

# **Artificially Impaired**

Scanning and optical character recognition limit artificial intelligence; beware these common problems



# **OCR limits AI success**

Gartner predicts corporate legal departments will spend 12% of in-house budgets on technology in 2025, a threefold increase from 2020 and up from just 2.6% in 2017. The same 2025 forecast warns that corporate legal departments will capture only 30% of the potential benefit of their contract life cycle management investments and advises them to "investigate, document and prioritize desired business outcomes and the necessary operational capabilities to achieve them."

Worldwide, artificial intelligence (AI) software revenue is soaring, but while AI holds great promise for legal teams, most oversee outdated contract repositories consisting of image-captured or PDF documents converted from a native editor (e.g., MS Word). Among the operational capabilities AI depends on in legal environments, a reliable data pool extracted from a contract repository tops the list for most teams.

Al technologies need these documents converted back into machine-readable formats, and optical character recognition (OCR) tools promise tremendous performance in doing so with many claiming 98% or 99% accuracy. Unfortunately, their paper legacies make these kinds of expectations unrealistic for most legal teams, but every situation is different.

Those contemplating AI investments or frustrated with tools they've already purchased should look to their contract repositories to understand how issues with OCR conversion may limit their success with AI. UnitedLex regularly uncovers OCR conversion errors while providing quality assurance to clients, and having traced countless errors back to specific terminology in specific sections of specific contracts, patterns have emerged. In the pages that follow, we've outlined some of the most common issues to look out for that can limit OCR conversion success and set AI investments up to underperform or fail outright.

## Common source-document problems:

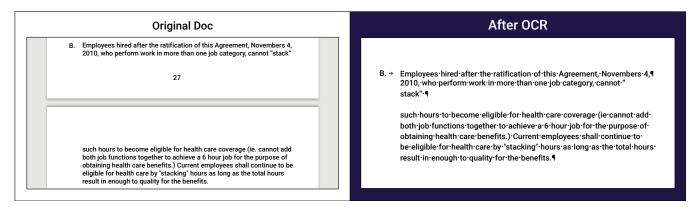
#### 1. Errors in table construction

Table fidelity might be lost between or across columns and rows, especially when affected by page breaks.

Original Doc		After OCR			
APPENDIX 1 LOCAL 1585 AFSOME CLASSIFICATION AND LEVEL SCHEDULE		APPENDIX 11 LOCAL 1585 AFSOME1 CLASSIFICATION AND LEVEL SCHEDULE1			
CLASSIFICATION AGRICULTURAL AND SPECIAL COUPMENT OPERATOR AGRICULTURAL AND SPECIAL COUPMENT OPERATOR AGRICULTURAL GROUP LEADER ANIMAL CARETAKER ANIMAL CARETAKER ANIMAL CARETAKER ANIMAL CARETAKER ANIMAL CARETAKER RAMONIST TRAINEE BANQUET SET-UP WORKERNELLOOG CENTER BUILDING SANITATION WORKI BUILDING SERVICES CUSTOWN HULDING SERVICES CUSTOWN HULDING SERVICES CUSTOWN HULDING SERVICES CUSTOWN COMPUS WORST ASSISTANT CENTRAL STERILIZATION WORKER COMMODITY COORDINATOR COOK RANGEMELLOGG CENTER COSTUME MAKER CONLOSE SERVICES ASSISTANT CONTO CONTRAINER CONLOSE SERVICES ASSISTANT CULLINARY PLATFORMATTENOWNY CUSTODIAN CUSTODIAN CUSTODIAN IN DECORATOR VU AWERS DINING ROOM B DINING SERVICE CASERATTENDANT	LEVEL 12 14 9 7 10 11 20 17 31 33 21 36 6 28 22 15 39 38 3 3 5 25 40 41 18 19	ACRICULTURAL AND SPECIAL COUPMENT OPERATOR AGRICULTURAL AND SPECIAL COUPMENT OPERATOR AGRICULTURAL AND SPECIAL COUPMENT OPERATOR AGRICULTURAL GROUP LEADER ANIMAL CARETAKER ANIMAL CARETAKER ANIMAL CARETAKER RESCARCH ANIMAL CARETAKER RESCARCH ARDORIST TRAINEE AVAMSU MUCERS TRANE BANQUET SET-UP WORKERNELLOOG CENTER BULDING SAKTAKA GROUP LEADER BULDING SERVICES CUSTOWN HULONG SAKACKA GROUP LEADER COMMODITY COORDINATOR COSTUME MAKER CONKRAMELLOGG CENTER CONK RANGEMELLOGG CENTER CONKONTAR LEKELLOGG CENTER CONK RANGEMELLOGG CENTER COSTUME MAKER CONLES HOUSE SERVICES ASSISTANT CULINARY PLATFORMATTENOWNY CUSTODIAN CUSTODIAN	LEVEL 124 99 74 104 29 201 171 21 331 221 334 339 339 399 399 399 399 399 409 409 419		
DINING SERVICES WORKER DINING SERVICES WORKER DRIVER BUYER ASSISTANT CYCLOTRON DRIVER COLLECTORALERI DRIVER COLLECTORAL FEED MILL OPERATORI FOOD MILL OPERATORE	37 44 23 13 16 4		189 199 374 441 234		

