



WHAT'S NEW IN 2.9

UPCOMING FEATURES AND FUNCTIONALITY

28 July 2020



WHAT'S NEW IN 2.9

- **Data Integration Improvements**
 - New MS Office Integration
 - New Box integration
 - Enhanced database export
 - New document viewer, supporting multiple document views
 - Data lookups can now be performed against CMIS sources
 - Improved OCR synthesis and paragraph detection
 - Classify, Extract, and Data review now support content type filtering
- **Improvements for Architects**
 - New Extractor weighting mechanisms
 - TF/IDF training builds training batch
 - Compile stats command for content containers
 - New TF/IDF variants for lexical classification
 - New weighted classification properties allow for new classification methodologies
 - New Inheritance Architecture
- **Separation Improvements**
 - Improved unstructured document separation training
 - Improved ESP auto-separation logic
 - New settings on document types to aid separation
 - New separation review UI
- **New Data Annotation Option in data review**
- **Code expressions support LINQ and String Interpolation**



DATA INTEGRATION IMPROVEMENTS

MS OFFICE INTEGRATION

NATIVELY INGEST MULTIPLE FILE FORMATS GENERATED BY THE MS OFFICE SUITE:

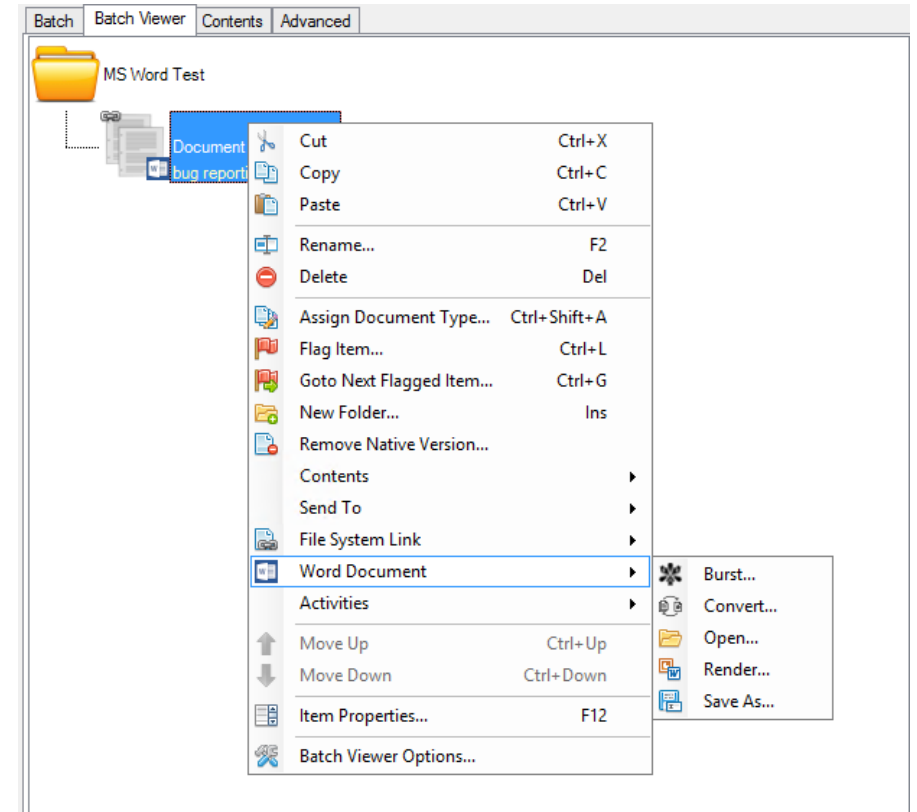
- Word document
- PowerPoint presentation
- Excel spreadsheet
- via CMIS, Outlook messages and mailboxes

NATIVELY PULL TEXT FROM ALL SUPPORTED FILETYPES

REMOVES THE NEED FOR SINGLE-THREADED RENDERING

MS OFFICE MUST BE INSTALLED ON MACHINE RUNNING DESIGN STUDIO

REPRESENTS BETTER INTEGRATION WITH CONTENT PRODUCTION SYSTEMS





NEW BOX INTEGRATION

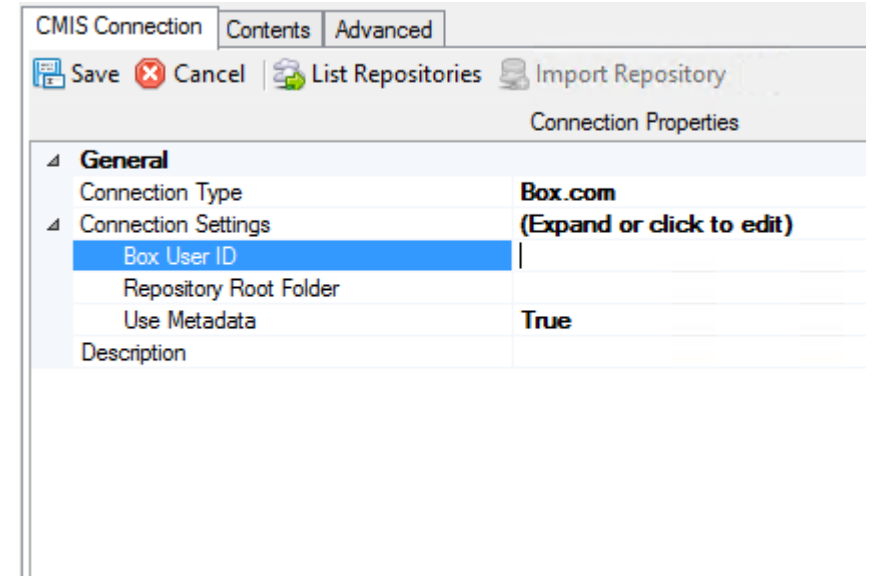
CONNECT DIRECTLY TO A BOX ACCOUNT TO IMPORT OR EXPORT FILES

WORKS WITH BOX BUSINESS OR ENTERPRISE

TAKES ADVANTAGE OF GROOPER'S EXISTING CMIS+ ARCHITECTURE, GIVING YOU ACCESS TO METADATA MAPPING AND EXPRESSIONS, CMIS LOOKUPS, ETC.

