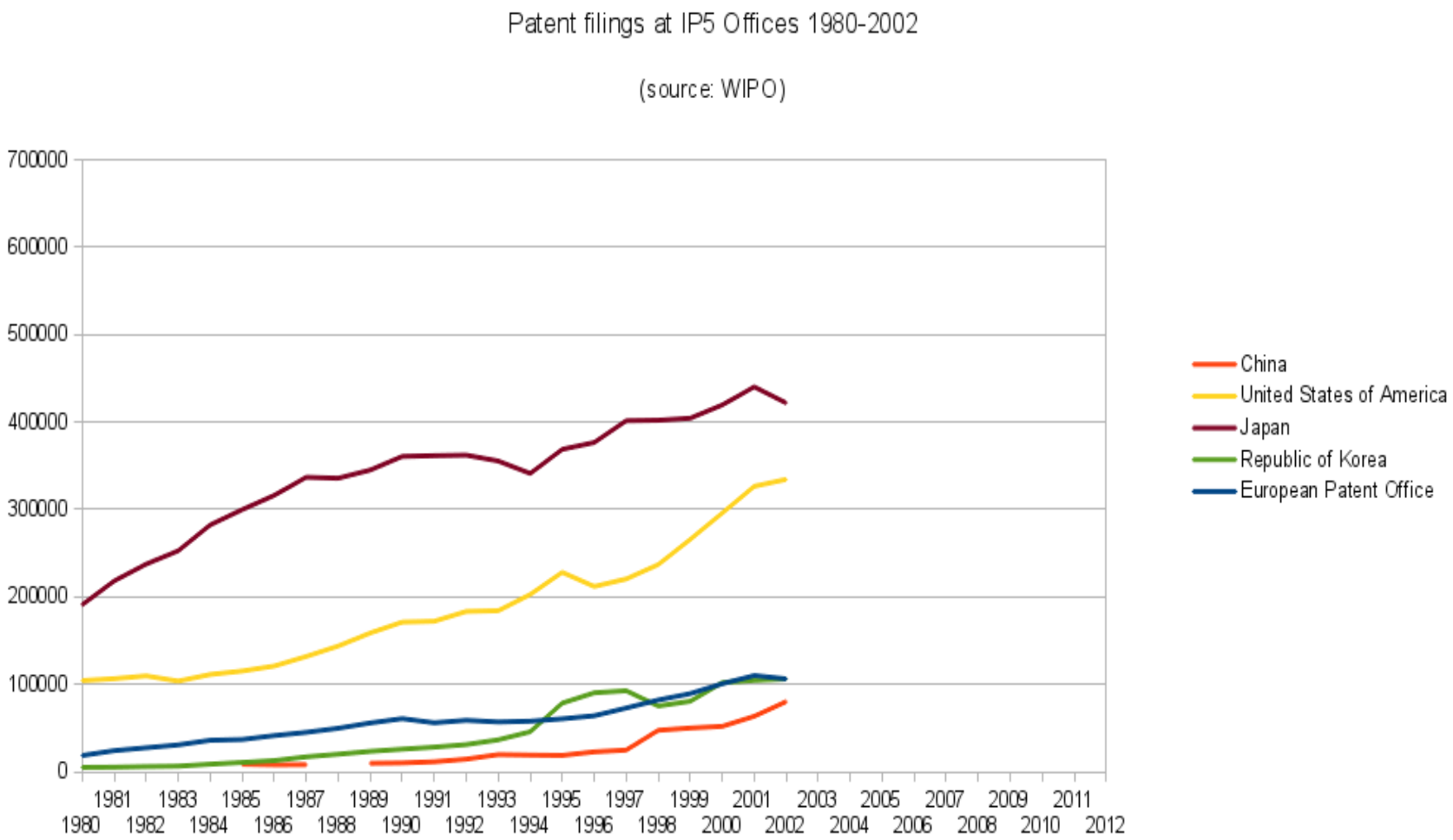
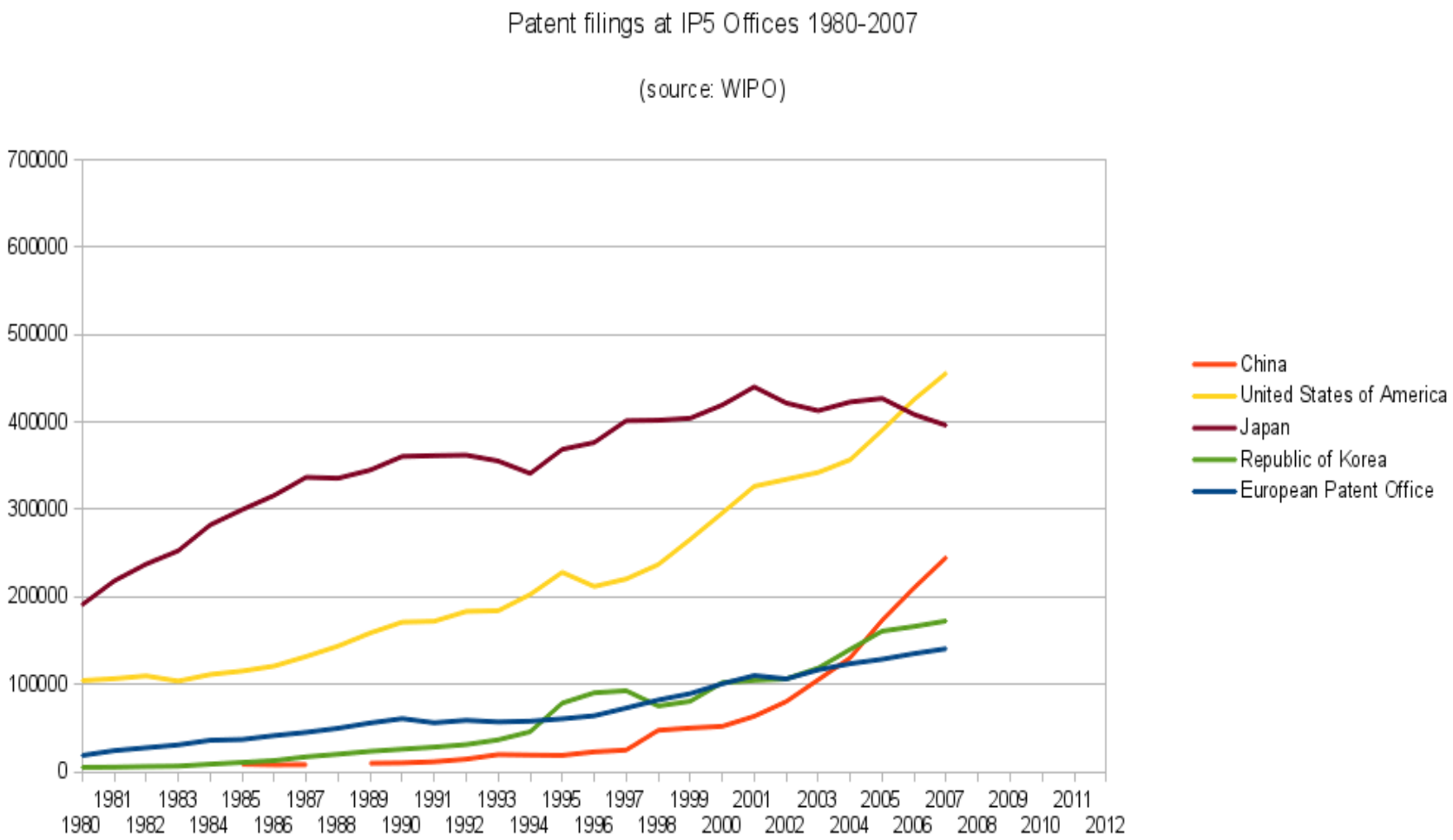
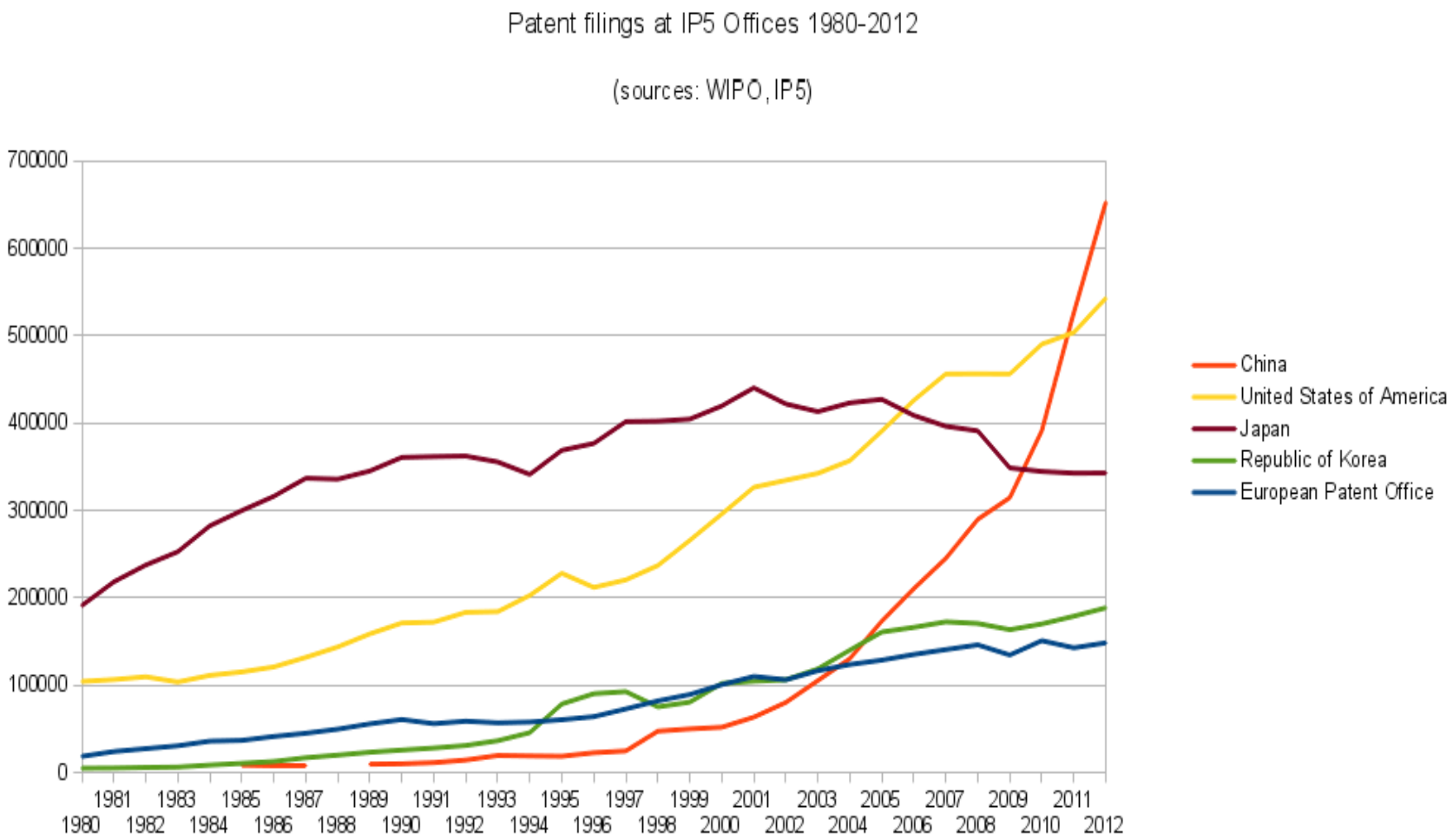
Towards automated mining of chemical structures in Chinese Patents

