Extracting Chemical Information within Documents - from Desktop to Enterprise

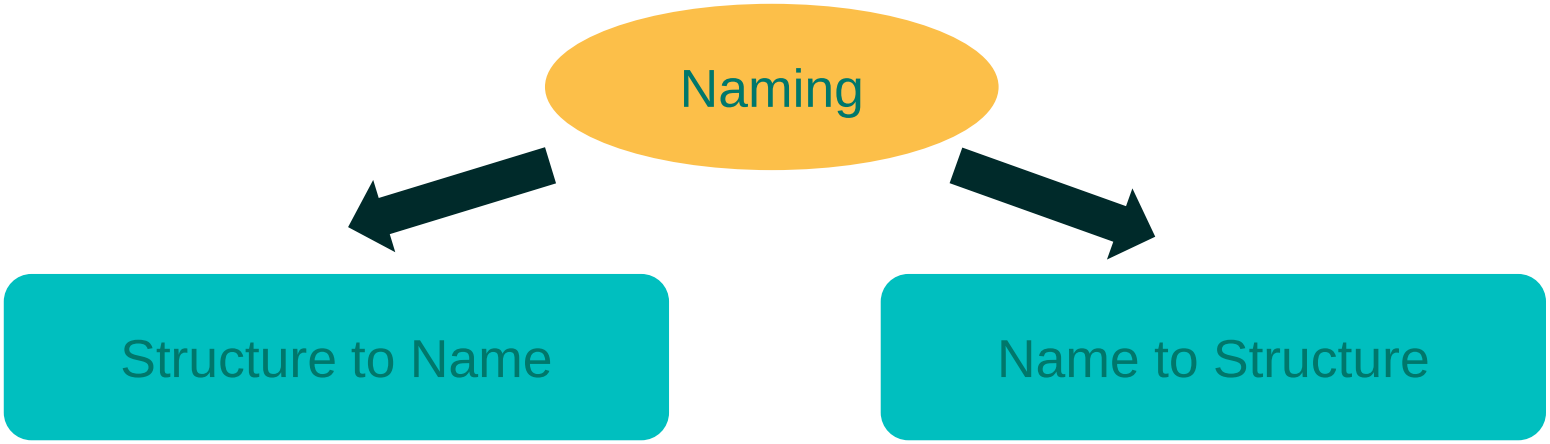
Wei Deng (David), Daniel Bonniot

ChemAxon US UGM
Sep 2013

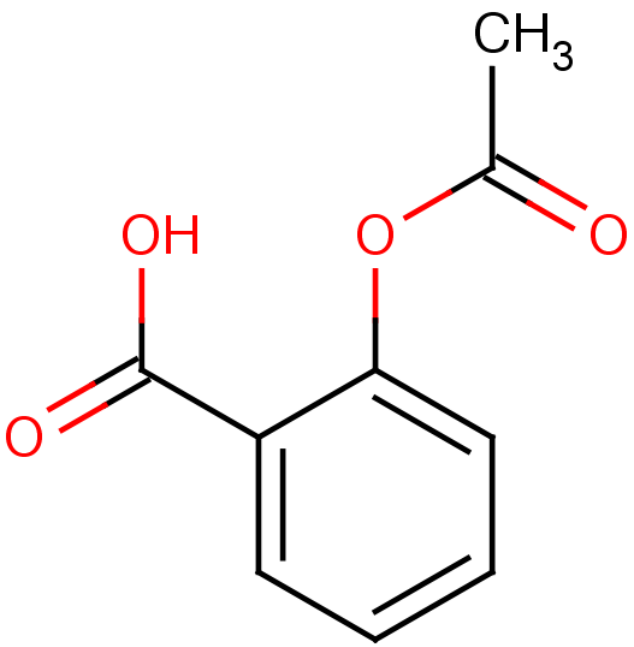
What is *Naming*

- Reliable chemical structure and name conversion
- Available in all ChemAxon products
- Can be used in various ways
- Backbone of ChemAxon's text mining tools

Backbone of Text Mining



Structure to “Name"



S2N

- 2-(acetyloxy) benzoic acid
- Aspirin
- 50-78-2
  11126-35-5
  11126-37-7
  2349-94-2
  26914-13-6
  98201-60-6