OUR FIRST-EVER INTEGRATION WITH A CLOUD ONLY PROVIDER; POSITIONS DEPLOYMENTS TO TAKE ADVANTAGE OF ONE OF THE LEADING STORAGE PROVIDERS

CAN USE FOLDERS WITH GOVERNANCE LEVEL ASSIGNED TO TAKE ADVANTAGE OF GOVERNANCE FEATURES

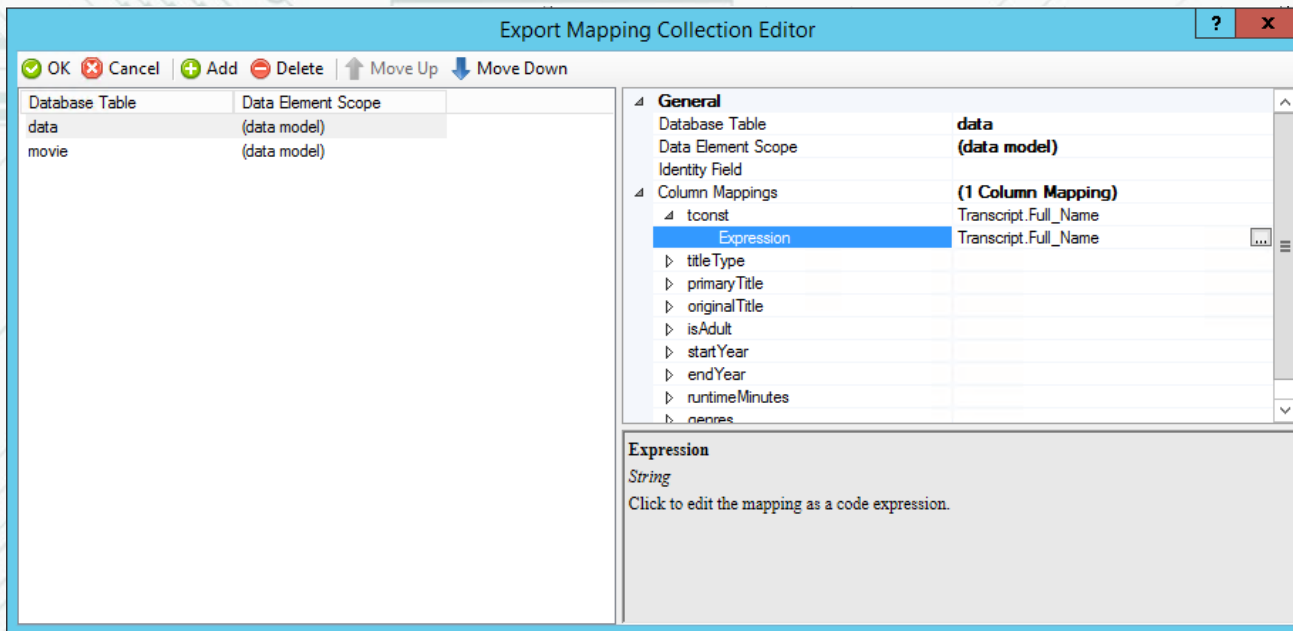
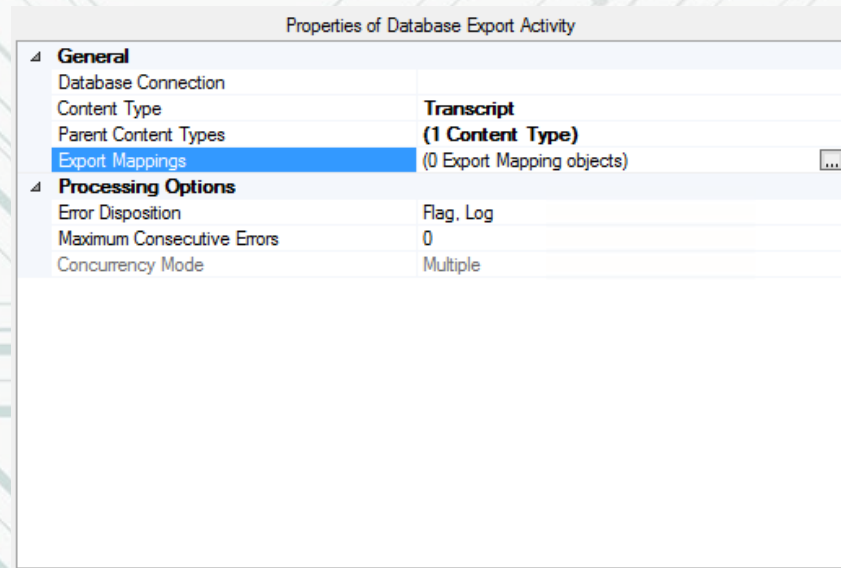




ENHANCED DATABASE EXPORT

Multiple exports can now be defined on a single database export step

- Must be within a single database; can span multiple tables
- Can choose separate data element scope on each table mapping object
- Greatly simplifies complicated or multipart database exports
- Supports SQL-server generated identity columns





NEW DOCUMENT VIEWER

**ALLOWS FOR DISPLAY OF MULTIPLE DOCUMENT
"RENDITIONS"**



NEW VIEWER: NATIVE RENDITION

Batch Folder Contents Advanced

Save Cancel Assign Document Type... Flag Item... Remove Native Version...

General	
Content Type	
Flagged	False
Description	
Attachment	
File Name	Piers Plowman.eml
File Name Extension	.eml
MIME Type	message/rfc822
File Type Description	Mail Message
MIME Type Info	
Mail Message	
Has Local Copy	True
Has PDF Version	False
Content Link	CMIS Document Link

Batch Folder

A Batch Folder is a container object within the hierarchical structure of a [Batch](#), which is used to represent logical folders and multipage documents.

Remarks

Batch Folders may contain [Batch Page](#) objects or other folders as children.