Daniel Bonniot de Ruisselet
ChemAxon

ICIC 2013, Vienna
October 16th 2013


Patent filings at IP5 Offices 1980-2002
(source: WIPO)
700000
600000
500000
400000
300000
200000
100000
0
1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
China
United States of America
Japan
Republic of Korea
European Patent Office


Patent filings at IP5 Offices 1980-2007
(source: WIPO)
700000
600000
500000
400000
300000
200000
100000
0
1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
China
United States of America
Japan
Republic of Korea
European Patent Office


Patent filings at IP5 Offices 1980-2012
(sources: WIPO, IP5)
700000
600000
500000
400000
300000
200000
100000
0
1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
China
United States of America
Japan
Republic of Korea
European Patent Office

Why Chinese patents matter

- Volume, exploding...
- Increasingly innovative
- Potential infrigment, lawsuits
  - Apple (2008, 2012, 2013), Schneider Electric, Samsung, ...
- Hard to access because of language

Why chemical mining matters

- Find interesting patent(s) using text search
  - Each patent can contain 100s of chemical names
  - Convert them automatically to structures
  - Enables chemical calculations
- Find interesting patent(s) using chemical structure search
  - Requires building a chemical database index
- Track structures accross multiple patents
  - Including multiple languages
  - Searching for prior art, infringment, …
  - Chemical similarity search
- ...

