Document to Structure History of Changes

You might want to also see the changes in Name to Structure, since they also affect Document to Structure.

November 29th, 2019: Document Annotator 19.25

Bug fixes

- PDF documents preprocessed by ABBY OCR were appearing as blurry.

June 24th, 2019: Document Annotator 19.14

Bug fixes

- Some CDX and MOL files linked from patent XML were ignored.
- When annotating patent XML, the size of some images had an invalid unit.
- Some annotated patent XML contained incorrect HTML markup for chemical structures.

June 3rd, 2019: Document to Structure 19.12

Bug fixes

- When TextToStructure reaches a timeout, processing would sometimes continue in the background, still using some CPU and memory resources longer than necessary.

April 8th, 2019: Document Annotator 19.9

Improvements

- OCR is not performed anymore when text processing is disabled in the Document to Structure options.

Bug fixes

- Some OSR hits were displayed on the wrong page of PDF documents in version 19.8.

March 8th, 2019: Document Annotator 19.7

Improvements

- Optical Structure Recognition is now supported on PDF documents with structures drawn using vector graphics.

Bug fixes

- Invalid HTML was generated in rare cases.

December 14th, 2018: Document Annotator 18.30

Bug fixes

- Document annotator could abort on some rare situations.

November 12th, 2018: Document to Structure 18.26

Improvements

- Names can now be recognized inside double brackets.

Bug fixes

- Name with multiple ending character after them had an incorrect "end" position property.

September 24th, 2018: Document to Structure 18.23

Bug fixes

- Temporary files were sometimes left over when using the Imago OSR tool.
- Some images where not OCR'ed in PDF documents that have already been partially OCR'ed.

August 21st, 2018: Document to Structure 18.20

Bug fixes

- OCR was not performed on some scanned PDF documents.

January 12th, 2018: Document to Structure 18.1

Improvements

- The processing of images in the JBIG2 format in PDF documents for OCR and OSR has been re-enabled.

Bug fixes

- Asian names with Unicode superscripts and Zero Width Space characters were not recognized.

October 13th, 2017: Document to Structure 17.25

Bug fixes

- The timeout option was ignored when using the chemaxon.naming.document. TextToStructure API.

September 22nd, 2017: Document to Structure 17.24

Bug fixes

- Processing was getting stuck in the presence of some names with a very large number of brackets.

September 14th, 2017: Document to Structure 17.23

Bug fixes

- The processing of images in the JBIG2 format in PDF documents for OCR and OSR has been temporarily disabled. It will be re-enabled by default once we require Java 8, which we expect to do in a few months time. In the mean time, if this feature is desired, a Java ImageIO JBIG2 plugin can be added to the classpath to enable this feature. Please contact us if you have any issue with this.

July 21st, 2017: Document to Structure 17.16

Bug fixes

- Some incorrect SMILES strings were detected.

July 6th, 2017: Document to Structure 17.14

Improvements

- Complex IUPAC names are now detected even inside a bracketed subsentence.

June 20th, 2017: Document to Structure 17.13

Improvements

- The new version of method TextToStructure.extract accepts a DocumentToStructureOptions instead of an option string, allowing higher-level usage and faster processing.

June 13th, 2017: Document to Structure 17.12

Improvements

- Processing of short text using chemaxon.naming.document.TextToStructure is about 10% faster.
- Class chemaxon.naming.document.D2S.Options is renamed to chemaxon.naming.document.DocumentToStructureOptions.

June 8th, 2017: Document to Structure 17.11

Improvements

- When processing short XML fragments, passing the option content-type=text/xml now leads to higher performance by skipping format detection entirely.

Bug fixes

- Option content-type was ignored in chemaxon.naming.document.TextToStructure

Document to Structure 17.03.13

New features

- The "Preparation N" identifier is now automatically detected in patents, in additions to exemplified compounds.

Document to Structure 17.02.20

Improvements

- PDF documents OCRed by PDF-XChange are not OCRed a second time, avoiding duplicate results and resulting in much faster processing.

Document to Structure 17.01.30

Bug fixes

- The 'end' character position property was missing or incorrect in names delimited by ":" characters.