#### 2. Unnecessary new-line characters within paragraphs

Paragraphs may be needlessly split by erroneous new-line characters. This often results from OCR rules that indicate when a new line should start with either a number or a capital letter. Once processed as a new line, this results in a loss of context as the paragraph or sentence is broken into pieces.



#### 3. Oversized headers/footers

Although many technologies can ignore the top/bottom half inch of documents, when headers or footers run over that area, they will be captured as part of the document text. This erroneously collected information may be combined within paragraphs that span page breaks causing loss of fidelity.



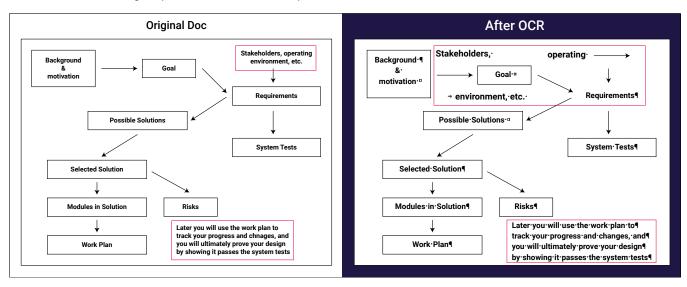
#### 4. Handwritten signatures, initials and notes

Handwritten signatures, initials, notes, etc. are usually captured by OCR software as images; however this is not always the case. Sometimes these are converted to random characters, resulting in incoherent data creation.

Original Doc		After OCR			
or as otherwise agreed by the parties For the University Multiple 10,341-202 U Michelle Sullivan Director of Labor Realtions	For the Union: <u>Hay of Picy Rozo</u> Kathyb j. Oppenheim Date UMPNC/President	or as otherwise agreed by the parties			
Nancy May Date Chief Nurse Executive	Vincent Schaub Date MNA Labor Representative	Director of Labor Realtions <sup>#</sup> Mancy May Date¶ → Vincent Schaub. → Date¶ Wrse <sup>#</sup> tive <sup>#</sup> Date¶			
		Signature Date Gate a			
Signature Date	Signature Date	Signature → Date → Signature → Date-MEMORANDUM-OF-UNDERSTANDING			