Putting it together

Chinese patents matter

&

chemical mining matters

→

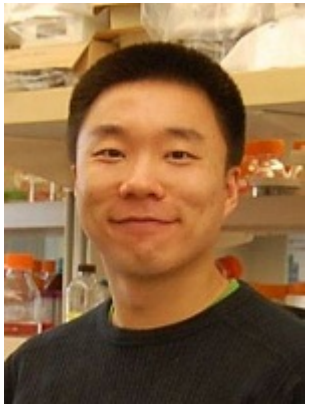
Chemical mining of chinese patents matters

ChemAxon?

- Cheminformatics, since 1998
- All of the top 15 global pharmas are customers

- Chemical database: indexing and searching
- English Name to Structure
- Document to Structure
- Missing piece: Chinese Name to Structure

Chinese Name to Structure



邓巍（Wei Deng, a.k.a. David)

Builds on english name to structure

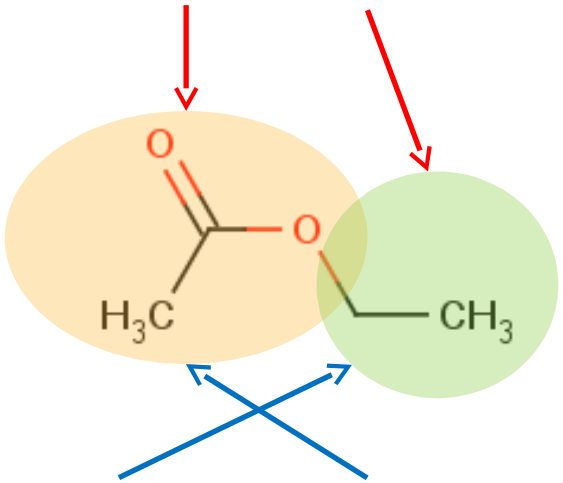
Specific dictionaries

Changes in algorithms...