NEW

“Name” to Structure

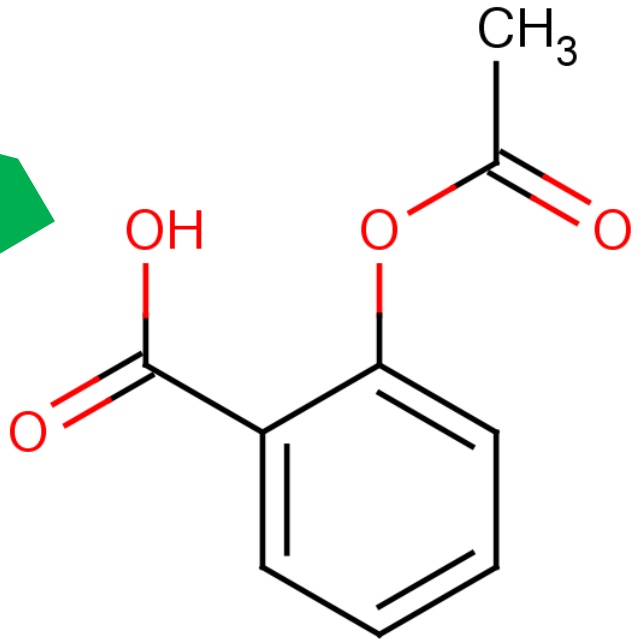
- **2-（乙酰氧基）苯甲酸**
- **阿司匹林**

- **2-(acetyloxy) benzoic acid**
- **Aspirin**
- Acetylsalicylate
- Easprin …

- 50-78-2

CN2S

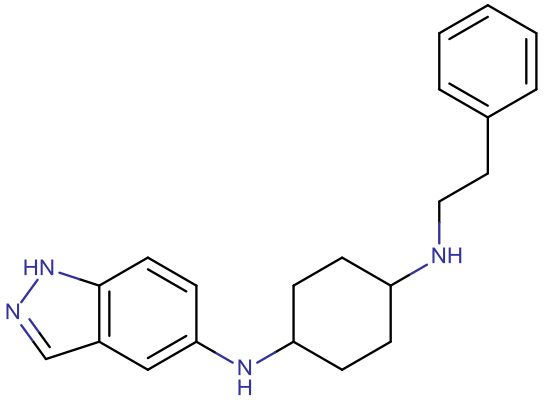
N2S



Matured yet still Improving

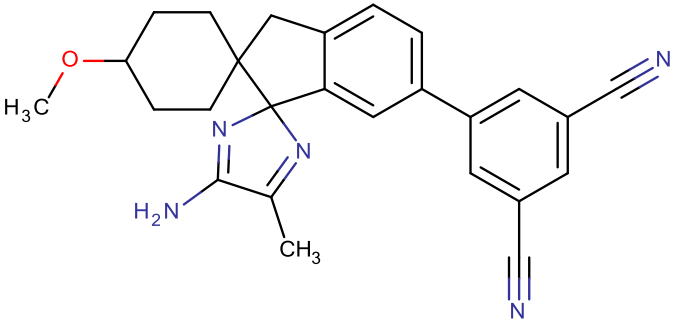
N1-(1H-5-Indazolyl)-N4-phenylethyl-1,4-cyclohexanediamine

6.1



5-(4''-amino-4-methoxy-5''-methyl-3'H-dispiro[cyclohexane-1,2'-indene-1',2''-imidazol]-6'-yl)benzene-1,3 -dicarbonitrile

6.2



What About Other Chemical Information?

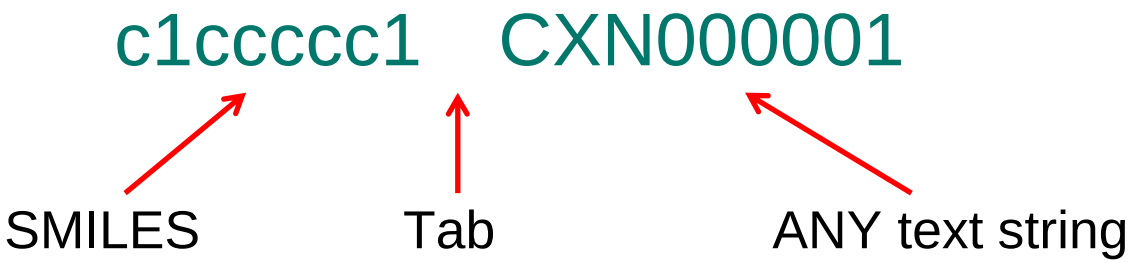
- Corporate compound ID number
- Unusual common names
- Names in other languages
- Other identifier (e.g. FICTS ID)