Batch Folder objects may be created in a variety of ways. If a [Separation Profile](#) is assigned in the [Scan](#) activity, then folders will be created in real time during scanning. The [Separate](#)

Piers Plowman.msg (2)

Piers Plowman.eml

Native (Mail Message)

From: Chris Dearer <cdearner@bisok.com>
Sent: Monday, October 7, 2019 2:09 PM
To: Chris Dearer <cdearner@bisok.com>
Subject: Piers Plowman

IN a somer seson,
Whan softe was the sonne
I shoop me into shroudes
As I a sheep weere
In habite as an heremite
Unholy of werkes,
Wente wide in this world
Wondres to here;
Ac on a May morwenyng
On Malverne hilles
Me bifel a ferly,
Of fairye me thoughte.
I was wery for-wandred,
And wente me to reste
Under a brood ban
By a bournes syde;
And as I lay and lenede,
And loked on the watres,
I slombred into a slepyng,
It sweyed so murye.

Thanne gan I meten
A merveillous swevene,
That I was in a wildernesse,
Wiste I nevere where,
And as I biheeld into the eest
An heigh to the sonne,
I seigh a tour on a toft
Trieliche y-maked,
A deep dale bynethe,
A dongeon therinne,
With depe diches and derke
And dredfulle of sighte.
A fair feeld ful of folk
Fond I ther bitwene,
Of alle manere of men,
The meene and the riche,
Werchyng and wandryng,
As the world asketh.

NEW VIEWER: ALTERNATIVE RENDITION

Batch Folder Contents Advanced

Save Cancel Assign Document Type... Flag Item... Remove Native Version...

RFC822 Source

General	
Content Type	
Flagged	False
Description	
Attachment	
File Name	Piers Plowman.eml
File Name Extension	.eml
MIME Type	message/rfc822
File Type Description	Mail Message
MIME Type Info	Mail Message
Has Local Copy	True
Has PDF Version	False
Content Link	CMIS Document Link

Batch Folder

A Batch Folder is a container object within the hierarchical structure of a [Batch](#), which is used to represent logical folders and multipage documents.

Remarks

Batch Folders may contain [Batch Page](#) objects or other folders as children.

Batch Folder objects may be created in a variety of ways. If a [Separation Profile](#) is assigned in the [Scan](#) activity, then folders will be created in real time during scanning. The [Separate](#)

Piers Plowman.msg (2)
Piers Plowman.eml

```
Content-Type: multipart/alternative; boundary="6a869d85c4844656aa2914b34f7a5f61"
From: Chris Dearer <cdearner@bisok.com>
To: Chris Dearer <cdearner@bisok.com>
Subject: Piers Plowman
Date: Mon, 7 Oct 2019 19:09:10 +0000
Received: from SN6PR1901MB2061.namprd19.prod.outlook.com
(2603:10b6:803:2d::12) by SN6PR1901MB2032.namprd19.prod.outlook.com with
HTTPS via SN4PR0701CA0032.NAMPRD07.PROD.OUTLOOK.COM; Mon, 7 Oct 2019 19:09:11
+0000
MIME-Version: 1.0

--6a869d85c4844656aa2914b34f7a5f61
Content-Type: text/html; charset="utf-8"

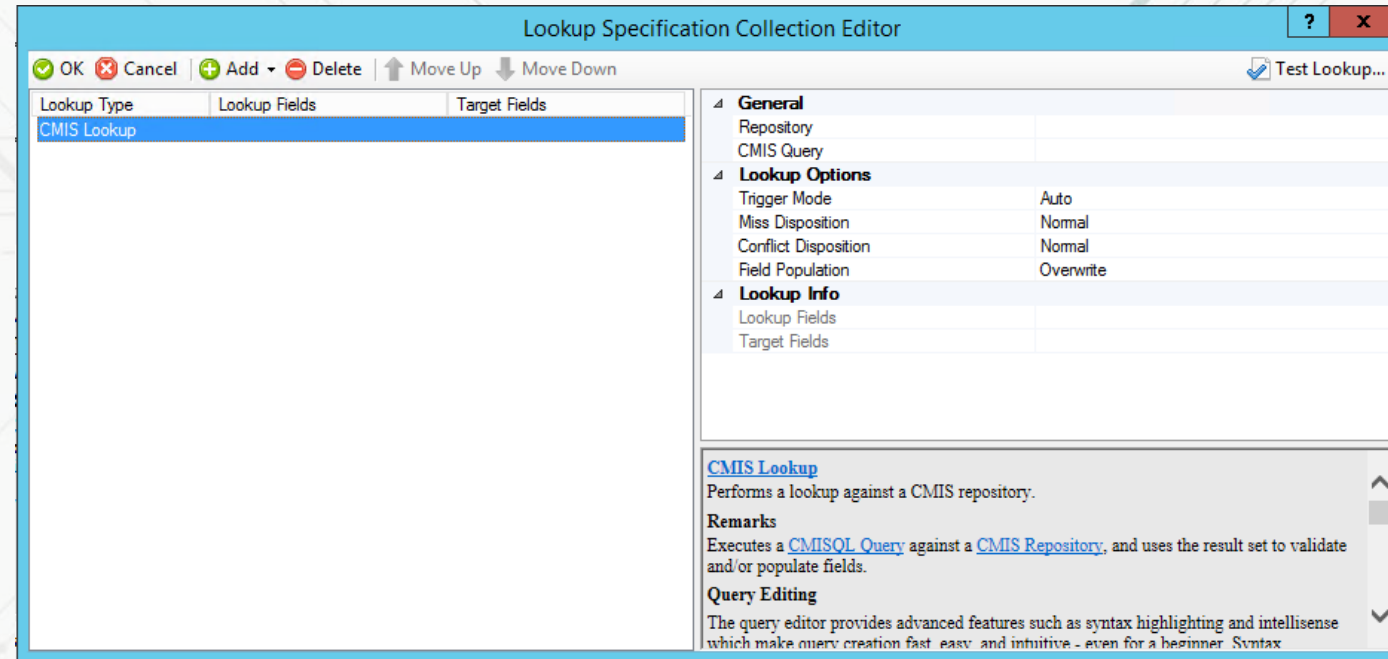
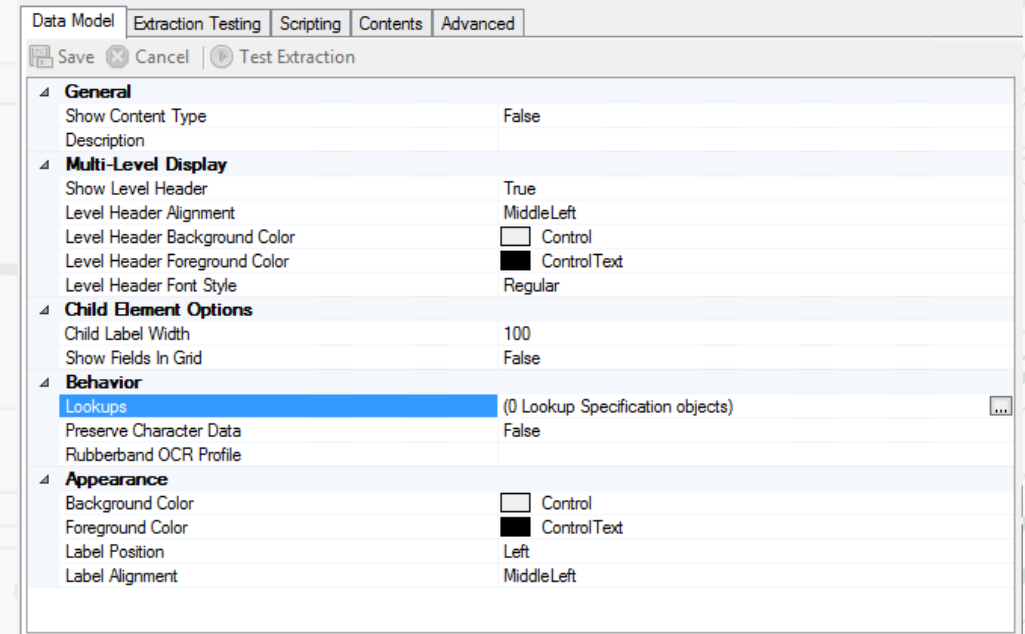
<html xmlns:v="urn:schemas-microsoft-com:vml" xmlns:o="urn:schemas-microsoft-com:office:office" xmlns:w="urn:schemas-microsoft-com:office:word" xmlns:m="http://schemas.microsoft.com/office/2004/12/omml" xmlns="http://www.w3.org/TR/REC-htm140"><head><meta name=Generator content="Microsoft Word 15 (filtered medium)"><style><!--
/* Font Definitions */
@font-face
{font-family:"Cambria Math";
panose-1:2 4 5 3 5 4 6 3 2 4;}
@font-face
{font-family:Calibri;
panose-1:2 15 5 2 2 2 4 3 2 4;}
@font-face
{font-family:"Roboto Light";
panose-1:0 0 0 0 0 0 0 0 0 0;}
/* Style Definitions */
p.MsoNormal, li.MsoNormal, div.MsoNormal
{margin:0in;
margin-bottom:.001pt;
font-size:11.0pt;
font-family:"Calibri",sans-serif;}
a:link, span.MsoHyperlink
{mso-style-priority:99;
color:#0563C1;
text-decoration:underline;}
a:visited, span.MsoHyperlinkFollowed
{mso-style-priority:99;
color:#954F72;
text-decoration:underline;}
span.EmailStyle17
{mso-style-type:personal-compose;
font-family:"Calibri",sans-serif;
color:windowtext;}
.MsoChpDefault
{mso-style-type:export-only;
font-family:"Calibri",sans-serif;}
@page WordSection1
{size:8.5in 11.0in;
margin:1.0in 1.0in 1.0in 1.0in;}
div.WordSection1
{page:WordSection1;}
--></style><!--[if gte mso 9]><xml>
<o:shapedefaults v:ext="edit" spidmax="1026" />
</xml><![endif]><!--[if gte mso 9]><xml>
<o:shapelayout v:ext="edit">
<o:idmap v:ext="edit" data="1" />
</o:shapelayout></xml><![endif]></head><body lang=EN-US link="#0563C1" vlink="#954F72"><div class=WordSection1><p
```