The Challenges

1. Chinese texts have no spaces
2. Ester & Salt

乙酸乙酯



Ethyl Acetate

The Challenges

3. English: name alterations

丁烷 → buta + ane → butane

4. Chinese: many Characters have different meanings

盐 = salt
酸 = acid
盐酸 = hydrochloric acid

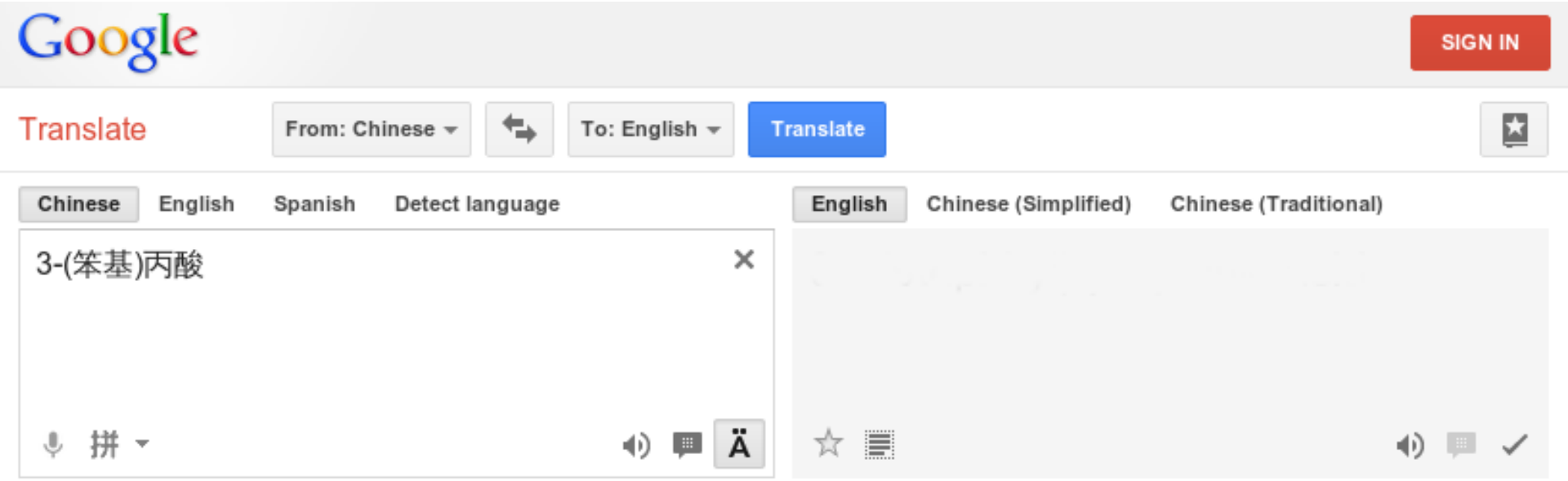
OCR Error correction

3-(笨基) 丙酸

苯 = benzene

苯基 = phenyl

丙酸 = proprionic acid



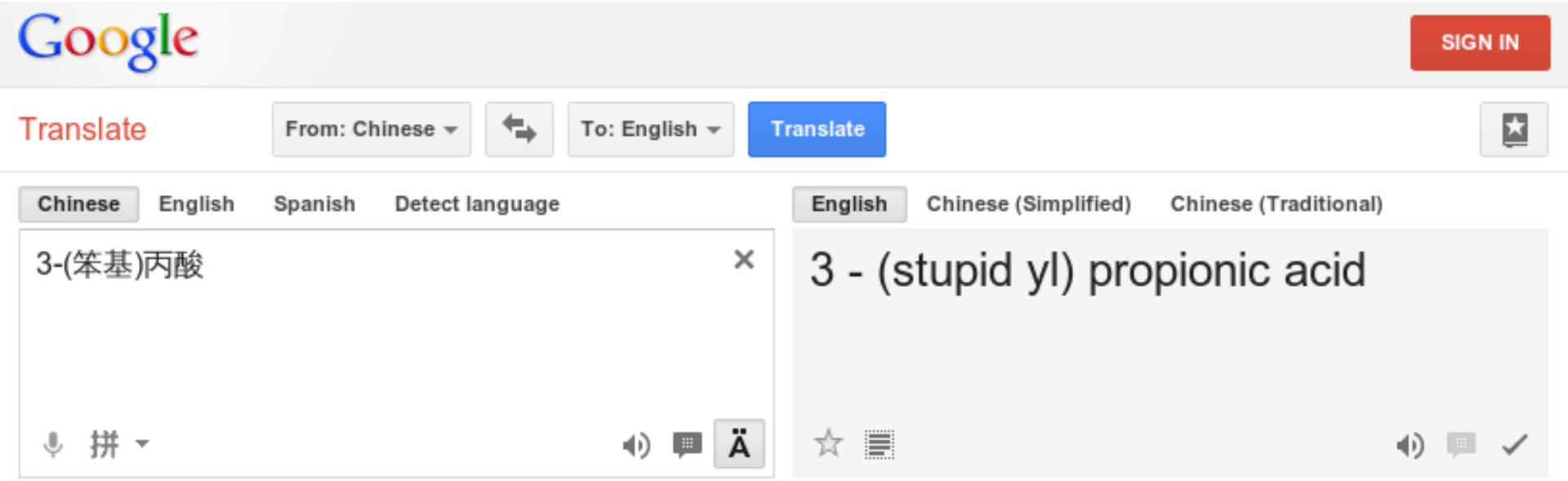
OCR Error correction

3-(笨基) 丙酸

苯 = benzene

苯基 = phenyl

丙酸 = proprionic acid



Chinese Document to Structure

- Additional challenge: no spaces
- 如式 I 所示的 {5-[2-(4- 正辛基苯基) 乙基]-2 ， 2- 二甲基 -1 ， 3- 二氧六环 -5- 基 } 氨基甲酸叔丁酯是合成芬戈莫德及其衍生物的重要中间体。

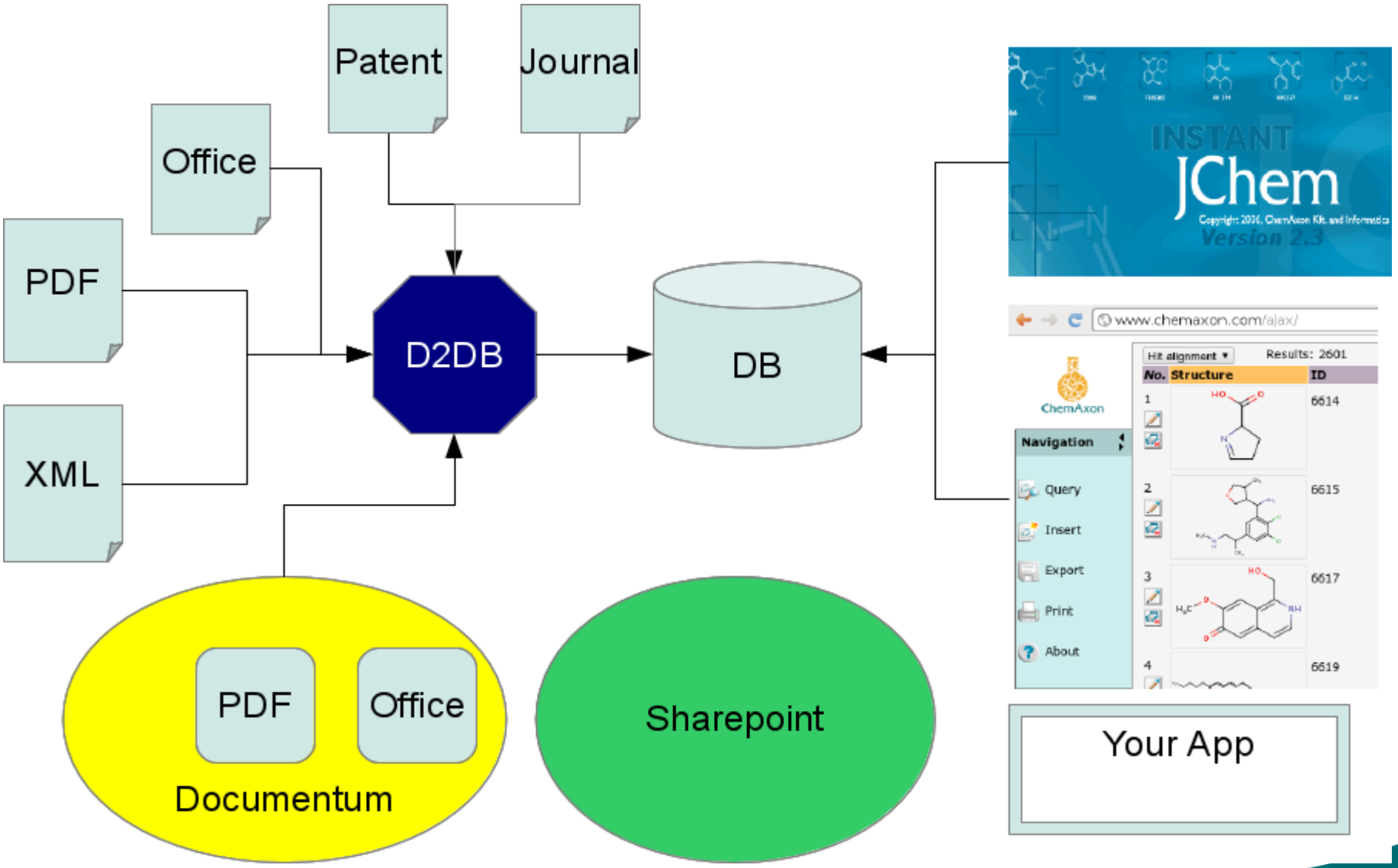
Chinese Document to Structure

- Additional challenge: no spaces
- 如式 I 所示的 **{5-[2-(4- 正辛基苯基) 乙基]-2 ， 2- 二甲基 -1 ， 3- 二氧六环 -5- 基 } 氨基甲酸叔丁酯**是合成芬戈莫德及其衍生物的重要中间体。