…

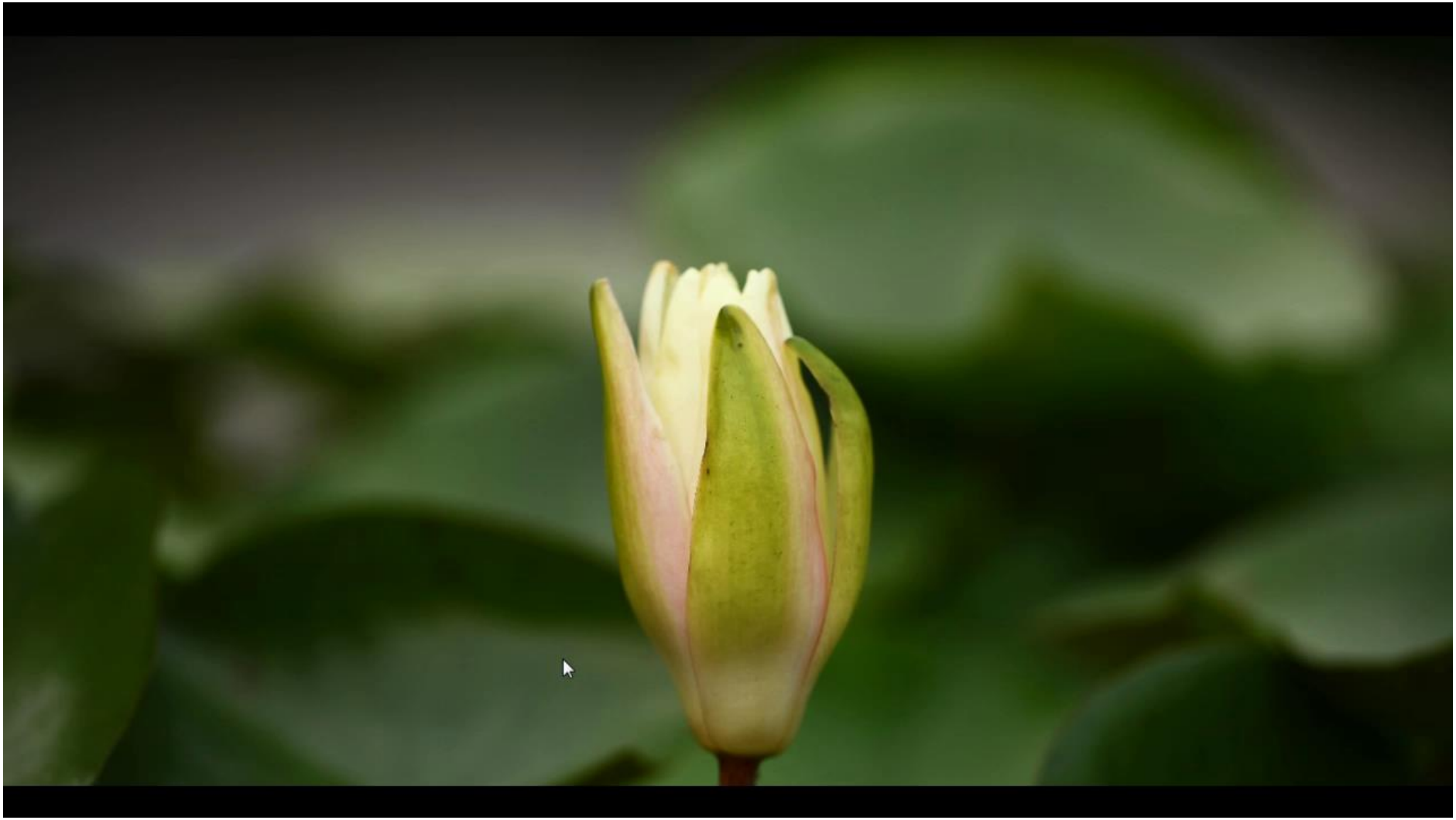
Customized Dictionary

- A SMILES file “custom_names.smi”
- Default location ChemAxon DIR
  e.g. in Windows 7 C:\Users\USERNAME\chemaxon\

- Format



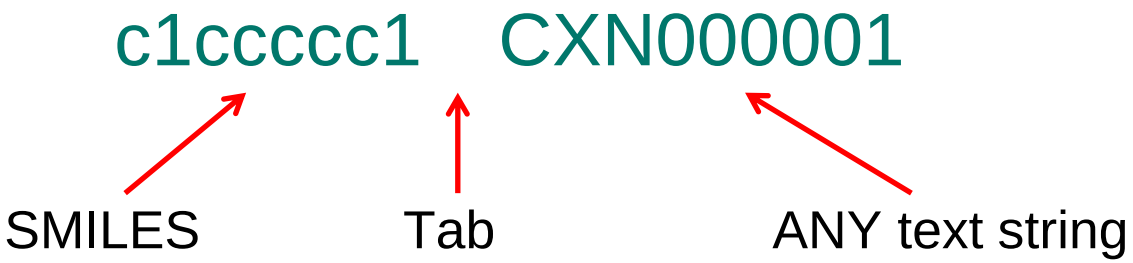
