Indexing chemical names and structures from documents: putting it all together

Daniel Bonniot de Ruisselet
ChemAxon

ChemAxon US UGM
September 26th 2012

ChemAxon's Naming Technology

- Structure to Name
- Name to Structure
- Document to Structure
- Document to Database

ChemAxon's Naming Technology

- Structure to Name
  - IUPAC Name, traditional names
  - Mature
  - Still upcoming: peptides, some fused systems
- Name to Structure
  - IUPAC, CAS and systematic names
  - Dictionary of common names and drug names
  - Support CAS Registry number (webservice)
  - Homology group: alkyl, aryl …
  - Future: Biological names, polymers, ...
- Accuracy and coverage constantly improving (...)
- Available from GUI, API and command-line

Name to Structure internals

- Dictionary of common and drug names
  - Uses nine different source dictionaries
  - Harmonized using weighted consensus method
  - 150K names for 40K unique structures
  - Custom dictionary and plug-in system
- Systematic names
  - Proprietary algorithm
  - ”Understands” systematic names
  - Example:

(2R)-2-methylsulfanyl-3-hydroxybutanedioate

Systematic name example

(2R)-2-methylsulfanyl-3-hydroxybutanedioate

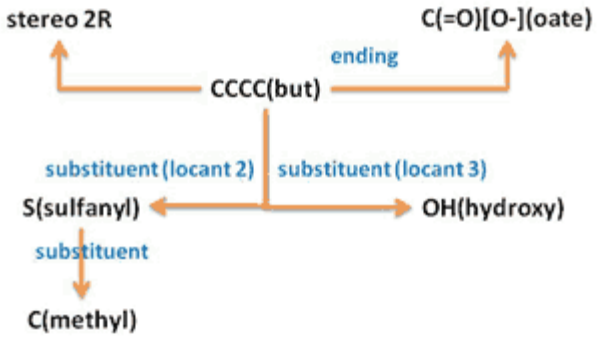
Systematic name: tokenization

(2R)-2-methylsulfanyl-3-hydroxybutanedioate

(2R)	Stereo Locant list
2	Locant list
meth	Structure token: C
yl	Suffix
sulfan	Structure token: S
yl	Suffix
3	Locant list
hydroxy	Structure token: O
but	Structure token: CCCC
ane	Suffix
di	Multiplier
oate	Suffix: C(=O)[O-]

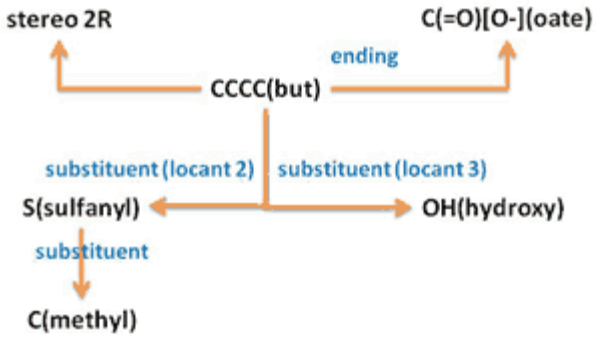
Systematic name: parsing

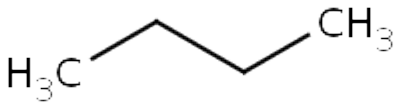
(2R)-2-methylsulfanyl-3-hydroxybutanedioate



Systematic name: parsing

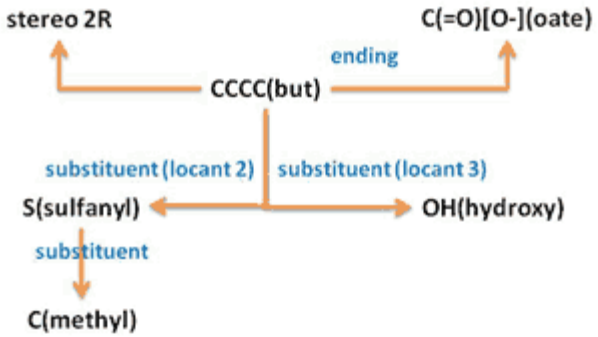
(2R)-2-methylsulfanyl-3-hydroxybutanedioate