Document to Structure 16.09.26

Improvements

- Structures mentioned in plural, for instance "pyrimidines", are now tagged with the "generic" type.
- The format of HTML documents can now be detected from the file contents only.

Document to Structure 16.08.29

Bug fixes

- The -groups option did not work for common names.

Document to Structure 16.08.08

Improvements

- The format of Microsoft Office documents can now be detected from the file contents only.

Document to Structure 16.08.01

Bug fixes

- A trailing name between brackets, often a synonym of the previous name, for instance "retinoic acid (vitamin A)", was sometimes misrecognized as a part of the previous name.

Document to Structure 16.07.25

Improvements

- Example numbers are now detected in Chinese language patents.

Bug fixes

- ? (water) is now recognized as a vernacular term in Chinese and Japanese.

Document to Structure 16.07.11

Improvements

- More syntaxes for example numbers are detected in patents.

Bug fixes

- A spurious warning was logged when recognizing an InChI string in a document.

Document to Structure 16.05.02

Bug fixes

- Some Japanese names were not recognized without the Chinese Name to Structure license.

Document to Structure 16.04.11

Bug fixes

- Some Asian names were wrongly starting with a closing bracket.

Document to Structure 16.03.07

Improvements

- The detection of Asian names in documents is improved.

Bug fixes

- Some invalid structures returned by an OSR tool could interrupt document to structure. They are now simply logged and processing continues.

Document to Structure 16.01.18

Bug fixes

- Asian names followed by some number formats were not recognized.

Document to Structure 16.01.04

Improvements

- Some names with spurious whitespace are better detected in documents.

Document to Structure 15.12.14

Improvements

- When detecting exemplified compound numbers in patents, more syntaxes are now detected.

Document to Structure 15.10.26

Bug fixes

- When processing Office documents with Optical Structure Recognition enabled, the following log message was issued: WARNING: Skipping non-existing image.

Document to Structure 15.09.07

Bug fixes

- The warning "JBIG2ReadParam not specified. Default will be used.", which appeared in the logs, has been fixed.

Document to Structure 15.07.20

Improvements

- The detection of asian names is improved in some documents.

Document to Structure 15.06.29

Bug fixes

- Processing of scanned PDFs (using OCR - Optical Character Recognition) was failing on Mac OS X.

Document to Structure 15.06.01

Improvements

- Names that contain extra spaces (for instance because of OCR or new lines without a - character before the break) are now better supported.
- Patent PDFs provided by LexisNexis's Univentio can now be processed using the built in OCR text instead of doing the OCR again, leading to about 10 times faster processing. This was already working for older Univentio PDFs but is now also supported for recent ones.

Document to Structure 15.05.25

Improvements

- Improvements in OCR error correction.

Document to Structure 15.05.18

Improvements

- Significant improvements are included in OCR error correction, especially for patents based on scanned images.

Document to Structure 15.05.04

New features

- Exemplified compound numbers are now extracted automatically when they are mentioned directly before the IUPAC name.

Improvements

- Improved detection of asian names in documents.

Document Annotator 15.04.20

Bug fixes

- Some TIFF images failed to be processed.

Document to Structure 15.03.09

Bug fixes

- Certain "custom" image types in PDFs failed to be processed by OSR tools.

Document to Structure 15.02.23

New features

- The location of OSR tools can now be specified by using Java system properties: chemaxon.naming.clide.path, chemaxon.naming.osra.path and chemaxon.naming.imago.path. These take precedence over the environment variables (CLIDE, OSRA and IMAGO), which are also supported.

Improvements

- The detection of Japanese names in documents has been improved.

Document to Structure 15.01.26

Improvements

- In the output of Optical Structure Recognizers, some non-standard labels such as X2 are now interpreted as R-groups. Those structures also become representable as SMILES.
- Some aliases detected by Optical Structure Recognition programs are expanded to the corresponding chemical group.

Document to Structure 15.01.19

Improvements

- Encrypted documents are now reported with a clearer message and a specific single exception: chemaxon.naming.document.EncryptedDocumentException

Bug fixes

- Failure to start OCR could lead to multiple entries in the log.

Document to Structure 15.01.12

Bug fixes

- A failure in writing temporary images to disk lead to duplicate entries in the log.