#### 5. Charts, graphs, diagrams and other images containing text/data

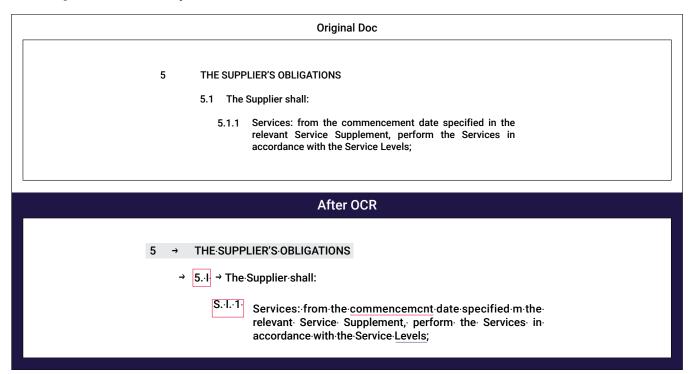
OCR software often misinterprets imagines containing text or data. Text often gets mixed with or removed from text boxes, and pictures can be anchored in incorrect locations. This causes completely incoherent rendering of portions or even complete sections of documents.



#### 6. Numbered lists

Numbered lists are often misread. These errors may be caused by a wide array of problems, including:

- Missing new line characters resulting in multiple list items being combined into one
- List numbers being incorrectly converted (e.g., S.1 instead of 5.1)
- · List numbers being converted into images
- · The inadvertent placement of a new line character within one item on a list
- Missing list values entirely



#### 7. Multiple columns to a page

Paragraphs may be needlessly split by erroneous new-line characters. This often results from OCR rules that indicate when a new line should start with either a number or a capital letter. Once processed as a new line, this results in a loss of context as the paragraph or sentence is broken into pieces.