Custom Dictionary Demo



Customized Dictionary

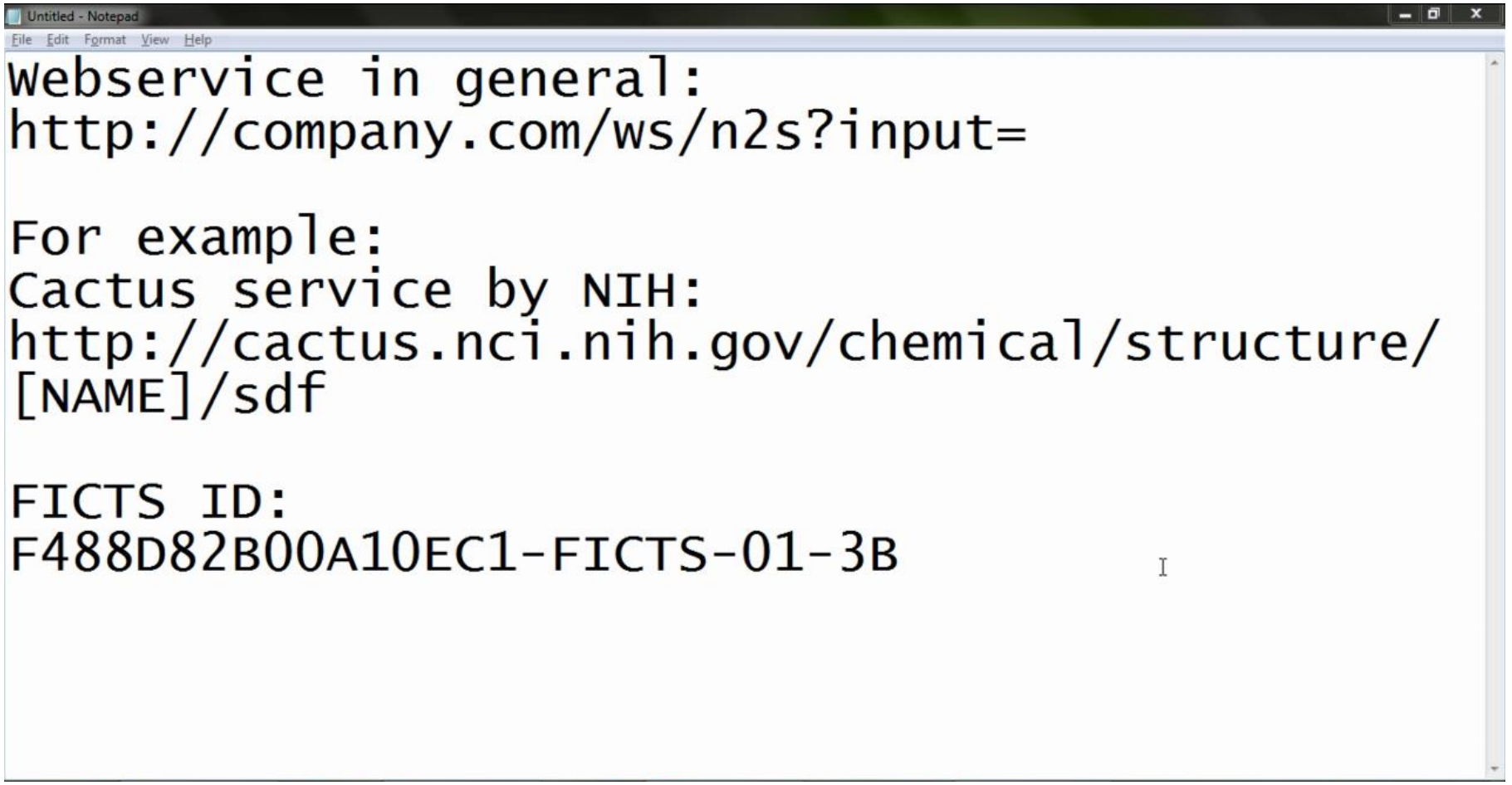
- A SMILES file “custom_names.smi”
- Default location ChemAxon DIR
  e.g. in Windows 7 C:\Users\USERNAME\chemaxon\

