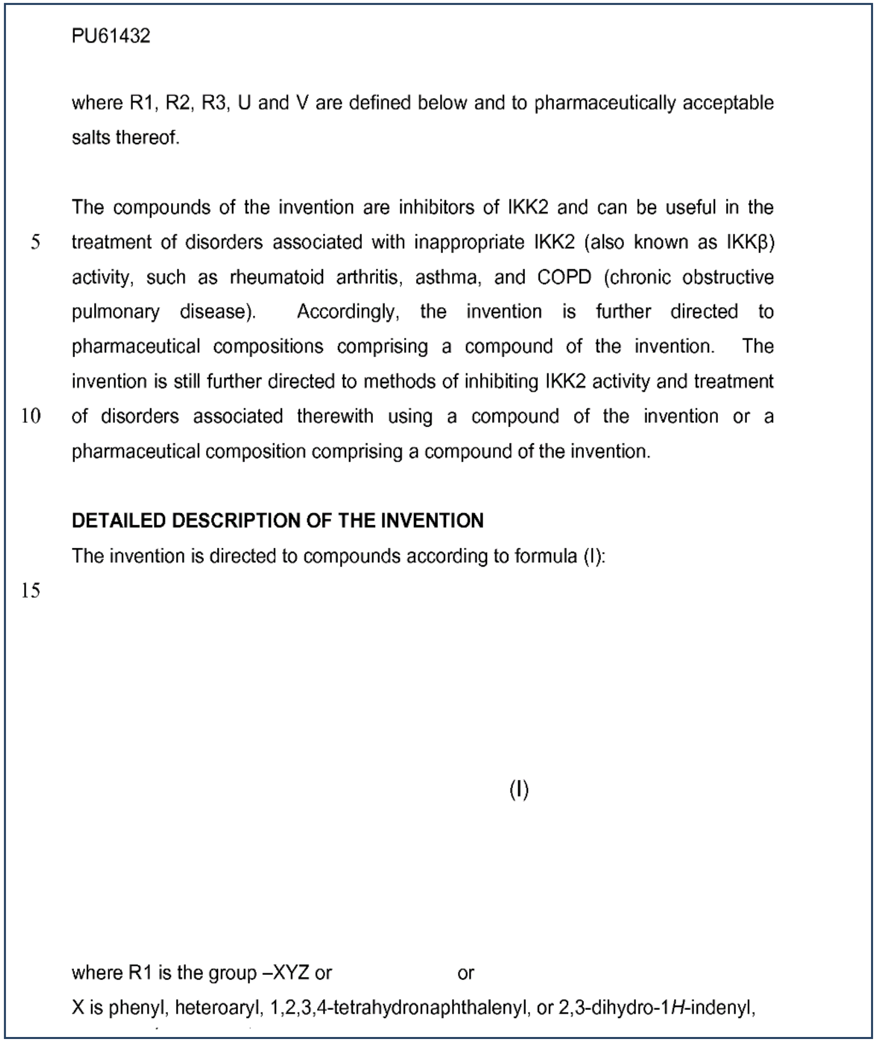






USPTO – How the Solution Works (cont'd)

Sample Page After Artifact Extraction



PU61432

where R1, R2, R3, U and V are defined below and to pharmaceutically acceptable salts thereof.

The compounds of the invention are inhibitors of IKK2 and can be useful in the treatment of disorders associated with inappropriate IKK2 (also known as IKKβ) activity, such as rheumatoid arthritis, asthma, and COPD (chronic obstructive pulmonary disease). Accordingly, the invention is further directed to pharmaceutical compositions comprising a compound of the invention. The invention is still further directed to methods of inhibiting IKK2 activity and treatment of disorders associated therewith using a compound of the invention or a pharmaceutical composition comprising a compound of the invention.

DETAILED DESCRIPTION OF THE INVENTION

The invention is directed to compounds according to formula (I):

(I)

where R1 is the group –XYZ or or

X is phenyl, heteroaryl, 1,2,3,4-tetrahydronaphthalenyl, or 2,3-dihydro-1*H*-indenyl,

Conversion Step 2 - OCR

- The modified page is then processed through a powerful OCR engine that can now accurately (99.6%) extract the text from the image.

Conversion Step 3 – Post Processing

- Once we have the OCR-extracted text and the associated image files with their coordinates, the post-processor analyzes the extracted text, using various advanced techniques to categorize and label the text.
- The post-processor creates well-formed XML in compliance with USPTO standards, using PE2E-conformant XML for the text and SVG Image XML for the images.
- The XML can be designed to match any XML Schema.

Building a Robust Conversion Process

- Tracking page topography
  - Paragraphs
  - Headings and Subheadings
  - Lists
- Identifying extraneous elements as boundary data
  - Line numbering
  - Headers/footers
- Tracking location to reinsert artifacts
- Identifying tag-able content & metadata
- Forms, and extracting data fields
- Linking artifacts back in

Automated Error Reporting and Quality Verification

- Building a process that doesn't stop on unexpected situations
- Logging and communicating unexpected situations
- Leveraging analytics to eliminate false positives
- Identifying anomalies for further review
- Reporting metrics on OCR quality and ambiguities

DCL™

Data Conversion Laboratory Inc.

Mark Gross, CEO
mgross@dclab.com
718-307-5710

Data Conversion Laboratory, Inc.
61-18 190th Street
Fresh Meadows, NY 11365
(718) 357-8700