	Original Doc			After OCR	
Contractor shall not enter into any subcontracts for any of the work scheduled under this contract other than described in Exhibit A without obtaining prior written approval from the Commission. Should the Contractor obtain prior written approval to enter into a subcontract with a qualified provider of services, the Subcontractor shall acknowledge the binding nature of this agreement and incorporate this agreement together with its attachments as appropriate. The Contractor must agree to be solely responsible for the performance of any Subcontractor Commission certifies at the time the contract is written that sufficient funds are available and authorized for expenditure to finance costs of this contract within the Commission's current appropriation or limitation. It is agreed that in the event the appropriation or funding	Commission is not obtained and continued at a level sufficient to allow for payments to the Contractor, for the services identified in Paragraph 2, the obligations of each party hereunder terminate upon delivery of written notice to the Contractor. Appropriation or limitation. It is agreed that in the event the appropriation or funding to the Commission is not obtained and continued at a level sufficient to allow for payments to the Contractor, for the services identified in Paragraph 2, the obligations of each party hereunder terminate upon delivery of written notice to the Contractor. The Contractor shall save and hold harmless the State of North Dakota and the Commission, its officers, agents, employees, and members, from all claims, suits, or actions of whatsoever nature resulting from or arising out of the activities of the Contractor	or its subcontractors, agents, or employees under this agreement. The Contractor shall obtain adequate liability insurance coverage, including, at a minimum, the maximum limits on damages established pursuant to N.D.C.C. §20-12.2-02. Notwithstanding the Contractor's responsibility for total management responsibility during the performance of the contract, the administration of the contract will require maximumcoordination between the Lignite Research Council, the Contractor. The Commission's Technical Representative (TR) will be designated on authority of the Contract. The types of actions within the purview of the TR's authority are to assure that the Contractor performs the technical	Contractor shall not enter into any subcontracts for any of the work scheduled under this contract other than described in Exhibit A without obtaining prior written approval from the Commission. Should the Contractor obtain prior written approval to enter into a subcontract with a qualified provider of services, the Subcontractor shall acknowledge the binding nature of this agreement and incorporate this agreement together with its attachments as appropriate. The Contractor must agree to be solely responsible for the performance of any Subcontractor Commission certifies at the time the contract is written that sufficient funds are available and authorized for expenditure to finance costs of this contract within the Commission's current appropriation or limitation. It is agreed that in the event the appropriation or funding	Appropriation or limitation. It is agreed that in the event the appropriation or funding to the Commission is not obtained and continued at a level sufficient to allow for payments to the Contractor, for the services identified in Paragraph 2, the obligations of each party hereunder terminate upon delivery of written notice to the Contractor. The Contractor shall save and hold harmless the State of North Dakota and the Commission, its officers, agents, employees, and members, from all claims, suits, or actions of whatsoever nature resulting from or arising out of the activities of the Contractor or its subcontractors, agents, or employees under this agreement. The Contractor shall obtain adequate liability insurance coverage, including, at a minimum, the maximum limits on damages established pursuant to N.D.C.C. §32-12.2-02.	requirements of the contract; to perform or cause to be performed inspections necessary in connection with the performance of the contract; to maintain both written and oral communications with the Contractor concerning the aspects of the written interpretations of the technical requirements of the statement of work; to monitor the Contractor's performance under the contract and notify the Commission of any deficiencies observed. contract and notify the Commission of any deficiencies observed. The Commission of any deficiencies observed. The Commission of any deficiencies observed. Contract and notify the State of North Dakota, and their duly authorized representatives, shall have access to the books, documents, papers and records of the Contractor which to the Contract on converted to the to the state of North Dakota, and the Office of the State of North Dakota, and their duly authorized representatives, shall have access to the contractor which to the
	<b>T</b> [	ſ	Commission is not obtained and continued at a level	Draft Contract for Private Entities Notwithstanding the Contractor's responsibility	the purpose of making audit, commons, excerpts, and
requirements of the contract; to perform or cause to be performed inspections necessary in connection with the performance of the contract; to maintain both written and oral communications with the Contractor concerning the aspects of the written interpretations of the technical requirements of the statement of work; to monitor the Contractor's performance under the commission of any deficiencies observed.	contract and notify the Commission of any deficiencies observed. The Commission, the State Auditor of the State of North Dakota, and the Office of the State of North Dakota, and their duly authorized representatives, shall have access to the books, documents, papers and records of the Contractor which are directly pertinent to the specific contract for the purpose of making audit, examination, excerpts, and transcripts.	the purpose of making audit, examination, excerpts, and transcripts. The parties agree that if any term or provision of this contract is declared by a court of competent jurisdiction to be illegal or in conflict with any law, the validity of the remaining terms and provisions shall not be affected, and the rights and obligations of the parties shall be construed & enforced as if the contract did not contain the particular term or provision held to be invalid.	and boundeed a fever sufficient to allow for payments to the Contractor, for the services identified in Paragraph 2, the obligations of each party hereunder terminate upon delivery of written notice to the Contractor. Appropriation or limitation. It is agreed that in the event the appropriation or funding to the Commission is not obtained and continued a a level sufficient to allow for payments to the Contractor, for the services identified in Paragraph 2, the obligations of each party hereunder terminate upon delivery of written notice to the Contractor.	contractor steporationary for total management responsibility during the performance of the contract, the administration of the contract will require maximumncoordination between the Lignite Research Council, the Commission and the Contractor. The Commission's Technical Representative (TR) will be designated on authority of the Corn all tech assist i continues past the Footer to include the actions text below before flowing to the 2 <sup>nd</sup> column	transcripts. The purpose of making audit, examination, excerpts, and transcripts. The parties are that if any term or provision of this contract is declared by a court of competent jurisdiction to be illegal or in conflict with any law, the validity of the remaining terms and provisions shall not be affected, and the rights and obligations of the parties shall be construed & enforced as if the contract did not contain the particular term or provision held to be invalid.

#### 8. Foreign-language and bilingual documents

Some OCR tools recognize non-English language and AI features will be able to identify the language within the document; however, the system must be set up to allow for this. Legal teams must be sure to configure the OCR system accordingly. Even when configured correctly, however, foreign language documents can confuse many OCR tools when mixed in with a large set of primary-language documents.

Documents that contain multiple languages on a single page may require special handling. Not all OCR vendors will be able to handle this situation. Teams managing these types of documents should be sure to research potential OCR software vendors and choose the best fit.