CMIS DATA LOOKUPS

Functions similarly to a Database lookup and allows for population or validation of Grooper fields based on metadata located on CMIS objects

- Allows for better integration with CMIS data sources
- Allows for treating CMIS data sources as containing queryable information

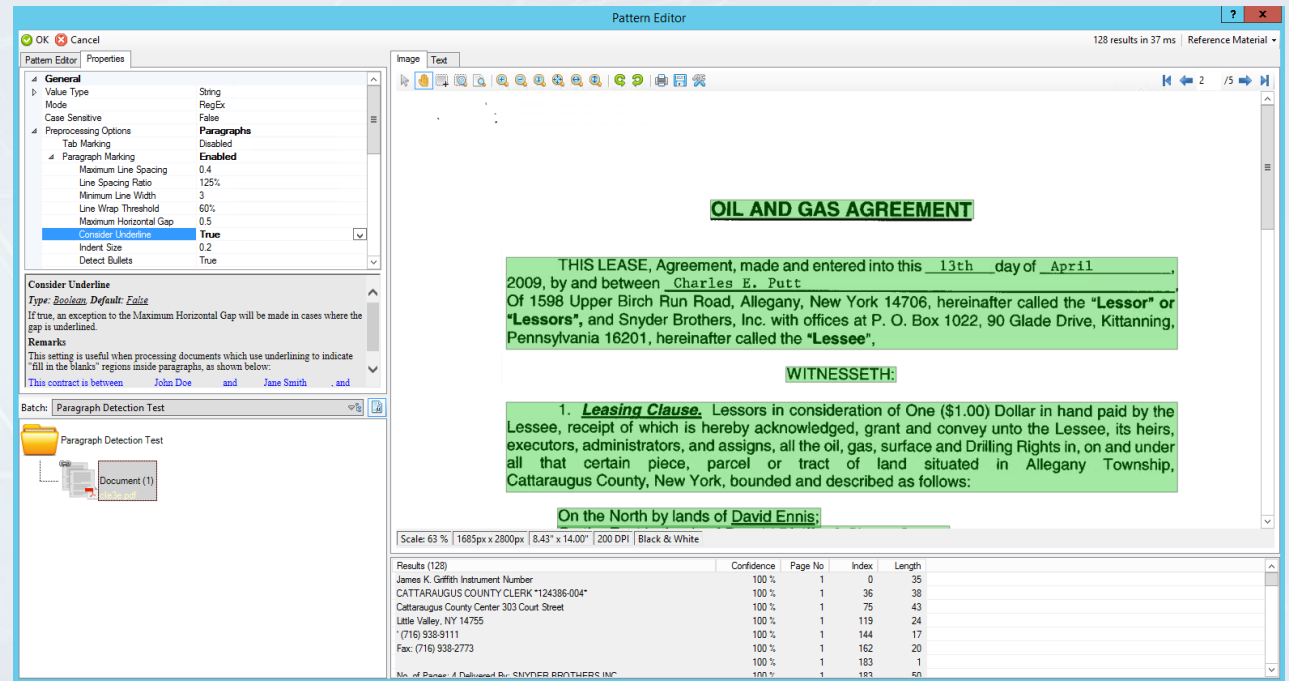




IMPROVED OCR SYNTHESIS AND PARAGRAPH DETECTION

Grooper can now detect paragraphs containing large sections of blank underlines correctly