- Format

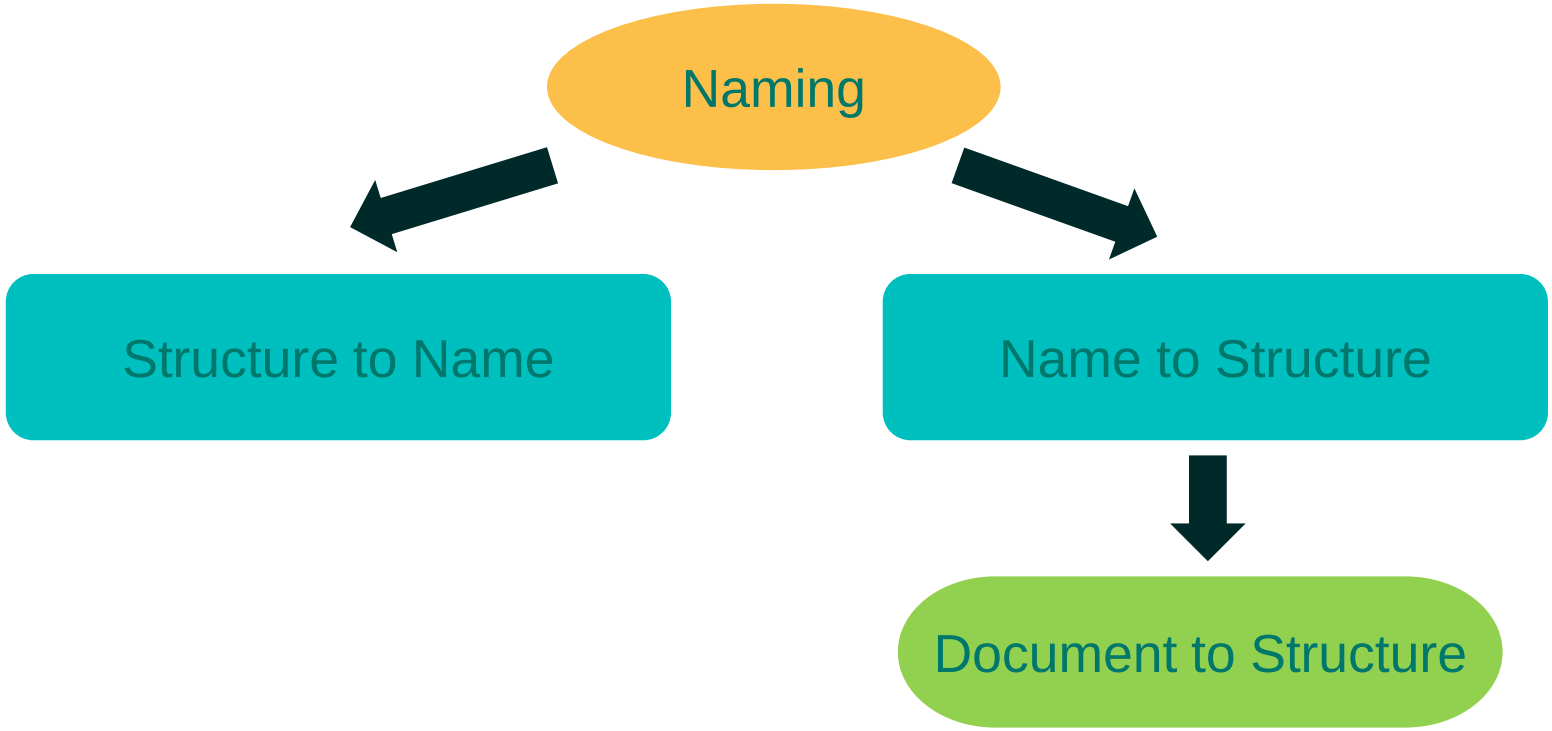


From Version 6.0, a custom **web service** can also be used

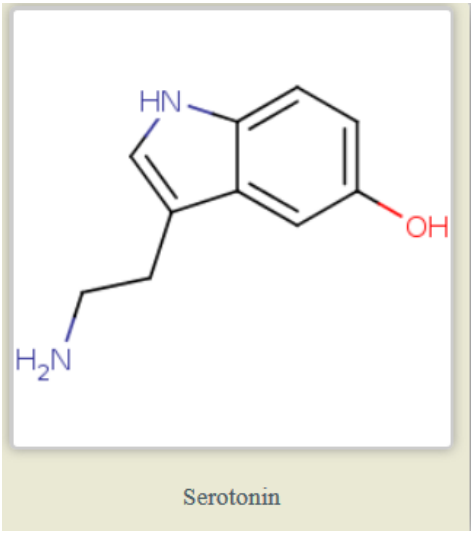
NEW *Naming* Web Service Demo



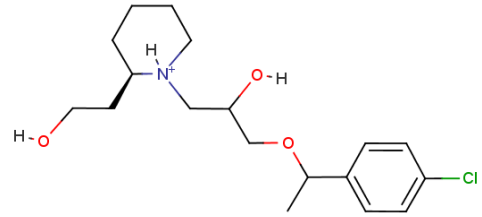
Chemical Information in Documents?