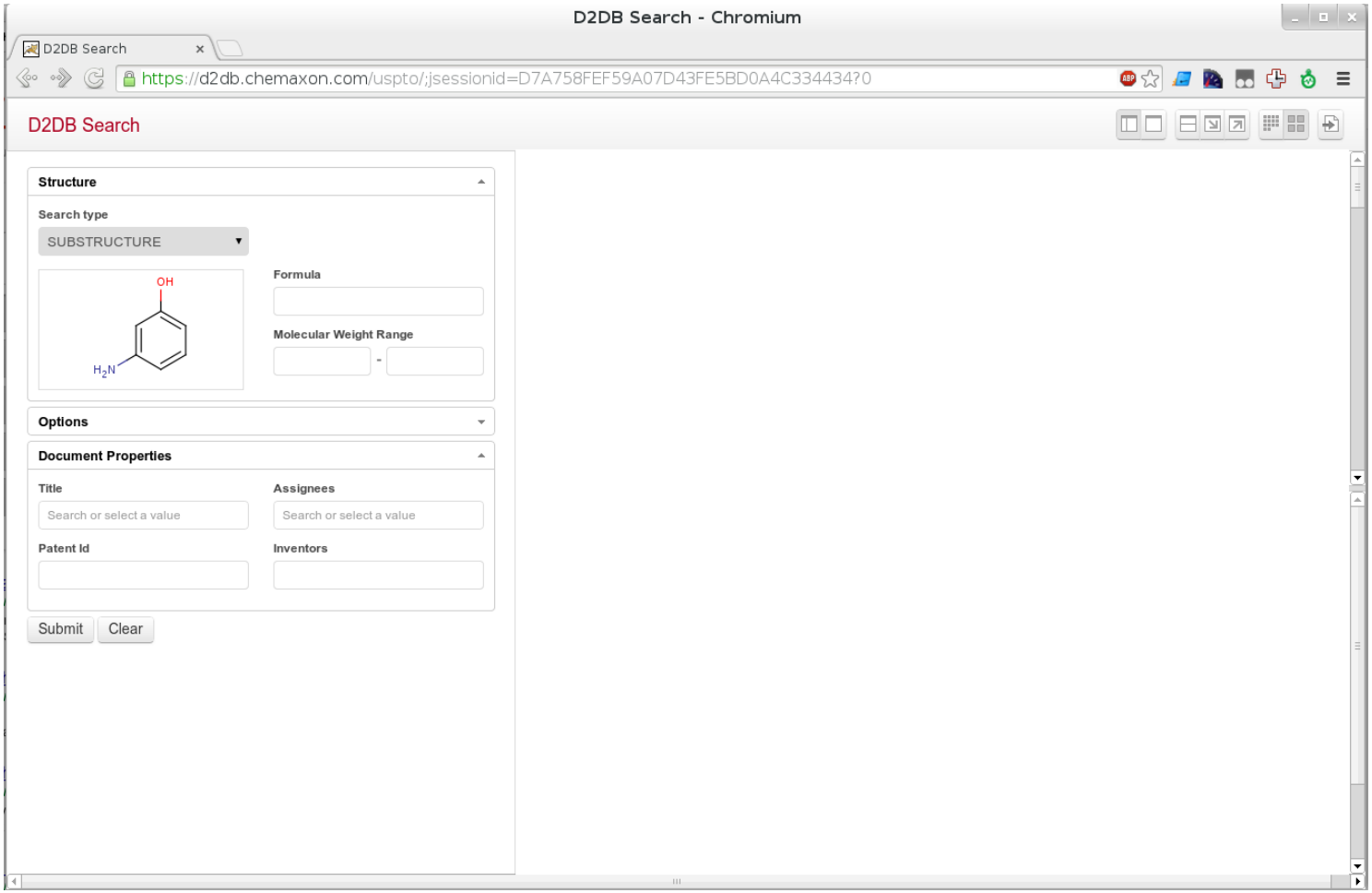
Chinese Document to Structure

- Additional challenge: no spaces
- 如式 I 所示的 **{5-[2-(4- 正辛基苯基) 乙基]-2 ， 2- 二甲基 -1 ， 3- 二氧六环 -5- 基 } 氨基甲酸叔丁酯**是合成芬戈莫德及其衍生物的重要中间体。
- XML Markup
  - Patent metadata
  - Encoding of characters
  - Tags (e.g. <p>)
- Document annotation