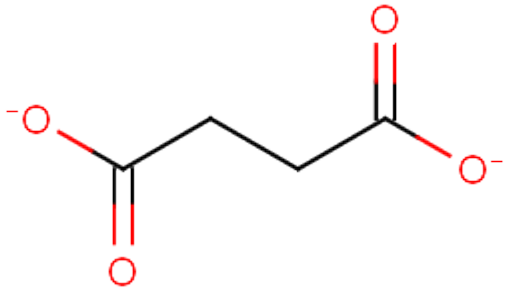




Systematic name: parsing

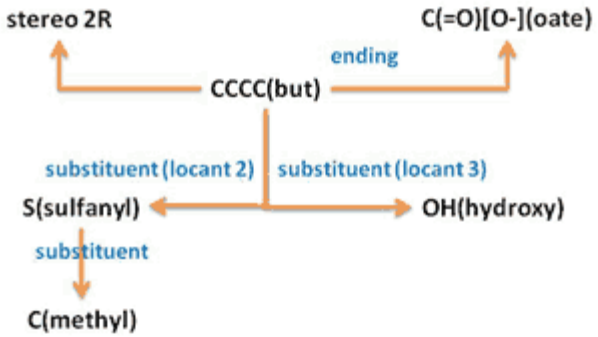
(2R)-2-methylsulfanyl-3-hydroxybutanedioate

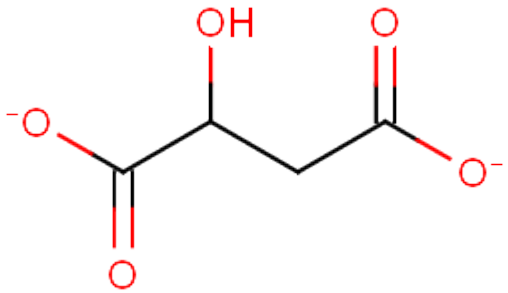




Systematic name: parsing

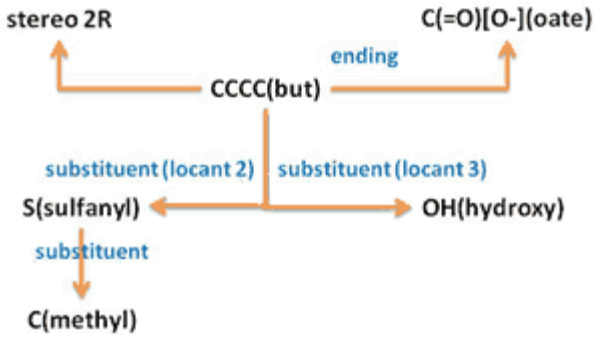
(2R)-2-methylsulfanyl-3-hydroxybutanedioate

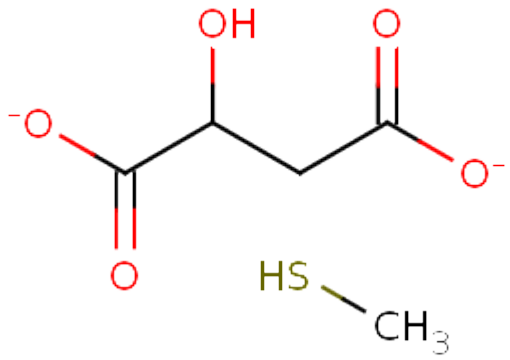




Systematic name: parsing

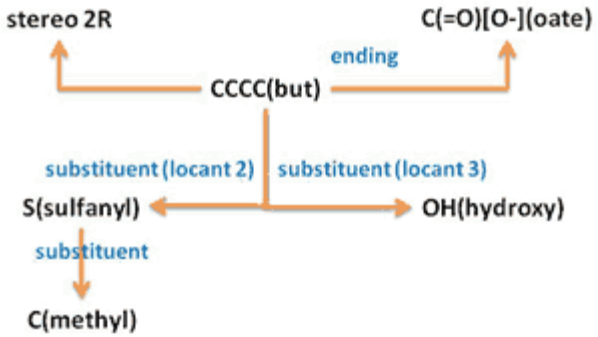
(2R)-2-methylsulfanyl-3-hydroxybutanedioate