Text Chemical Information in Documents



Structure Images in Documents



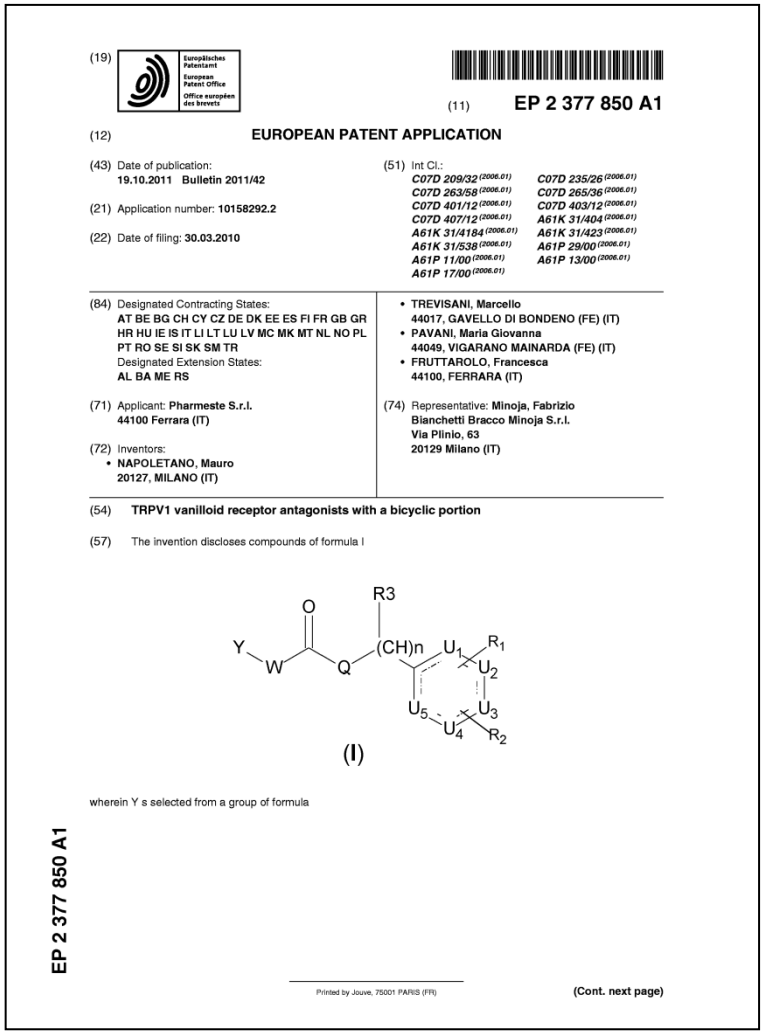
OSRA (NIH)

6.1 Supports NEW

CLiDE (Keymodule)

Imago (GGA)

Non-Text Document



(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

(21) Application number: **10158292.2**

(22) Date of filing: **30.03.2010**

(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 31/4184 (2006.01) *A61K 31/423* (2006.01)
A61K 31/538 (2006.01) *A61P 29/00* (2006.01)
A61P 11/00 (2006.01) *A61P 13/00* (2006.01)
A61P 17/00 (2006.01)

(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

(I)

wherein Y s selected from a group of formula

EP 2 377 850 A1

Printed by Jouve, 75001 PARIS (FR)

(Cont. next page)

OCR

Full Text

Automatic OCR Error Correction

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

↓

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

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

↓

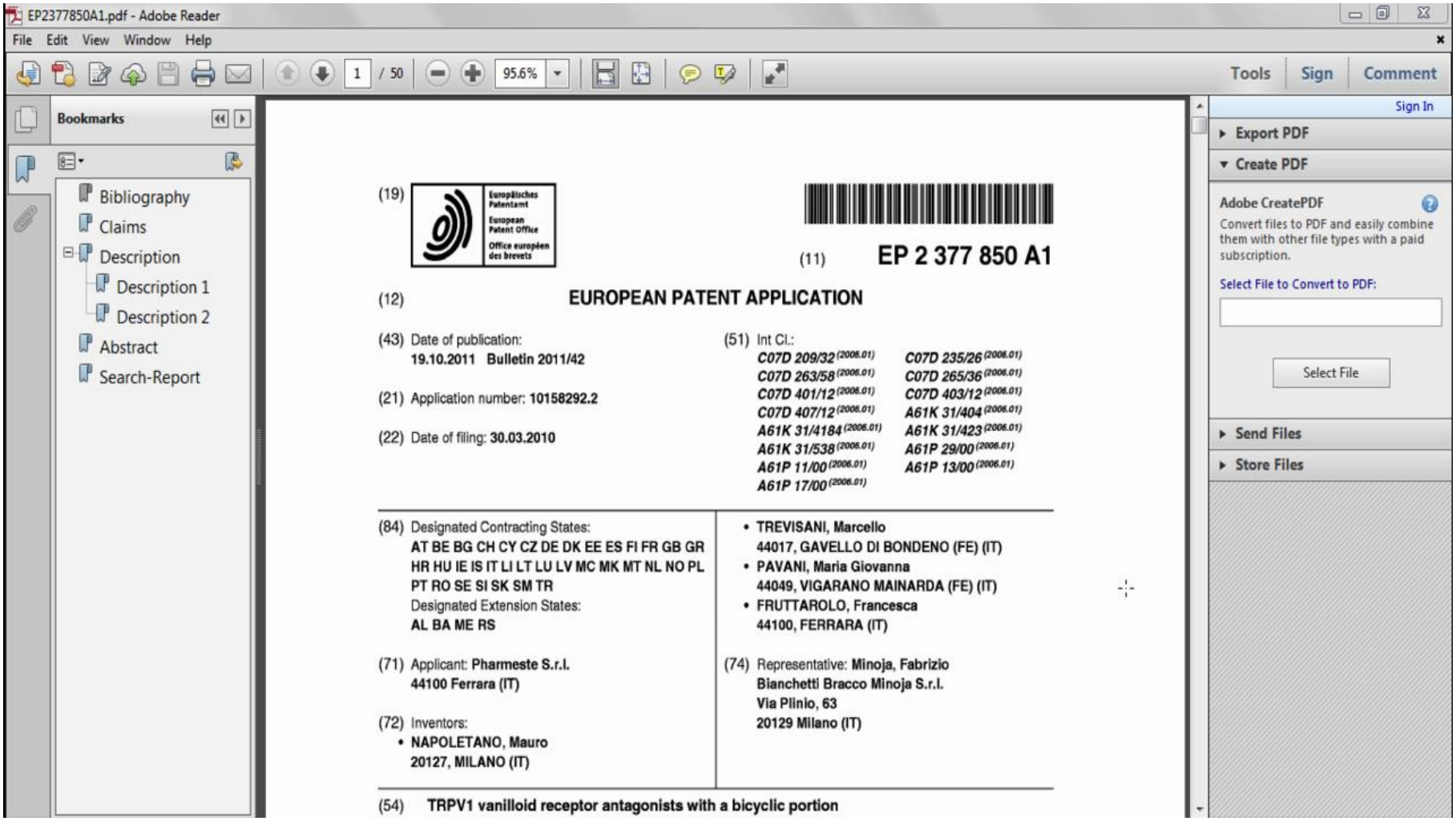
***N*-benzyl-*N*-[3-(1H-tetrazol-5-yl)phenyl]propanamide**