There are additional various improvements to OCR Synthesis





IMPROVEMENTS FOR ARCHITECTS



COMPILE STATS

COMPILE STATS PROVIDES
COMPREHENSIVE STATISTICS ON
CLASSIFICATION AND EXTRACTION
ACROSS AN ENTIRE CONTENT MODEL

ASSISTS ADMINISTRATORS IN
DEVELOPING AND TROUBLESHOOTING
CONTENT MODELS, ALLOWING FOR MORE
TARGETED WORK AND SHORTER
DEVELOPMENT TIMES.

4,344 OBJECTS:

OBJECT TYPE	COUNT
Batch Page	985
Data Type	862
Data Format	755
Training Page	752
Batch Folder	308
Data Field	151
Page Type	138
Document Type	109
Folder	91
Data Model	88
Form Type	74
Lexicon	13
Content Category	8
Batch Process Step	4
Field Class	4
Batch	1
Batch Process	1

109 DOCUMENT TYPES:

DOCUMENT TYPE	CHILD COUNT	FIELD COUNT	XTRACT COUNT	FIELD NAMES
ATTACH*ACK	18	1	0	DocType
ATTACH*BLM Attachment	29	1	0	DocType
ATTACH*EXHIBIT	23	1	0	DocType
ATTACH*SIG	8	1	0	DocType
BLM*CORR BLM	2	1	0	DocType
BLM*LEASE CORR BLM	2	1	0	DocType
BLM*TITLE ASMT CORR BLM	3	1	0	DocType
BLM*TITLE CA CORR BLM	2	1	0	DocType
BLM*TITLE MERGER CORR BLM	2	1	0	DocType
CORR*CORR EHL	53	4	3	DocType, Sender, Recipient, Date Sent
CORR*CORR FAX COVER SHEET	12	1	0	DocType
CORR*CORR LTR	79	4	3	DocType, Sender, Recipient, Document Date
CORR*CORR MAIL RECEIPT	19	2	1	DocType, Document Date
CORR*CORR MEM	53	4	3	DocType, Memo Date, Sender, Recipient
CORR*CORR MISC	7	2	1	DocType, Document Date
HIST*HIST COMM RPT	20	2	1	DocType, Date



CONTENT TYPE FILTERING SUPPORTED IN EXTRACTION, CLASSIFICATION, AND DATA REVIEW ACTIVITIES

- Appears in “index navigator settings” on Data Review Activity
- Allows for more focused or partial classification, extraction, and review
- Allows classification, extraction, and review to proceed in stages for larger or more complicated projects
- Supports targeted expansion or improvement
- Supports tiered classification decisioning
- Allows for classification to target only unclassified documents (“All,” “Classified”, or “Unclassified”)

Index Navigator Settings

OK Cancel

General	
Flag Invalid Items	False
Processing Level	Level 1
Display Parent Folder(s)	False
Auto-Load Next Invalid Document	False
Content Type Filter	(0 Content Type objects)
Flag Messages	(empty)

Content Type Filter

Type: List of [Content Type](#)

An optional list of [Content Types](#) to which the review process should be restricted.

Remarks

When this list is empty, all documents in scope will be included in the review process. If a list is provided, only document types matching one of the list entries will be included.

Properties of Extract Activity

General	
Mode	Normal
Default Content Type	
Flag Invalid Items	False
Content Type Filter	(0 Content Type objects)
Processing Options	
Error Disposition	Flag, Log
Maximum Consecutive Errors	0
Concurrency Mode	Multiple

Properties of Classify Activity

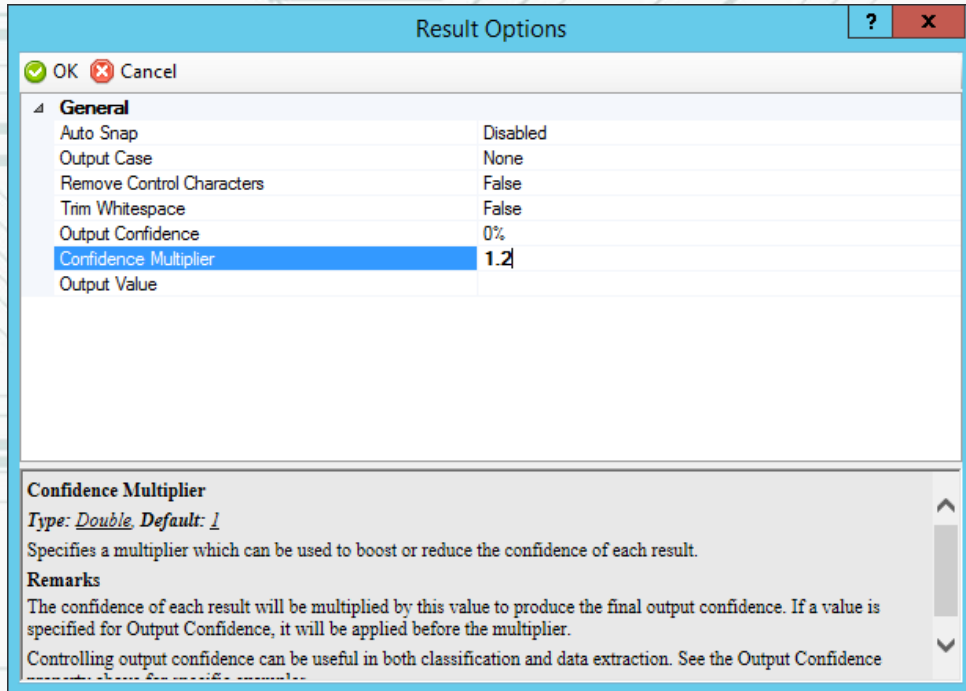
General	
Content Model Scope	Blump University
Classification Level	DocType
Output Level	DocType
Apply To	All
Reclassify Mode	Overwrite
Suppress Candidate List	False
Model Refresh Interval	60
Processing Options	