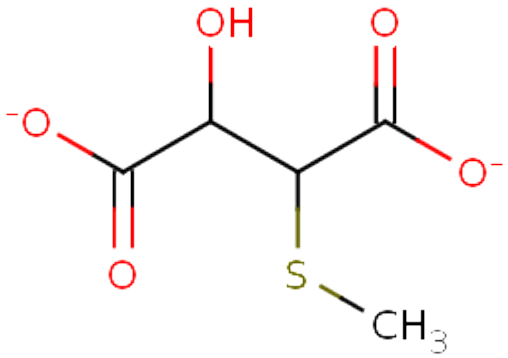




Systematic name: parsing

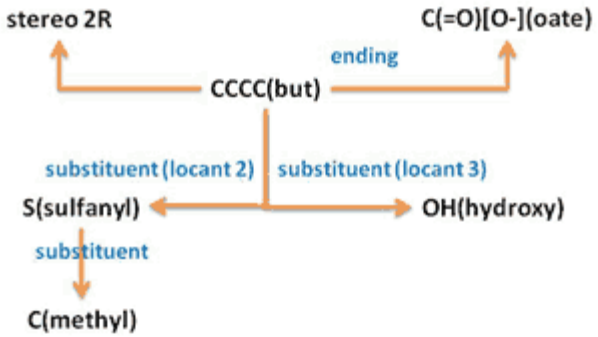
(2R)-2-methylsulfanyl-3-hydroxybutanedioate

