Challenges in Next Generation Scientific and Patent Information Mining

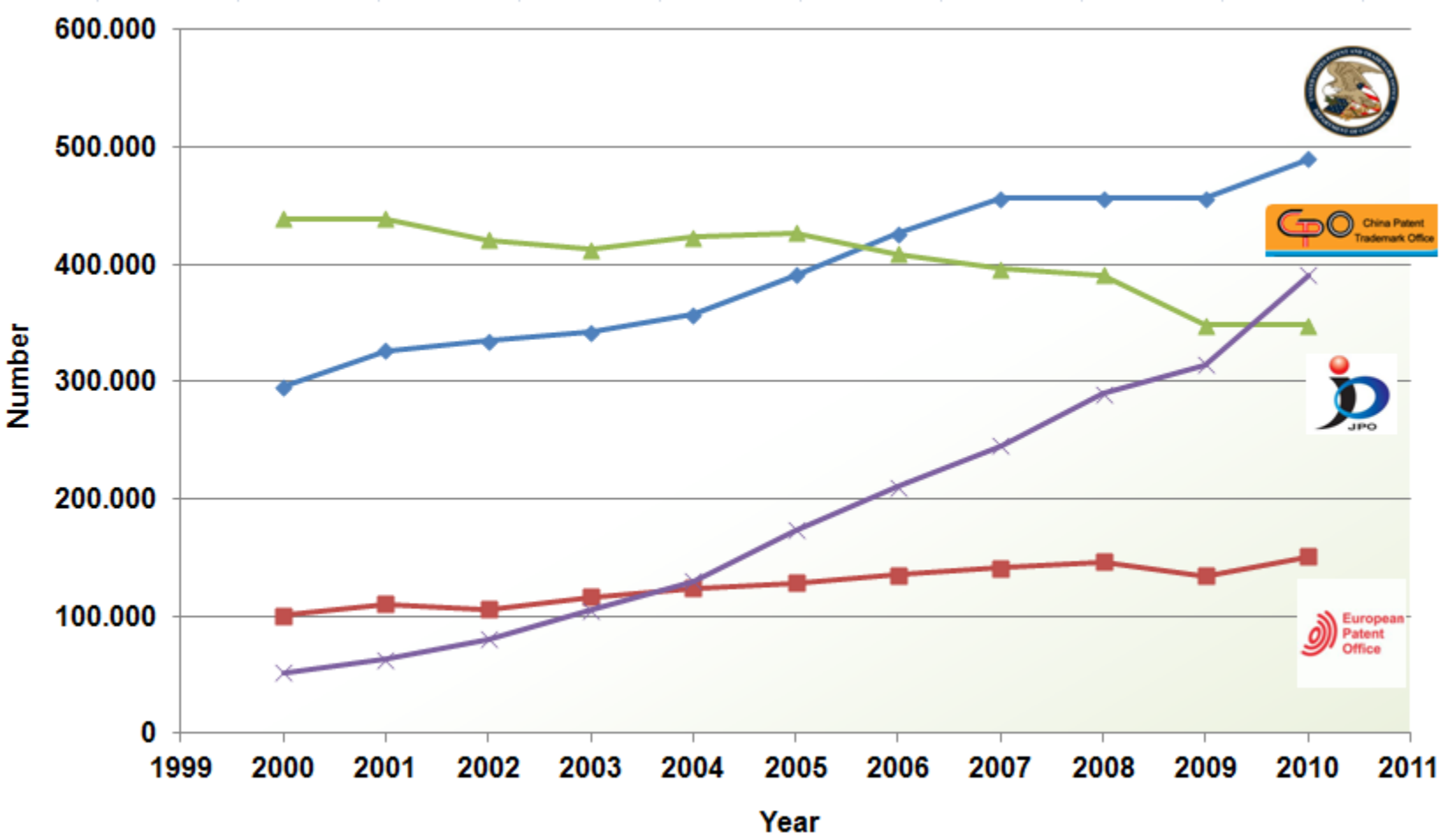
Josef Eiblmaier (InfoChem), Hans Kraut (InfoChem), Larisa Isenko (InfoChem), Heinz Saller (InfoChem), Peter Loew (InfoChem)

ICIC 2011 Barcelona, October 23 – 26

Outline

- Patents, a buried treasure?
- Chemical NER
- CDX work-up
- Image recognition
- Combining the techniques:
  ChemProspector