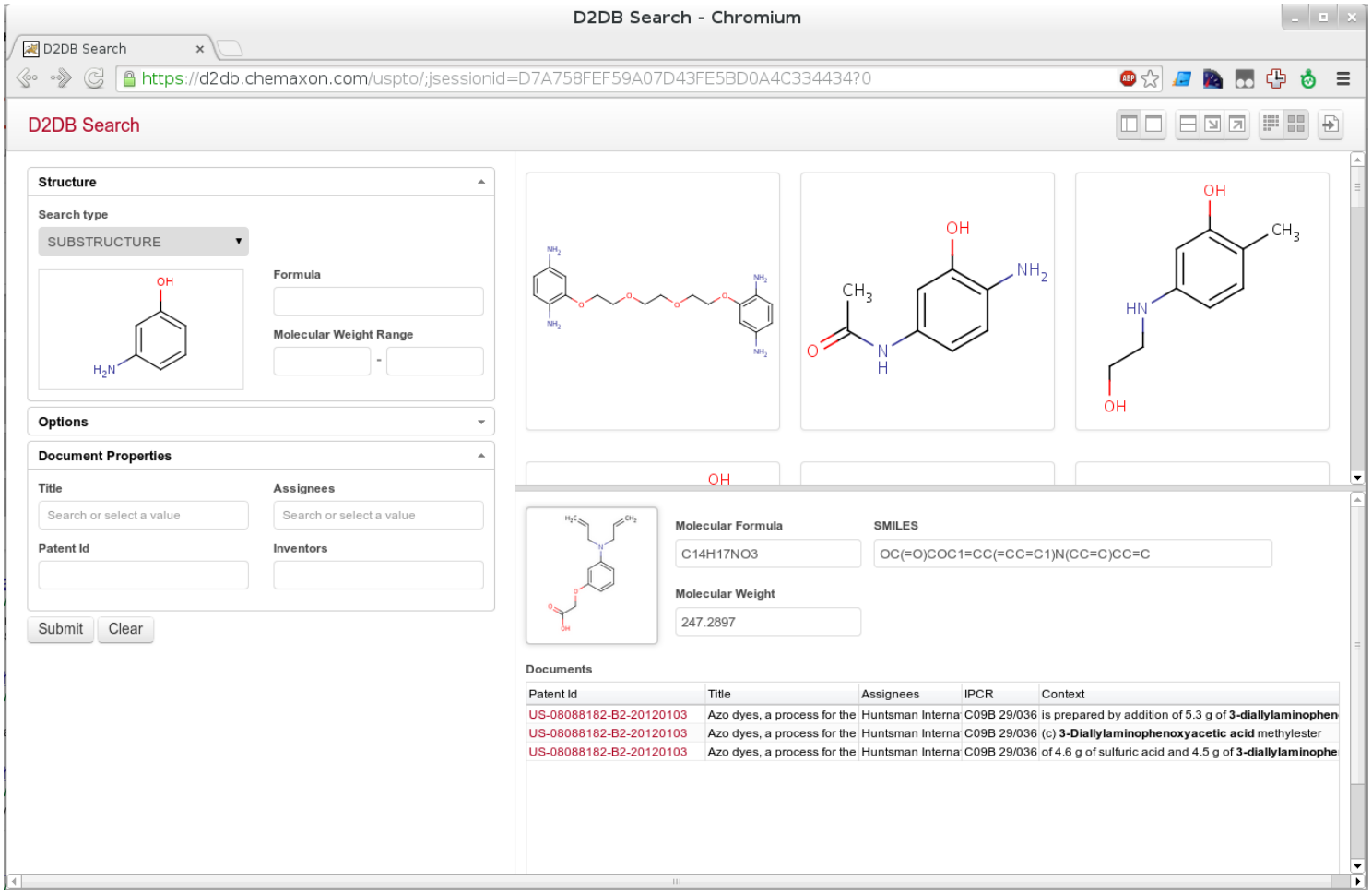
Document to Database



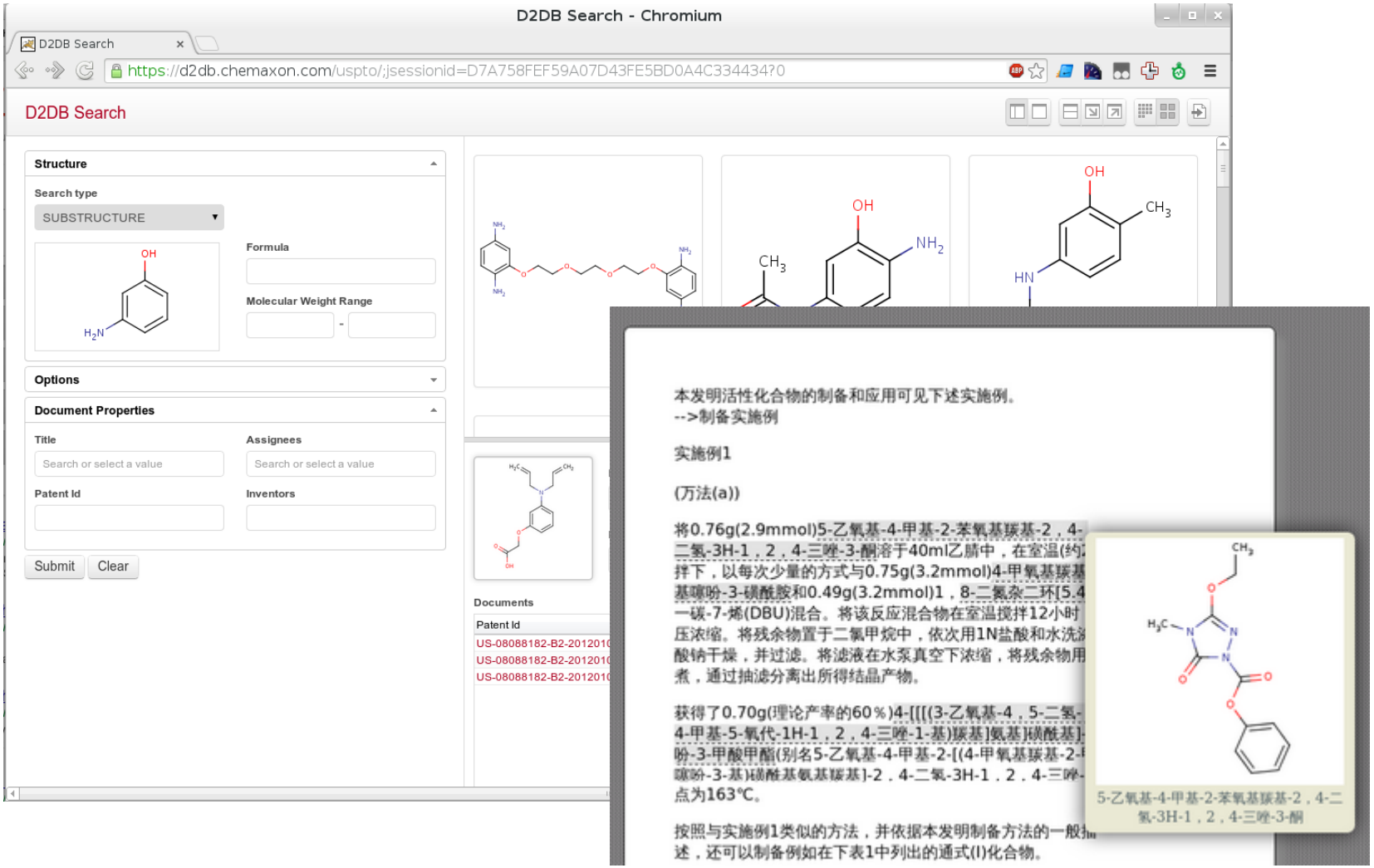
Document to Database



Document to Database



Document to Database

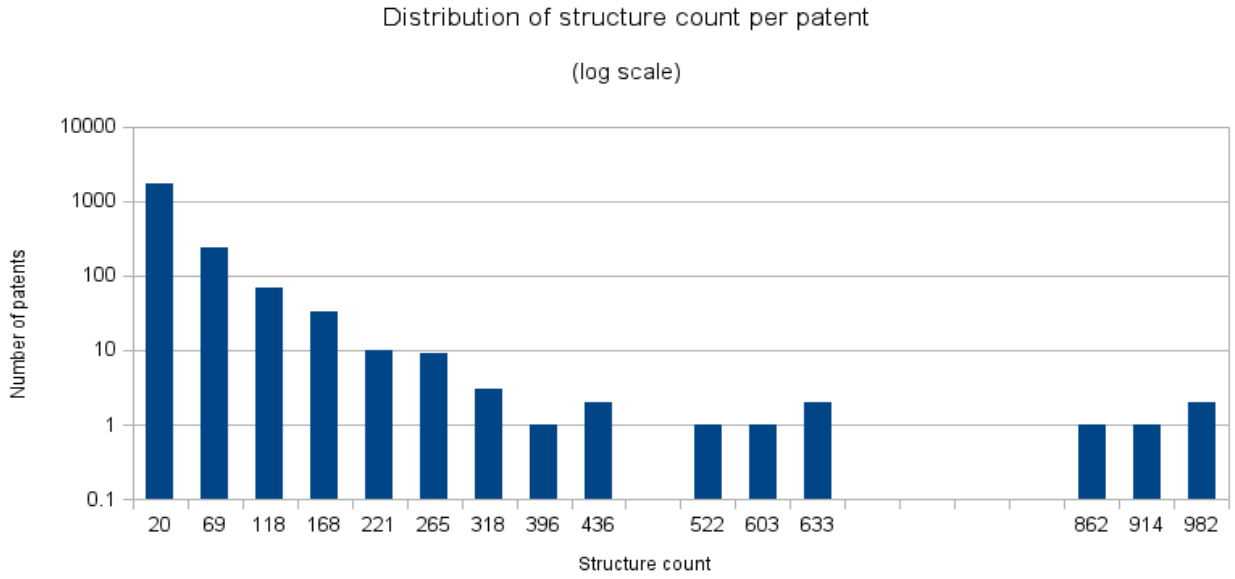


Validation 1: Chinese name to structure

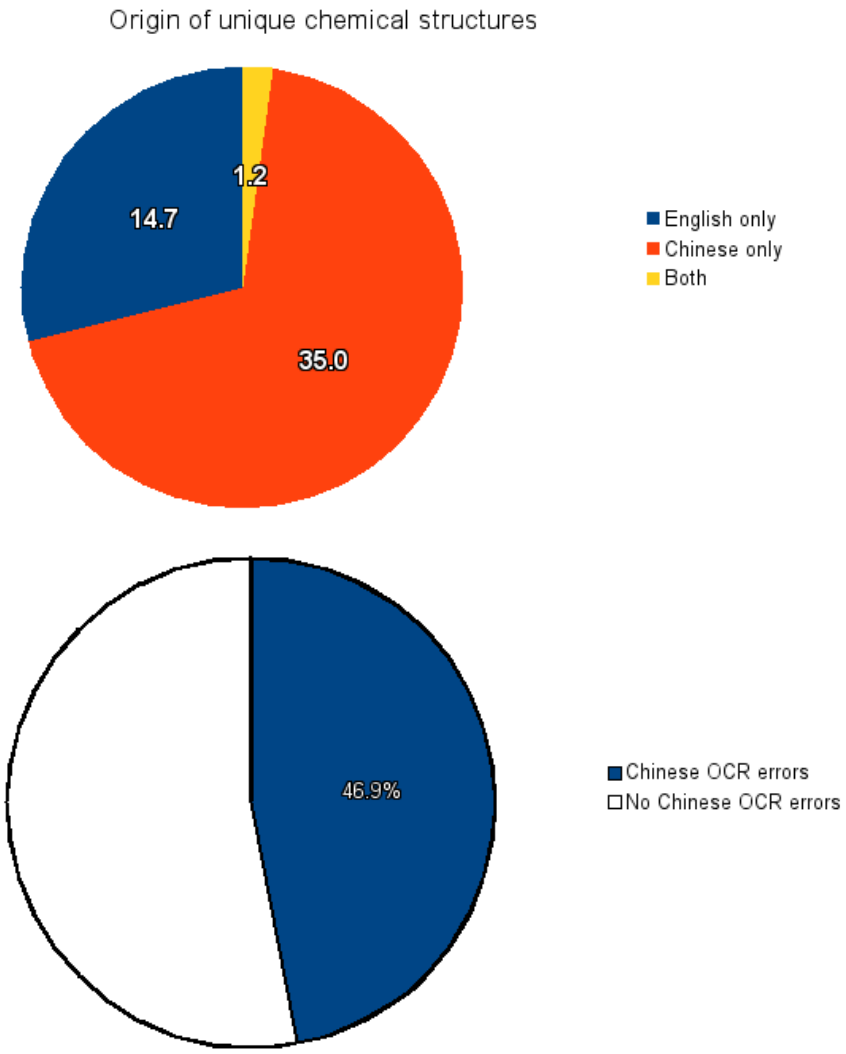
- Test set: 38,600 Chinese names + CAS number
- Contains unusual, incorrect, ambiguous names, radicals, inorganic salts,
- Conversion rate = 59 – 79 %
- Accuracy = 91%

Validation 2: Chinese patents

- 54K chinese patents with automated english translation
- Filter: structures with at least 20 heavy atoms, and patents with at least 20 structures
- Remains: 2108 patents



Validation 2: Chinese patents



Conclusions

- Patent volume in chinese is booming
- It is important to mine & monitor it
- Automated solutions are needed, but hard
- General purpose auto translation is not enough
- Chinese N2S already gives better results
- ChemAxon can build solutions for specific workflows
- More collaboration with patent providers is needed to keep improving quality and solutions