我们日前止在研究开友中文化字名称的**OCR**白动纠错工力能

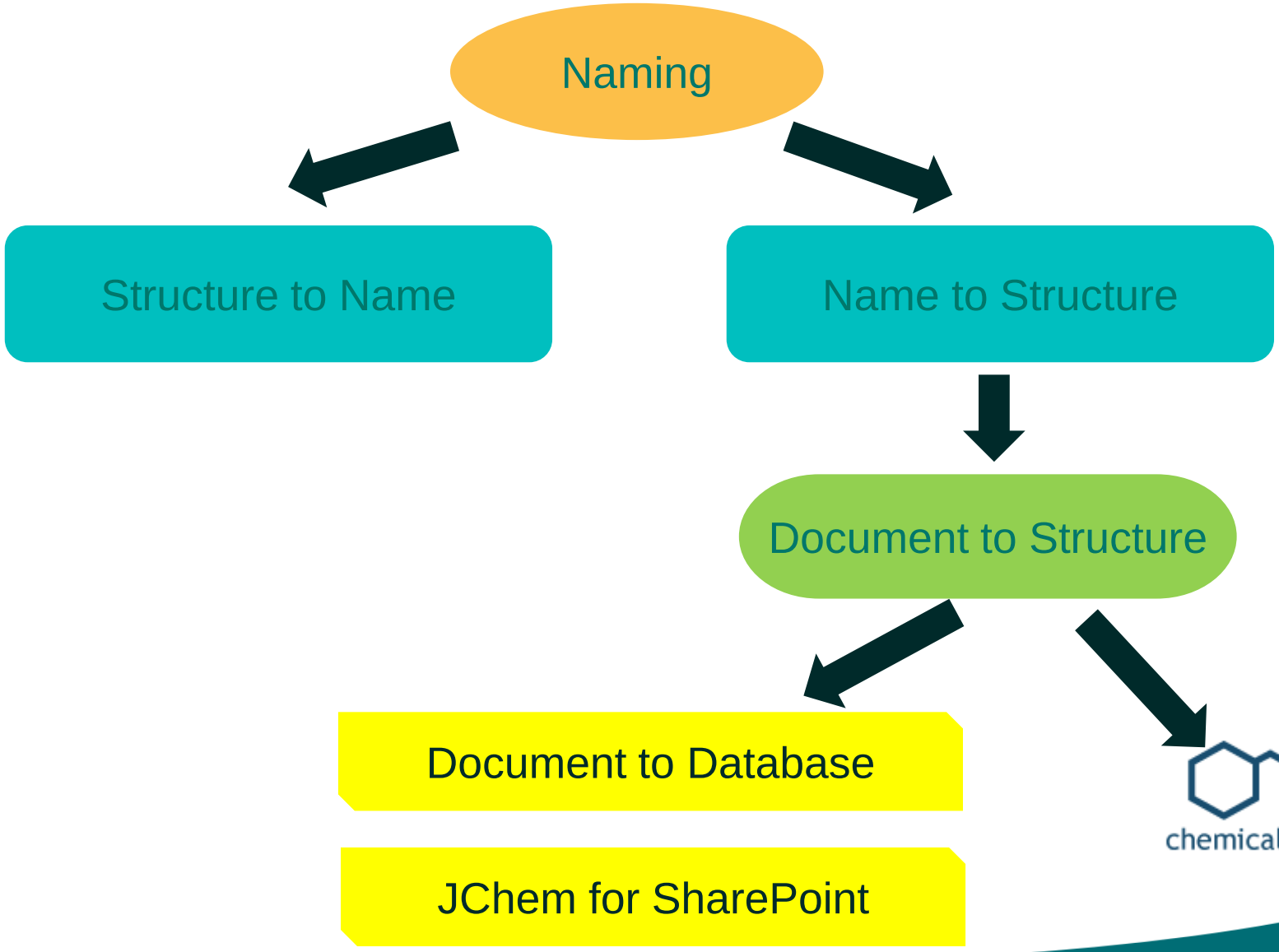
↓

我们目前正在研究开发中文化学名称的**OCR**自动纠错功能

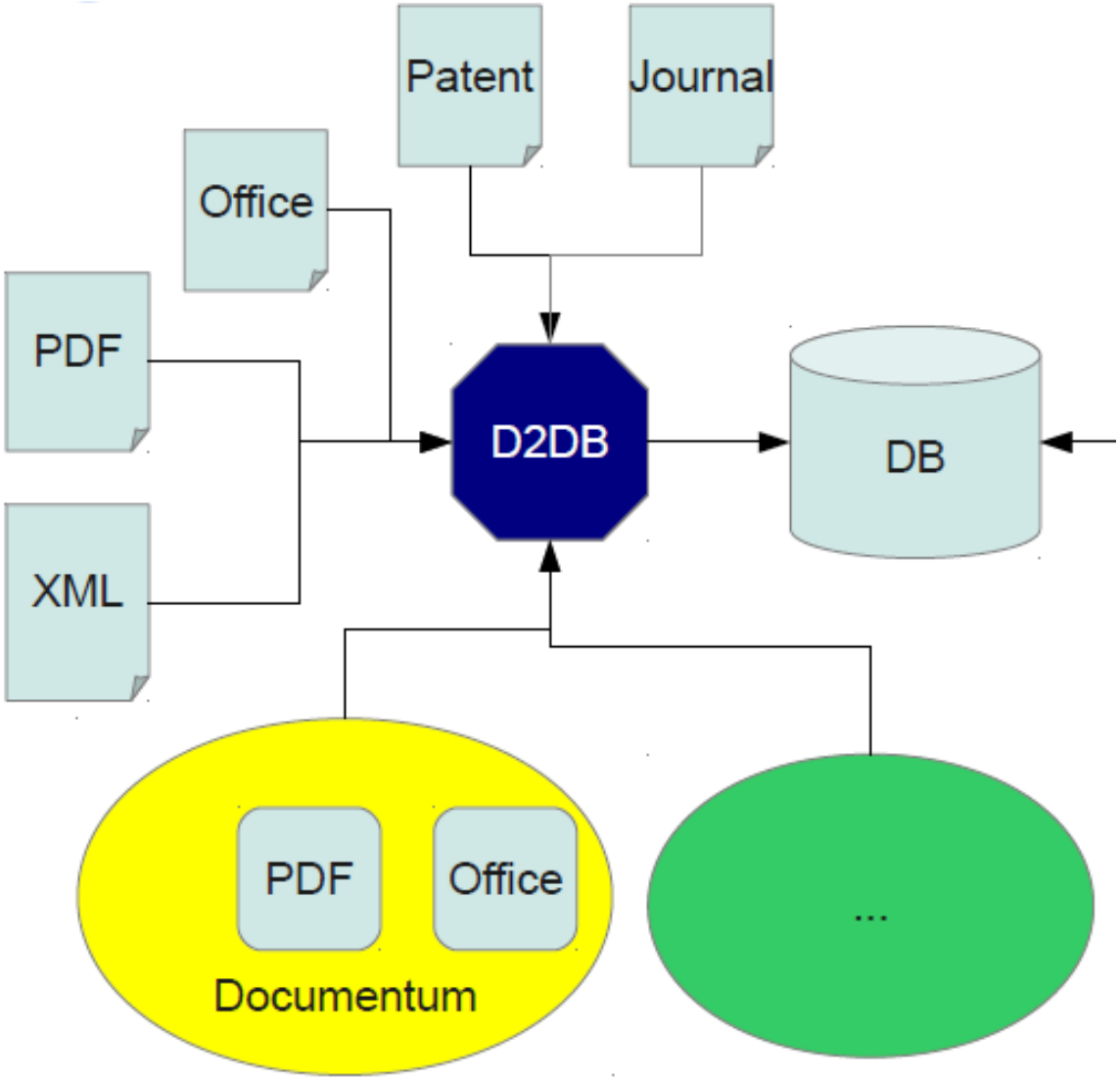
Document to Structure Demo



Too Many Documents to Manage?