NEW EXTRACTOR WEIGHTING MECHANISMS

CONFIGURABLE OUTPUT CONFIDENCE AND OUTPUT MULTIPLIERS ON DATA TYPES

- Manually set extraction confidence and multiply confidence by a set amount
- Allows for waterfall extraction techniques: most reliable methods can be backed up by less-reliable methods by using "order by confidence"





WEIGHTED CLASSIFICATION PROPERTIES AND WATERFALL CLASSIFICATION

ADVANCED CLASSIFICATION STRATEGIES FOR COMPLICATED DOCUMENT SETS

- Positive extractors now classify documents with the confidence of their referenced extractor
- Using the confidence multiplier and output confidence configuration items on data types allows for the creation of waterfall classification strategies
- This allows for documents that are highly lexically dissimilar within a group or highly lexically similar across groups to be targeted with better accuracy (e.g. legal documents, oil and gas leases vs lease acknowledgements, etc.)

Document Type | Data Element Overrides | Weightings | Contents | Advanced

Save | Cancel | Compile Stats... | Purge Training... | Rebuild Training...

General

Name Field	
Allow Training	True
Page Scope - Data Extraction	(unlimited)
Base Content Type	Document
Description	

Classification

Positive Extractor	Reference: Positive Rule
Negative Extractor	(none)
Minimum Pages	0
Maximum Pages	0
Scoring Limit	0%

Separation

Pagination	Unstructured
Combine Contiguous Documents	False
Prioritize EPI	True
Training Scope	Normal
Secondary Page Extractor	(none)

Result Options

OK | Cancel

General

Auto Snap	Disabled
Output Case	None
Remove Control Characters	False
Trim Whitespace	False
Output Confidence	0%
Confidence Multiplier	1.2
Output Value	

Confidence Multiplier
Type: *Double* Default: *1*
Specifies a multiplier which can be used to boost or reduce the confidence of each result.

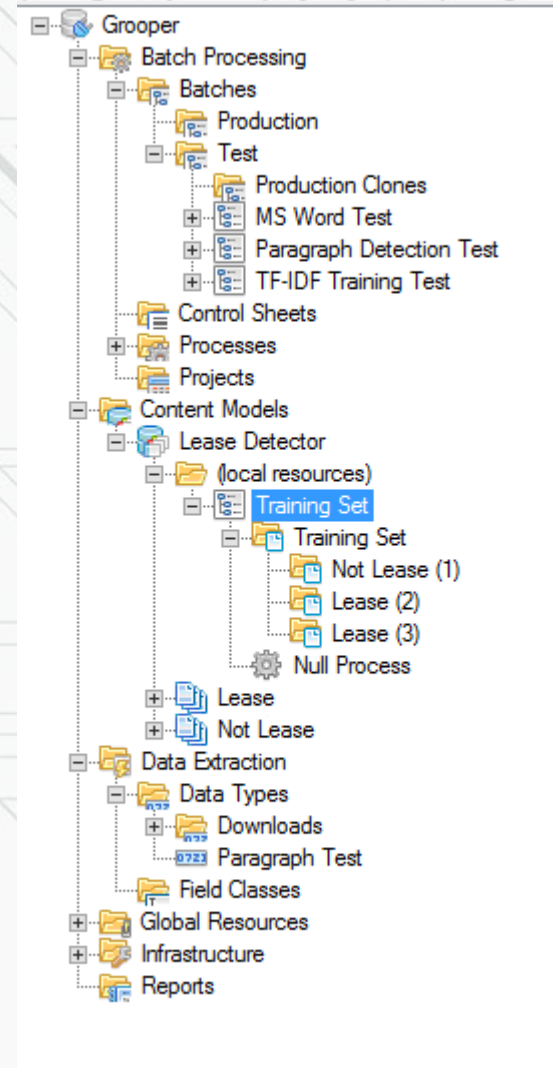
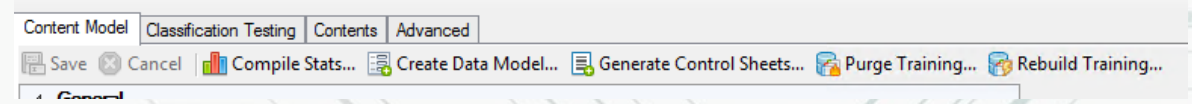
Remarks
The confidence of each result will be multiplied by this value to produce the final output confidence. If a value is specified for Output Confidence, it will be applied before the multiplier.
Controlling output confidence can be useful in both classification and data extraction. See the Output Confidence