No. Applications Filed (Selected Patent Offices)

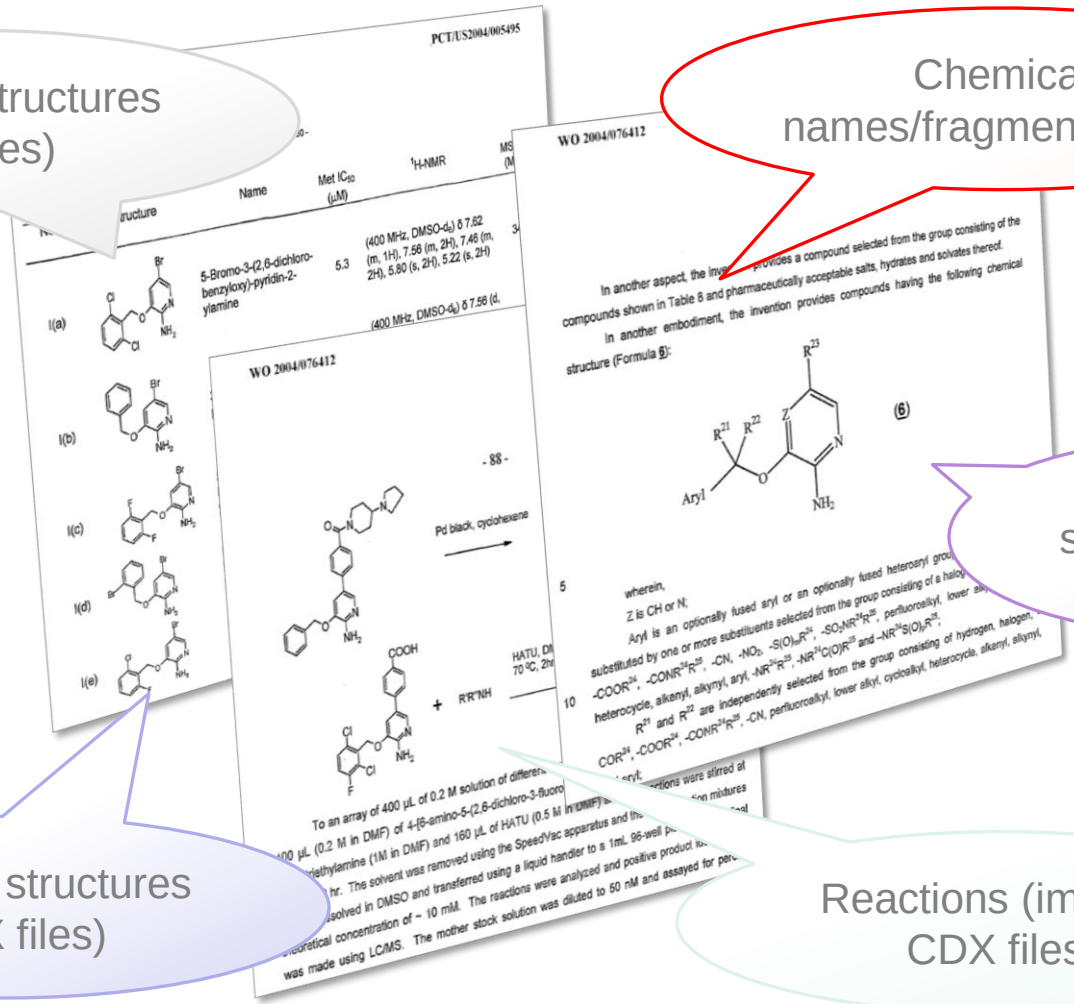


Source: EPO (http://www.epo.org/searching/asian/trends_de.html), USPTO (http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm)

Patents, a Buried Treasure?

- Scientific knowledge contained in patents comprises about 70 – 80 percent of all scientific knowledge
- In most cases, a patent is the first publication in the research process
- Important for competitor monitoring, technology assessment, R&D portfolio management
- Access to external technological knowledge

Chemical Pharmaceutical Patents: a Buried Treasure?