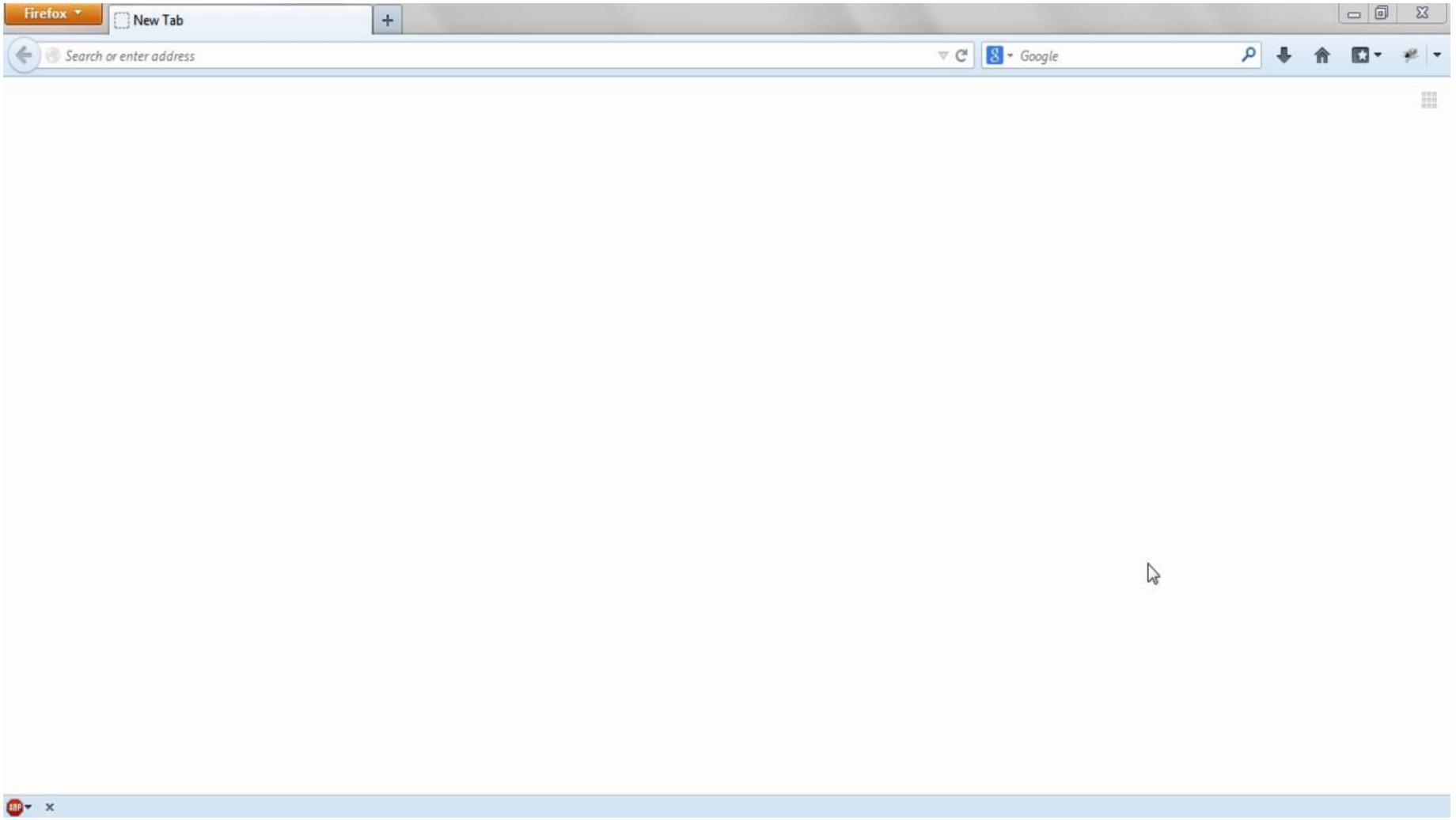


Document to Database

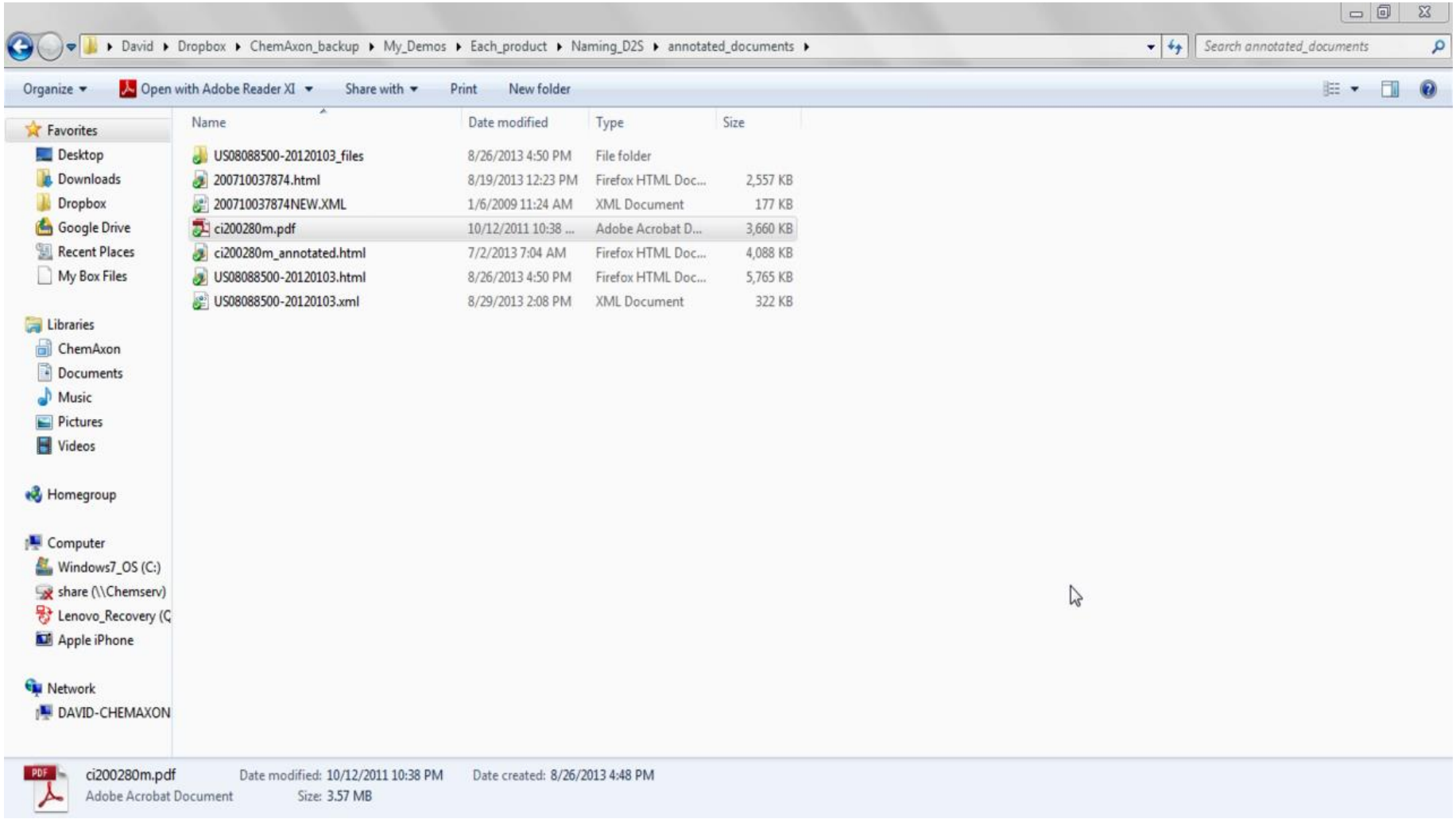


- Chemical information
- Document metadata
- Searchable
- Customizable
- Document annotation

Document to Database Demo



Document Annotation



In Summary, *Naming* Becomes

- Faster and more accurate
- Multilingual
- More versatile
- Backbone of Text Mining

What's Coming for *Naming*?

- Other language support in N2S and D2S
- D2S: MultiCore OCR/OSR
- D2DB: OpenText, Google Drive

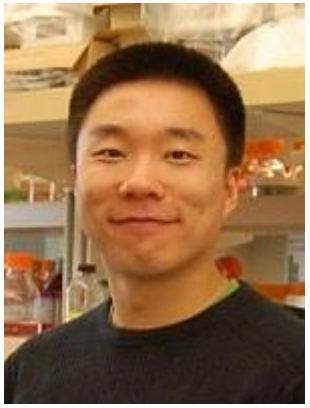
Acknowledgements

Daniel Bonniot

N2S

邓巍

CN2S



Gustavo Santucho

Mario Burdman