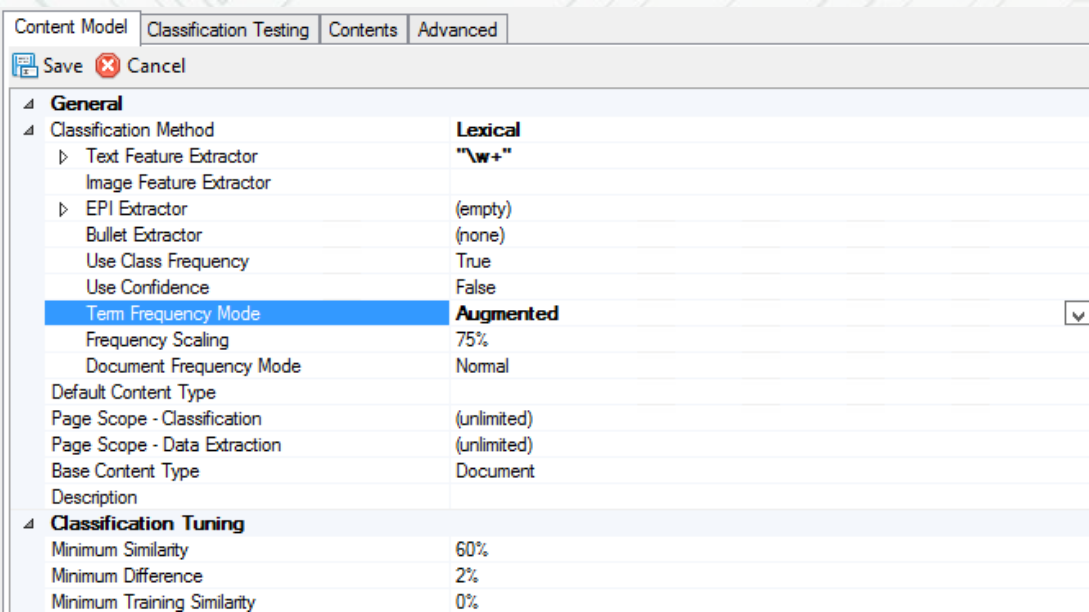


AUTOBUILT TF/IDF TRAINING BATCH

TF/IDF training now automatically builds training batch, allowing for easier training after changes to feature collection or TF/IDF settings.

“Rebuild training” allows for tuning and A/B testing using identical training set and document training decisions





NEW TF/IDF VARIANTS

- 'Use confidence' allows extraction confidence to factor into TF/IDF feature scoring
- 'Augmented' TF mode (in addition to logarithmic/'sublinear') allows for TF/IDF to partially account for document length (standard TF/IDF tends to favor longer documents)
- Smooth TF/IDF selectable under "Document Frequency Mode"

Use Confidence

Type: *Boolean*, Default: *False*

Specifies how the confidence of each feature instance should be calculated into the term frequency.

Term Frequency Mode

Type: *TfModes*, Default: *Normal*

Specifies how the term frequency (TF) of features should be calculated. Can be one of the following values:

- **Normal** - Term frequency is normalized to the size of the document. This mode provides matching which is completely independent of document size. In some classification scenarios, however, this mode can produce undesirable false positives, as it allows a very short document to match a very long document with a high degree of similarity.
- **Logarithmic** - Term frequency is scaled logarithmically. This mode is provided for backwards compatibility.
- **Augmented** - Term frequency is normalized to the frequency of the most-frequent feature. Provides an alternative to Normal mode where the size of the document plays more of a role in classification decisions. This mode also provides a Dampening Factor

Document Frequency Mode

Type: *IdfModes*, Default: *Normal*

Specifies how the inverse document frequency (IDF) of features should be calculated. Can be one of the following values:

- **Normal** - Standard IDF.
- **Smooth** - Smooth IDF.



NEW INHERITANCE ARCHITECTURE

DATA ELEMENT PROFILES ARE GONE; LONG LIVE DATA ELEMENT OVERRIDES

- Data element profiles have been replaced by data element overrides
- Every property is now overridable
- Data Element Overrides tab includes built-in tester
- A fuller realization of Grooper's inheritance-based architecture; allows for more robust, more scalable data models.
- Saves time and reduces complexity when architecting large systems

The screenshot displays the Grooper software interface. On the left, there are two panels: 'Data Element Overrides' and 'Property Overrides'. The 'Data Element Overrides' panel shows a tree view for 'LEASE' with sub-elements 'Description', 'Filename', and 'Lease Number'. The 'Property Overrides' panel shows a table of properties for 'General', 'Appearance', and 'Behavior'. The 'General' section includes 'Value Type' (String), 'Value Extractor' (none), 'Minimum Confidence' (0%), and 'Description'. The 'Appearance' section includes 'Alignment' (Left), 'Display Width' (300), 'Error Color' (255, 192, 192), 'Background Color' (Window), 'Foreground Color' (WindowText), 'Visible' (False), 'Label Position' (Left), and 'Label Alignment' (MiddleLeft). The 'Behavior' section includes 'Auto Complete' (Disabled) and 'Input Mask' (Disabled). Below these panels is a 'Data Element Profile' section with a description and a list of used properties.

The main window shows a 'Data Field Preview' with the text 'No fields are visible.' Below that is a 'No Issues' status bar. The main document view shows a scanned document with the following text:

James K. Griffith Instrument Number *124386-004*
CATTARAUGUS COUNTY CLERK
Cattaraugus County Center 303 Court Street
Little Valley, NY 14755
(716) 938-9111
Fax: (716) 938-2773

No. of Pages: 4 Delivered By: SNYDER BROTHERS INC

Receipt No. 124386 Return To:
SNYDER BROTHERS INC
P O BOX 1022
KITTANNING PA 16201
ENVELOPE (4)