Chemical structures (images)

Chemical names/fragments (text)

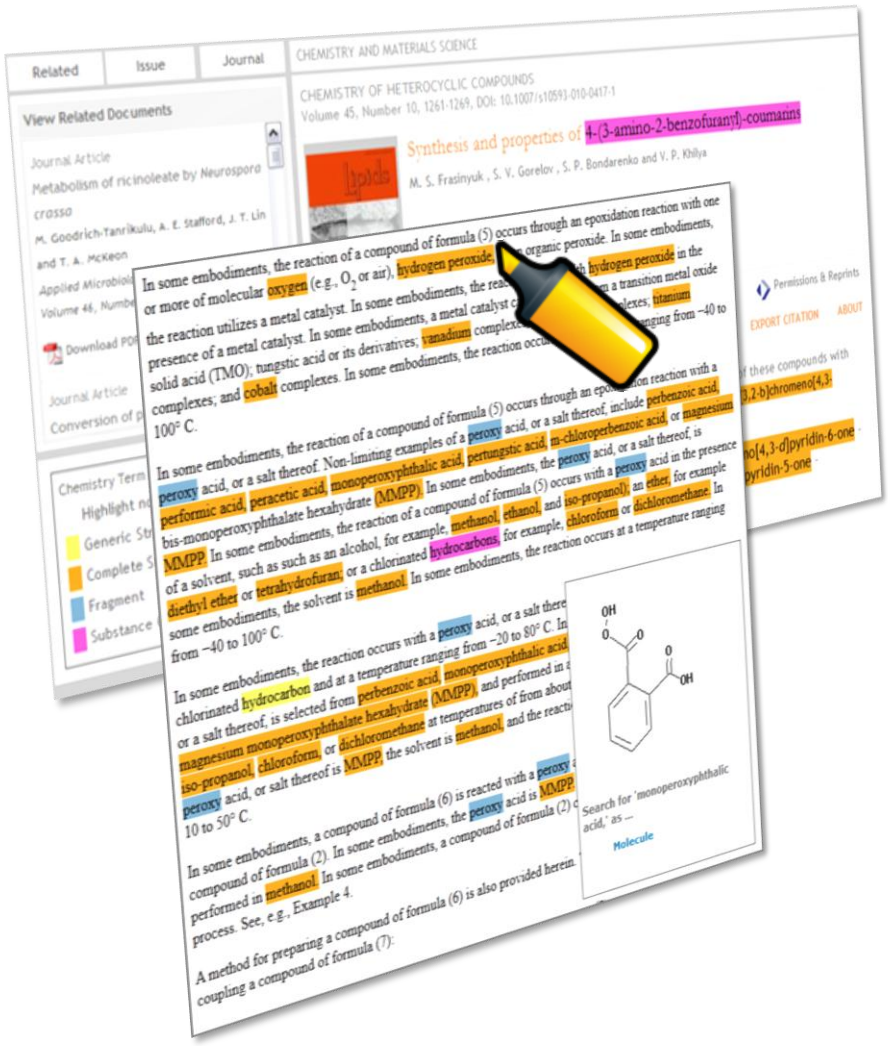
Markush structures (text, images, CDX)

Chemical structures (CDX files)

Reactions (images, CDX files)

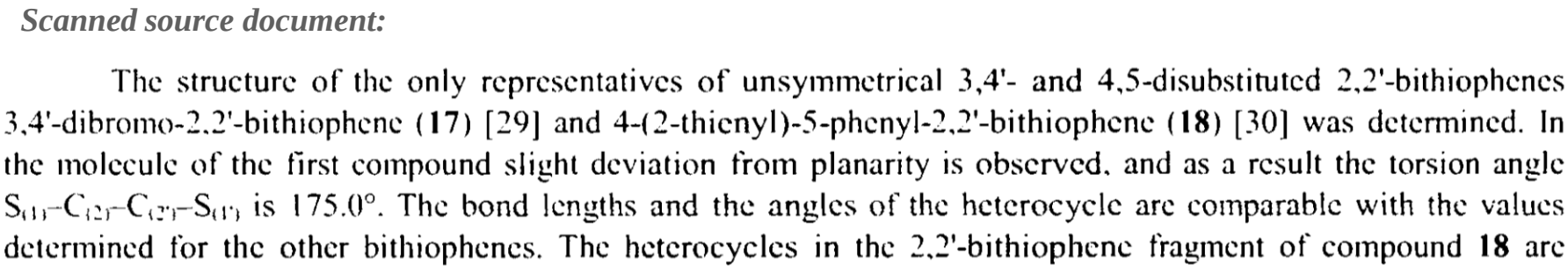
Chemical Named Entity Extraction: the Compulsory Exercise?