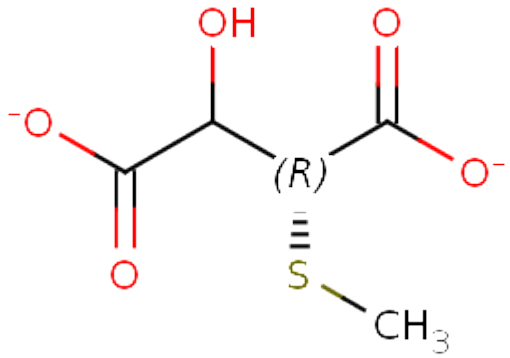




Systematic name: structure generation

(2R)-2-methylsulfanyl-3-hydroxybutanedioate





OCR error correction

(2R)-2-rnethylsulfany1-3-hydr0xybutanedi0ate

OCR error correction

(2R)-2-rnethylsulfany1-3-hydr0xybutanedi0ate

(2R)-2-methylsulfanyl-3-hydroxybutanedioate

OCR error correction

(2R)-2-rnethylsulfany1-3-hydr0xybutanedi0ate

(2R)-2-methylsulfanyl-3-hydroxybutanedioate

Ar-benzyl-Ar-[3-(lH-tetrazol-5-yl)phenyl]propanamide

?-benzyl-?-[3-(?H-tetrazol-5-yl)phenyl]propanamide

OCR error correction

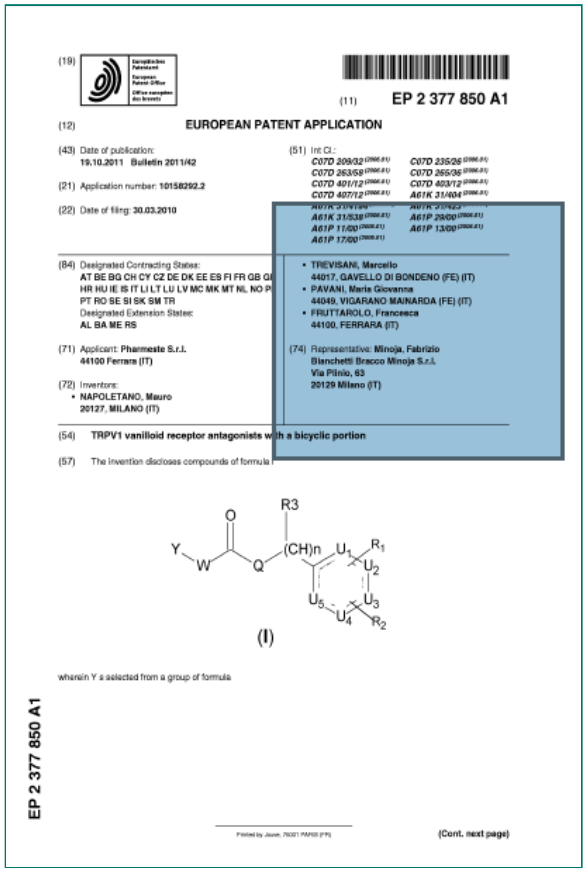
(2R)-2-rnethylsulfany1-3-hydr0xybutanedi0ate

(2R)-2-methylsulfanyl-3-hydroxybutanedioate

Ar-benzyl-Ar-[3-(lH-tetrazol-5-yl)phenyl]propanamide

***N*-benzyl-*N*-[3-(1H-tetrazol-5-yl)phenyl]propanamide**

From Document to Structures



(19) Europäisches Patentamt European Patent Office Office européen des brevets

(11) EP 2 377 850 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication: 19.10.2011 Bulletin 2011/42

(51) Int Cl.: C07D 209/32 (2006.01) C07D 235/26 (2006.01) C07D 263/58 (2006.01) C07D 265/36 (2006.01) C07D 401/12 (2006.01) C07D 403/12 (2006.01) C07D 407/12 (2006.01) A61K 31/404 (2006.01) A61K 31/538 (2006.01) A61P 25/00 (2006.01) A61P 11/00 (2006.01) A61P 13/00 (2006.01) A61P 17/00 (2006.01)

(21) Application number: 10158292.2

(22) Date of filing: 30.03.2010

(84) Designated Contracting States: AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO SE SI SK SM TR
Designated Extension States: AL BA ME RS

(71) Applicant: Pharmeste S.r.l. 44100 Ferrara (IT)

(72) Inventors:
- NAPOLETANO, Mauro 20127, MILANO (IT)
- TREVISANI, Marcello 44017, GAVELLO DI BONDENO (FE) (IT)
- PAVANI, Maria Giovanna 44049, VIGARANO MAINARDA (FE) (IT)
- FRUTTAROLO, Francesca 44100, FERRARA (IT)

(74) Representative: Minoja, Fabrizio Bianchetti Bracco Minoja S.r.l. Via Plinio, 63 20129 Milano (IT)