Document to Structure 15.01.05

Improvements

- Unknown format options now lead to a failure instead of being ignored.
- Numbers in subscript and superscript HTML and XML tags (<sub> and <sup>) are now interpreted.

Bug fixes

- Japanese and Chinese names were not recognized in documents when followed by a fullwidth number inside brackets.

Document to Structure 14.12.15

Bug fixes

- Assignee metadata was not extracted from some recent USPTO XML patents.

Document to Structure 14.12.01

Improvements

- Numeric character entities in metadata of XML documents are now decoded.

Document to Structure 14.11.24

Bug fixes

- Some SMILES-like words containing "Br" or "Cl" were wrongly detected as SMILES.

Document to Structure 14.11.17

New Feature

- The structures recognized by OSR tools (CLiDE and OSRA) are now represented in a uniform way, for instance using real R-group atoms instead of just aliases.

Bug fixes

- Some OLE structures embedded by JChem for Office were extracted as empty structures.

Document to Structure 14.09.29

Bug fixes

- A failure to process some Powerpoint documents was fixed.

Document to Structure 14.09.22

Improvements

- Logging messages are improved to facilitate investigating problematic documents.

Document to Structure 14.09.15

Improvements

- Detection of Chinese chemical names in documents was improved.
- Names (especially Chinese) are better detected in XML patent files from IFI Claims.

September 8th, 2014: Document to Structure 14.09.08

Bug fixes

- The molecule name extracted from Chinese and Japanese documents was in a few cases not the actual name but another close-by word.
- Some invalid structures were generated from some chinese and japanese documents with unusual character combinations.

September 1st, 2014: Document to Structure 14.09.01

New Features and Improvements

- The embedded OCR information in PDFs from the TotalPatent system (by LexisNexis) is now detected and used automatically. This leads to massively faster processing for such PDFs, from 10 to 30 times faster!

August 25th, 2014: Document to Structure 14.8.25

New Features and Improvements

- Names split between two paragraphs can now be detected and converted to structure. While ideally this situation should not occur in proper semantic documents, it does occur in automatically generated ones, for instance in patent documents distributed by patent providers.
- The context field can now include text from a previous paragraph. This is useful in particular in patents, where this sometimes includes the example number.

Bug fixes

- Google patent PDFs were only processed until page 20.

August 18th, 2014: Document to Structure 14.8.18

Bug fixes

- Processing an HTML file with CLiDE enabled failed when an image contained no structure.

August 4th, 2014: Document to Structure 14.8.4

New Features and Improvements

- Added the osrResolution=N option to specify the resolution in DPI of images processed by OSRA. This is only needed for single image files, not for images inside PDF, Office or HTML documents.

Bug fixes

- The character encoding of HTML5 documents was not detected automatically.

July 7th, 2014: Document to Structure 14.7.7

New Features and Improvements

- Chinese and Japanese name are better detected in documents.
- Optical Structure Recognition is now performed on images referenced in HTML documents. This is supported when the image is accessible as a file (relatively to the HTML document file), and when the image is represented by a data URI.
- The BMP image format is now supported for calling Optical Structure Recognition tools using the document to structure API.
- The ending position of the names in text documents (HTML, XML, TXT) is now also included in the properties of returned structures.

Bug fixes

- Some embedded structures in Powerpoint documents were not extracted.

April 18th, 2014: Document to Structure 6.3.0

New Features and Improvements

- OSRA 2.0 is supported.
- A format option has been introduced allowing the user to select an optical structure recognition tool: CLiDE, OSRA or Imago when more than one is installed on the computer. For instance, `d2s:osra` will request OSRA to be used.

Bug fixes

- The encoding of HTML files was not always detected from the META tag.
- When the extraction of OLE embedded structures from Office documents was disabled using the `d2s:-ole` format option, the optical structure recognition of images was disabled as well.
- The character position field of structures was sometimes higher by a few characters when extracted from HTML documents with CRLF (\r\n) line endings.
- The CLiDE optical structure recognition tool was not automatically detected on 64 bit versions of Windows.
- When using OSRA, some structures with implicit hydrogens were wrongly filtered out.
- After processing only a part of an HTML document, the processing of the next HTML document could give some incorrect results.