- **Named entity extraction (IC*ANNOTATOR*)**
  - Dictionary based
  - Morphology based
  - Context based
- **Name to structure conversion**
- **Tagging of source documents**
  - Highlighting
  - Enrichment with additional attributes
  - Linking (deep linking, cross document)



Challenges NER: OCR‘ed Documents

Scanned source document:



The structure of the only representatives of unsymmetrical 3,4'- and 4,5-disubstituted 2,2'-bithiophenes 3,4'-dibromo-2,2'-bithiophene (**17**) [29] and 4-(2-thienyl)-5-phenyl-2,2'-bithiophene (**18**) [30] was determined. In the molecule of the first compound slight deviation from planarity is observed, and as a result the torsion angle $S_{(1)}$–$C_{(2)}$–$C_{(2')}$–$S_{(1')}$ is 175.0°. The bond lengths and the angles of the heterocycle are comparable with the values determined for the other bithiophenes. The heterocycles in the 2,2'-bithiophene fragment of compound **18** are

OCR results:

Tile structure of the only reprcsentativcs o f unsyrnrnctrical 3,4'- and 4,5-disubstitutcd 2,2'-bithiophcncs 3,4'-dibromo-2.2'-bithiophcne (17) [29] and 4-(2-thienyl)-5-phcnyl-2,2'-bithiophene (18) [30] was determined. In the molecul c of the first compound slight deviation from planarity is observed, and as a result the torsion angle S,~,-C~:~-Cn.~-S~n is 175.(I~ The bond lengths and the angles of the heterocycle arc comparable with the values determined lbr thc other bithiophenes. The heterocyclcs in thc 2,2'-bithiophene fragmcnt of compound 18 are

Challenges NER: ‘Digital Born‘ PDF‘s

PDF document:

[7α,8α,3′,4′]-N′-(Phenyl)succinimido-6,14-*endo*-etheno-6,7,8,14-tetrahydrothebaine (VI) was synthesized using the method described in [16];

PdfToText results:

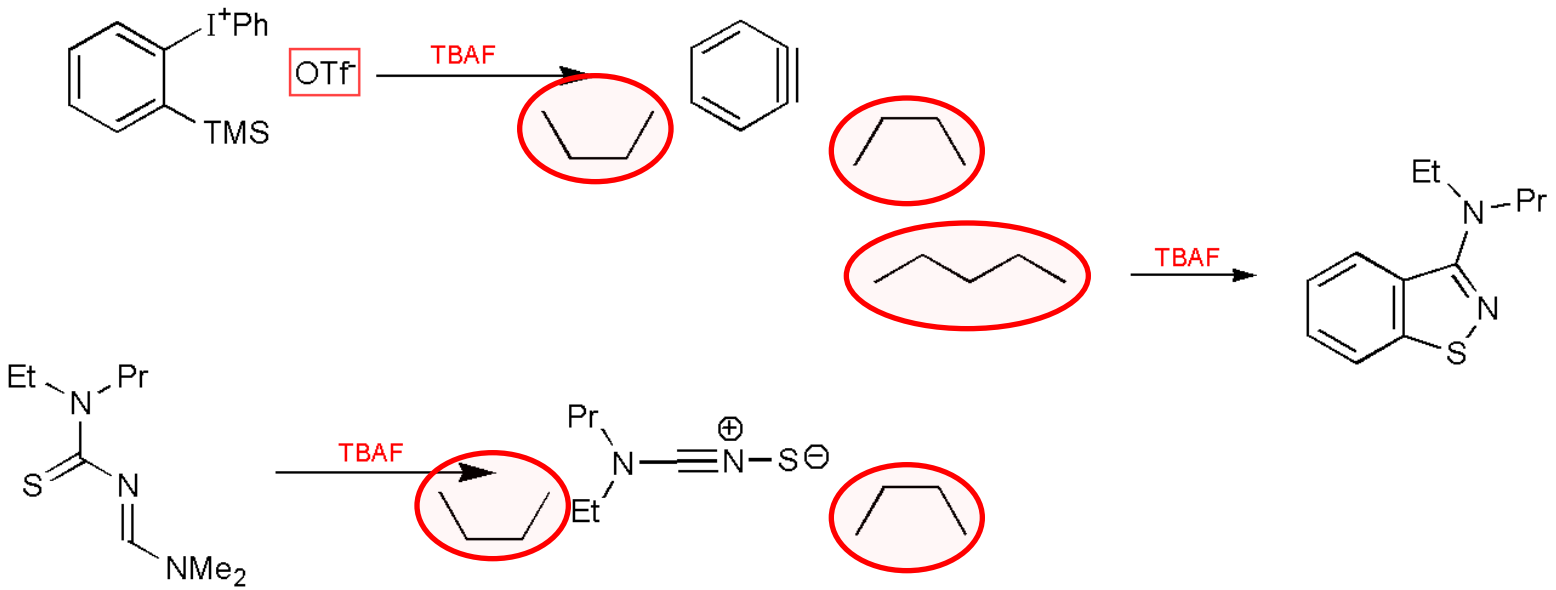
[7**a**,8**a**,3**¢**,4**¢**]-N**¢**-(Phenyl)succinimido-6,14-endo-etheno….