谢谢！

Extra information

谢谢！

Automatic OCR Error Correction

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

↓

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

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

↓

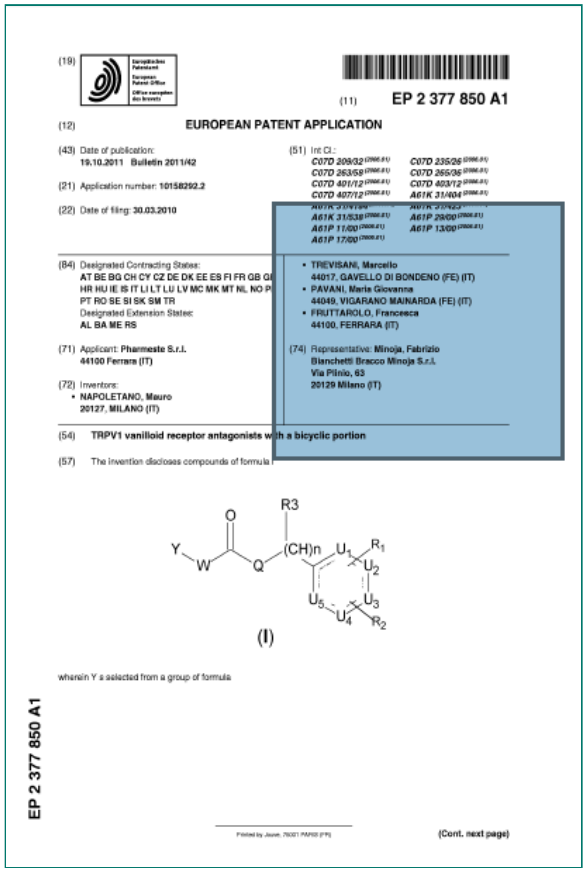
N-benzyl-*N*-[3-(1H-tetrazol-5-yl)phenyl]propanamide

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

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 31/4178 (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)

(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

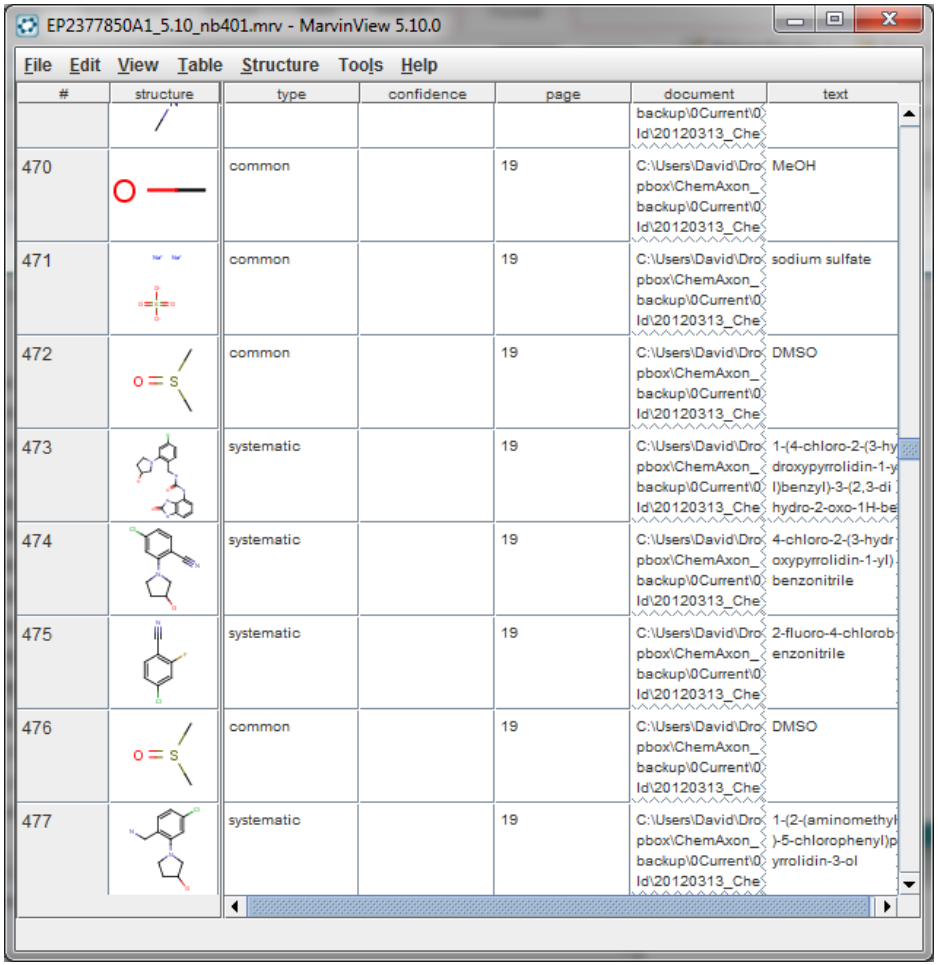
(57) The invention discloses compounds of formula

(I)

wherein Y a selected from a group of formula

EP 2 377 850 A1

(Cont. next page)



EP2377850A1_5.10_nb401.mrv - MarvinView 5.10.0

File Edit View Table Structure Tools Help

#	structure	type	confidence	page	document	text
					backup\0Current\0 Id\20120313_Che	
470		common		19	C:\Users\David\Dro pbox\ChemAxon_ backup\0Current\0 Id\20120313_Che	MeOH
471		common		19	C:\Users\David\Dro pbox\ChemAxon_ backup\0Current\0 Id\20120313_Che	sodium sulfate
472		common		19	C:\Users\David\Dro pbox\ChemAxon_ backup\0Current\0 Id\20120313_Che	DMSO
473		systematic		19	C:\Users\David\Dro pbox\ChemAxon_ backup\0Current\0 Id\20120313_Che	1-(4-chloro-2-(3-hy droxypyrrolidin-1-y l)benzyl)-3-(2,3-di hydro-2-oxo-1H-be
474		systematic		19	C:\Users\David\Dro pbox\ChemAxon_ backup\0Current\0 Id\20120313_Che	4-chloro-2-(3-hydr oxypyrrolidin-1-yl) benzonitrile
475		systematic		19	C:\Users\David\Dro pbox\ChemAxon_ backup\0Current\0 Id\20120313_Che	2-fluoro-4-chlorob enzonitrile
476		common		19	C:\Users\David\Dro pbox\ChemAxon_ backup\0Current\0 Id\20120313_Che	DMSO
477		systematic		19	C:\Users\David\Dro pbox\ChemAxon_ backup\0Current\0 Id\20120313_Che	1-(2-(aminomethyl)-5-chlorophenyl)p yrrolidin-3-ol

Non-searchable patent (50 pages)

Structure (text + image) + location

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

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?
  - ...