#### 9. Background images

If background images contain text, it may be captured and placed within the document. Images may also be captured and placed as separate images anchored within the document.



#### 10. Inverted/Upside-down scans

OCR software may not recognize the situation when it encounters upside down documents or pages. The OCR process often renders the verbiage incorrectly, resulting in gibberish from the conversion.

Original Doc	After OCR
15.14 Headings. The headings in this Agreement are for the convenience of the parties only, and are in no way intended to define or limit then scope or interpretation of the Agreement or any provision hereof.	5.14 The parties only, and are for the Agreement are for and are for this Agreement or the parties only, any provision hereof Headings. The hereof.
15.15 Relationships an independent contractor and nothing in this Agreement shall render either party an employee, agent or partner of the other, except for any specific Customer appointment(s) of as its agent under a Customer maintenance or software license agreement with a third party as set forth in attachments to Exhibit A. Neither party shall be liable for the acts or omissions of the other.	5.15 Relationships an in thing in attachment wither party as its to Exhibit A. Neithe or appointenance or party as set for party shall render partner party and nother maintenance othe other party as set for omissionships an employee, agreement wither a thing in this Agreement contractor an inder party as set for a ther appointenance othird party an employee, agreement
15.16 Entire Agreement. This Agreement and its exhibits constitute the entire agreement between and Customer, and supersede any prior or contemporaneous communications, representations or agreements between the parties, whether oral or written, regarding the	contracts to Exhibit A. Neither omissionships agreement wither party as set for party an employee, agent shall be liable for omissions or
subject matter of this Agreement. Customer's additional or different terms and conditions shall not apply. The terms and conditions of this Agreement may not be changed except by an amendment signed by an authorized representative of each party.	5.17 Customer's additions, whether of this Agreement. The subject matten, representational or agreementational or agreement and subject may not by and its contemporal or cons, whethe subject may not an authorized exhibits
15.17 Applicable Law. This Agreement is made under and shall be construed in accordance with the law of B giving effect to that jurisdiction's choice of law rules.	exhibits be changed by and communications, whethe ent. This Agreement between ther of each parties, represent.

## Surround OCR with quality assurance:

Machine AI and analytic accuracy will only be as precise as the data available to it. Incorrect or inaccurate captured text skews the analysis and sets the technology up to fail by decreasing the technology's precision rate (how accurate) and recall rate (how complete). Documents converted directly to PDFs or e-signature documents are relatively clean and convert back from PDF to a machine-readable version very well. Most errors arise from image capture or scanned documents. Buyers should be wary that many problems can arise, take care to choose the best solutions for their situation, and leverage talented people on both sides of the OCR conversion process.

Those planning to invest in new AI solutions and other technology relying on significant OCR conversions should first try to better understand their contract repositories. Are they flush with multiple languages, paper-scanned pdfs, complex page formats, or other potential challenges? Better understanding what needs OCR conversion can help legal teams identify the best tools for the job and determine if simple steps could address shortcomings of their contract repository or if other solutions might be more effective.

Legal teams should assess their levels of risk when it comes to OCR precision and recall and respond with an appropriate quality assurance effort to monitor performance. Putting smart people to work to spot check and look for patterns of problematic conversions can help to pinpoint and address or eliminate problematic data. This approach offers legal teams the best way to improve the precision and recall of OCR efforts and give their Al investments a fighting chance to meet expectations.

### About UnitedLex

UnitedLex is a data and professional services company delivering outcomes that create value for high-performing law firms and corporate legal departments in the areas of digital litigation, intellectual property, contracts, compliance, and legal operations.

Founded in 2006 with a mission to push the boundaries of legal innovation, we provide solutions that enable measurable performance improvement, risk mitigation, revenue gain, cost reduction and digital business synergy. Our team of 3,000+ legal, data and technology professionals supports our clients from operational centers around the world.

For more information, visit <u>https://unitedlex.com</u>