DATE: 07/16/2009

Time: 03:07 PM

Document Type: LEASE

Parties To Transaction: O&G-PUTT TO SNYDER BROTHERS INC

Scale: 50% | 1685px x 2200px | 8.43" x 11.00" | 200 DPI | Black & White



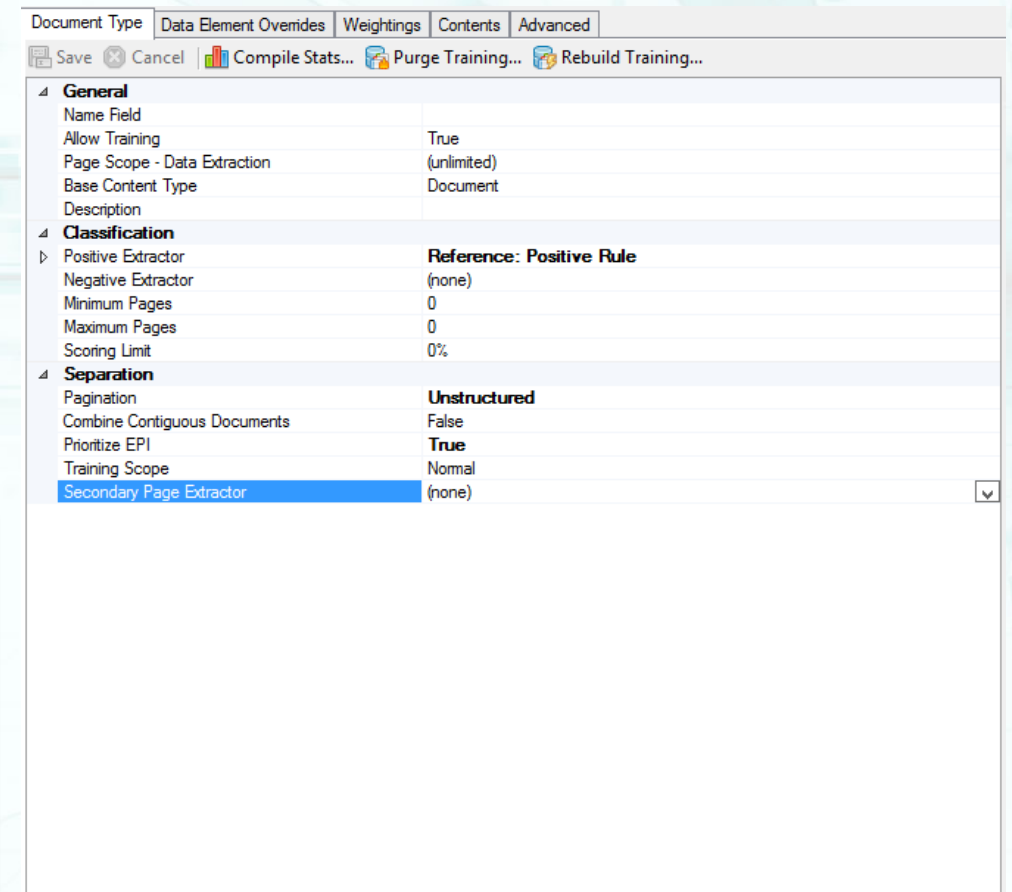
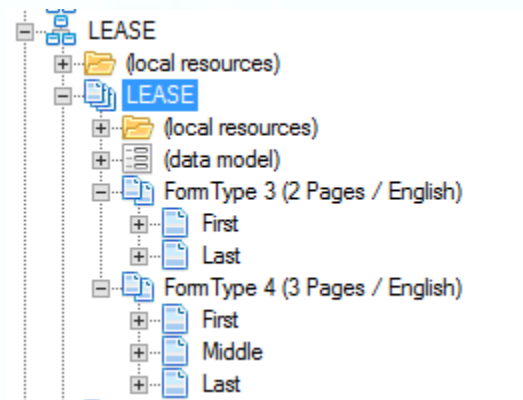
SEPARATION IMPROVEMENTS



ESP AUTO-SEPARATION IMPROVEMENTS

SEPARATION IS MORE ROBUST AND ACCURATE THAN EVER

- Improved ESP auto-separation logic
- New Secondary Page Extractor
- Unstructured Documents now consider only first, middle, last pages; each n-page version is no longer considered a different formtype



New Separation review UI

SIMILAR TO ESP AUTO-SEPARATION TESTER

The screenshot displays the 'Grooper Attended Client' window. The interface includes a menu bar (Processing, Task, Batch, Help), a toolbar with icons for Start, Stop, Open Task, Close Task, Complete Task, and New Batch. A list of 138 items is shown on the left, with columns for Name, Document Type, Page No, EPI, and Similarity. Below this is a 'Document Types' section with a search table. The main window displays a 'LIMITED OWNERSHIP REPORT' with the following details:

LIMITED OWNERSHIP REPORT

Dated: May 5, 2017
 Updated: April 23, 2019
 Research Dated: May 4, 2017
 Research Update: April 17, 2019

LANDS COVERED BY THIS REPORT: **WELD COUNTY, COLORADO**

Township 1 North - Range 67 West, of the 6th P.M.
 Section 33: Lot B of Recorded Exemption RECX-16-0076, being a part of the SE¼

Containing 71.45 acres, more or less

SURFACE OWNERSHIP:

WW, LLC
 9457 S. University Boulevard, Apt. 401
 Highlands Ranch, CO 80126

MINERAL OWNERSHIP:

OIL AND GAS OWNERSHIP	INTEREST	LEASEHOLD STATUS
Timothy J. Pivonka 2116 14 th Street SE Loveland, Colorado 80537	34.00%	HELD BY PRODUCTION

At the bottom, a status bar shows: Status: Task Open | BIS\cleaner on OKCSSND01 | Repository: Grooper | User Idle Time: 00:00:28 | Batch: Separation Test | Process: Core Process | Step: Separate Review | Review: 0 tasks pending



OTHER IMPROVEMENTS



NEW DATA ANNOTATION OPTION IN DATA REVIEW

DISPLAYS EXTRACTED DATA AT THE EXTRACTION LOCATION ON THE DOCUMENT

- Greatly speeds up review
- Configurable properties include color, location
- Can disable for simple extraction cases or documents.

James K. Griffith
CATTARAUGUS COUNTY CLERK
Cattaraugus County Center 303 Court Street
Little Valley, NY 14755
(716) 938-9111
Fax: (716) 938-2773

Instrument Number
124386-004

No. of Pages: 4
Delivered By: SNYDER BROTHERS INC

Receipt No. 124386
Return To:
SNYDER BROTHERS INC
P O BOX 1022
KITTANNING PA 16201
ENVELOPE (4)

DATE: 07/16/2009
Time: 03:07 PM

Document Type: LEASE

Parties To Transaction: LEASE
G-PUTT TO SNYDER BROTHERS INC

Town/City: AL - Allegany

<u>Deed Information</u>	<u>Mortgage Information</u>
Consideration: \$0.00	Mortgage Amount:
Transfer Tax: \$0.00	Basic Mtge. Tax:
RETT No: 03172	Special Mtge. Tax:
	Additional Mtge. Tax:
State of New York Cattaraugus County Clerk	Mortgage Serial No.:

This sheet constitutes the Clerk endorsement required by Section 316-A(5) & Section 319 of the Real Property Law of the State of New York. DO NOT DETACH

Activate Windows



CODE EXPRESSIONS: LINQ AND STRING INTERPOLATION SUPPORT IN EXPRESSIONS

DO MORE WITH EXPRESSIONS

- .NET string interpolation now supported in expressions
- LINQ now supported in expressions

Expressions	
Default Value Expression	<code>\$"Today is {DateTime.Today}"</code>
Generated Default Value	Today is 1/27/2020 12:00:00 AM
Calculate Expression	
Validate Expression	