(54) TRPV1 vanilloid receptor antagonists with a bicyclic portion

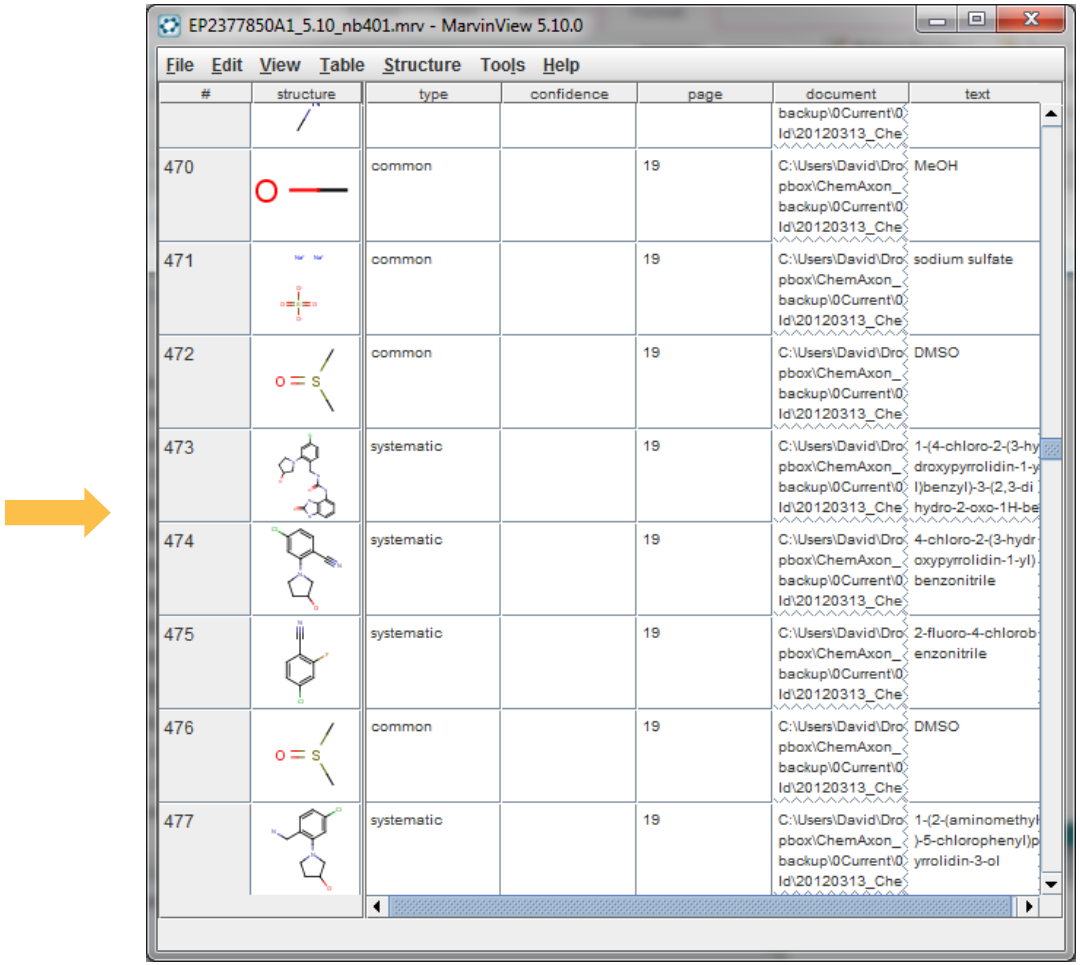
(57) The invention discloses compounds of formula

(I)

wherein Y a selected from a group of formula

EP 2 377 850 A1

(Cont. next page)



Non-searchable patent (50 pages)