CDX Workup: Nothing is as it Seems!

- **For various sources CS ChemDraw files are available on large scale**
  - MRW's (e.g. Science of Synthesis)
  - Journals (e.g. Thieme, Wiley, Springer)
  - Patents (e.g. USPTO: 2000 - present)
- **The contained structures/reactions can be worked-up automatically (*ICSchemeProcessor*)**
- **But be careful: nothing is as it seems!**

CDX Work-up: Nothing is as it Seems!

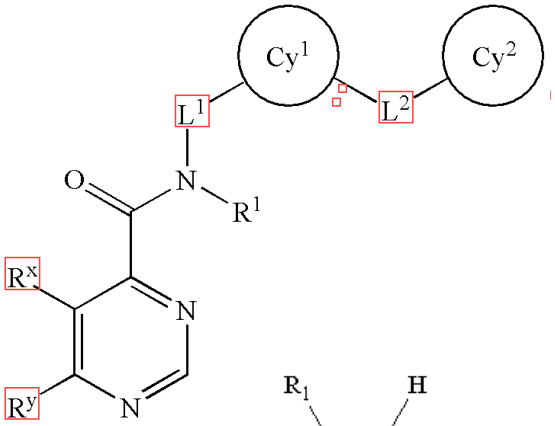
» **Reaction arrows / forked arrows / brackets**

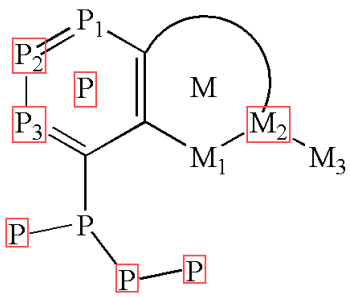


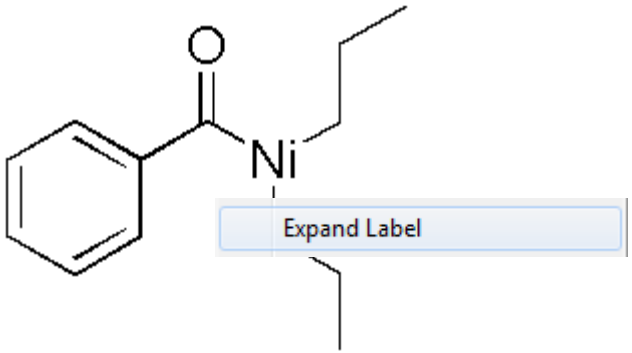
CDX Work-up: Nothing is as it Seems!

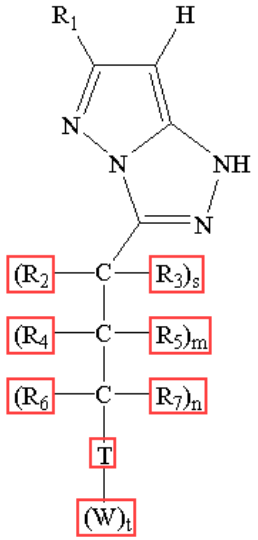
» **Unresolvable labels**

- Labels not defined
- Element symbols used as R-group labels
- Ambiguous fragment labels (e.g. molecular formula)