Structure (text + image) + location

ChemAxon

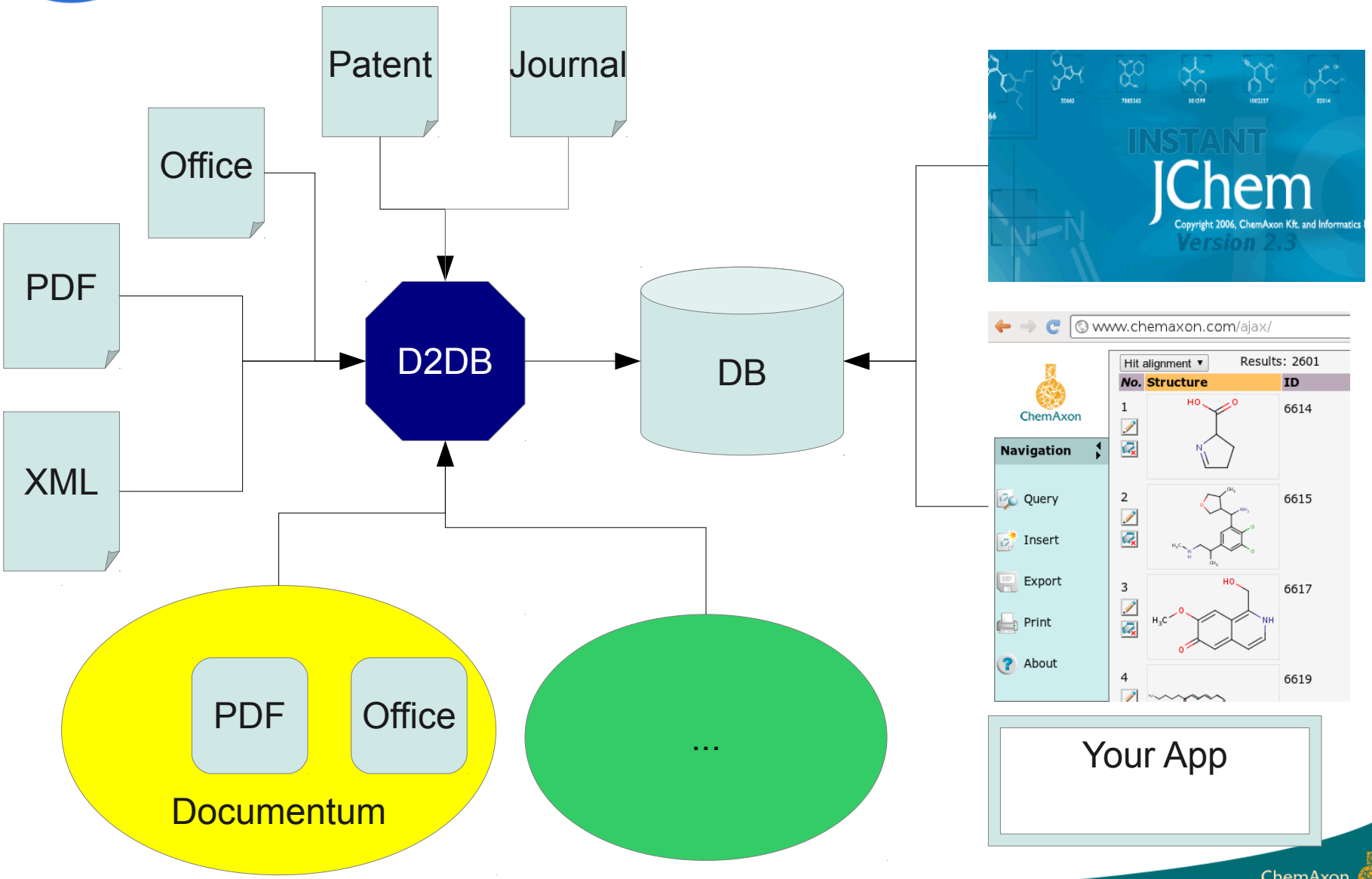
ChemAxon's "Document to Structure"

- Extract chemical information from documents
  - Names: powered by the Naming Technology
  - Also import SMILES, InChI, CAS number …
  - Images: OSRA, ...
  - Works with **scanned non-searchable PDF**
  - Returns structures and their **location** in the document

ChemAxon's "Document to Structure"

- Supported formats:
  - MS Office document: doc, docx, ppt, pptx, xls, xlsx, odt …
  - Embedded structure objects (ChemDraw, Symyx, Marvin, …)
  - PDF, text, XML, HTML

“Document to Database”



ChemAxon's "Document to Database"

- Data in DB:
  - Structures
  - Source (name, smiles, embedded, …) and location
  - Documents, Authors, Metadata...
- Questions:
  - What structures appear in a specific document?
  - What documents contain a structure/substructure/...?
  - What documents written since 2010 in location X contain substructure S?
  - ...

ChemAxon's "Document to Database"

- Customizable:
  - Metadata extracted from documents
  - Interface (IJC forms, webapp)
- Demo:
  - One month of US patents
  - 85K unique structures from systematic names
  - 1M occurrences

Summary

- Extensive, improving naming technologies (n2s, s2n)
- Increasing support for Document mining (d2s, d2db, SharePoint)
- Putting it all together → going large scale, extract and use valuable chemical information
- Still component-based and responding to our users requests