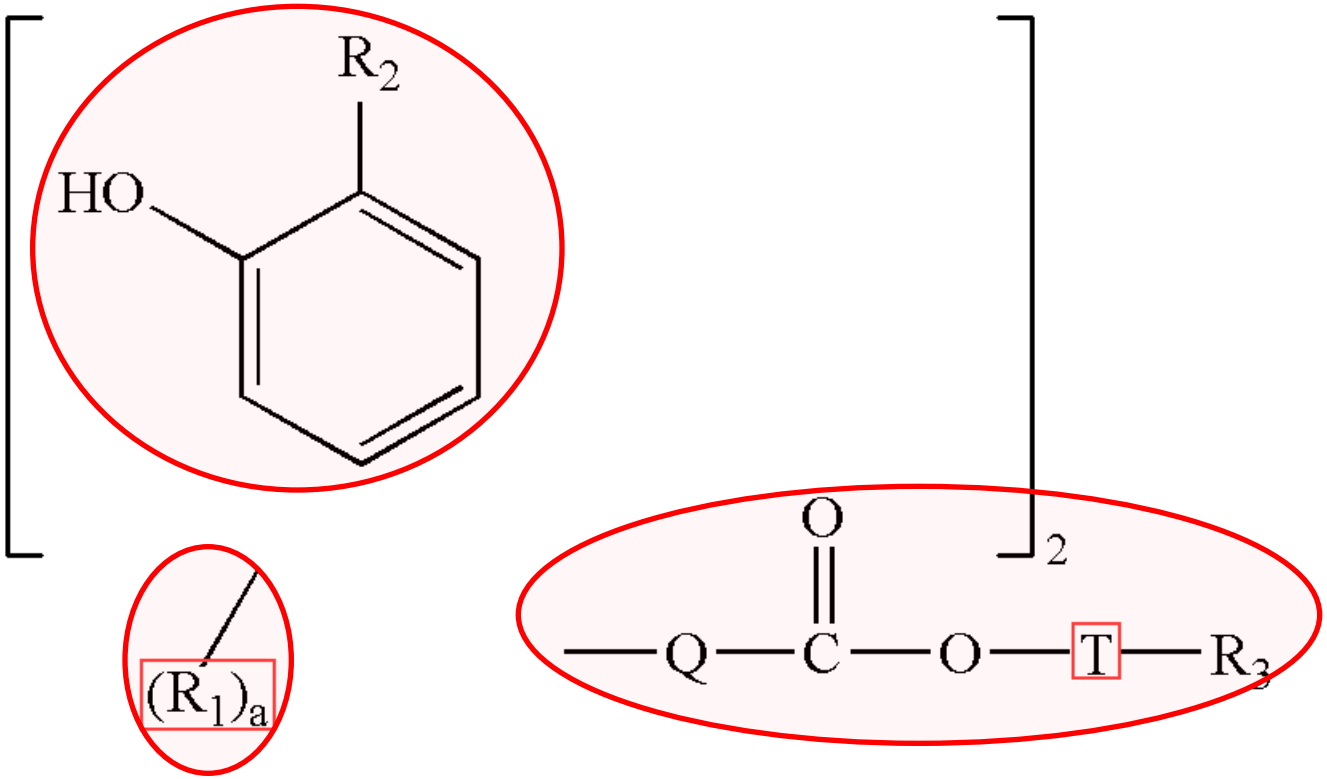






CDX Work-up: Nothing is as it seems!

» **Variable points of attachment**



CDX Work-up: Nothing is as it Seems!

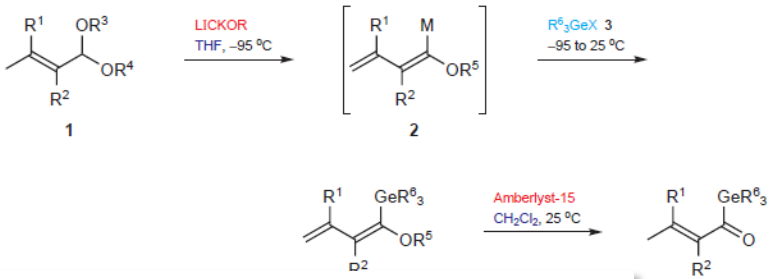
» **R-group enumeration**

- in the CDX file
- unstructured in the text
- structured in the text as table (XML)

5.1.16.6.1 **Method 1:**
Synthesis from Germyl Enol Ethers by Hydrolysis

The reaction of α,β-unsaturated *O,O*-acetals **1** with the butyllithium/potassium *tert*-butoxide superbase (LICKOR) gives 1-alkoxy-1-metallobuta-1,3-diene intermediates **2** via 1,4-elimination. The sequential trapping of these organometallic species with trialkyl(halo)germanes **3** affords 1-alkoxy-1-germylbuta-1,3-dienes **4** in moderate to excellent yields. Hydrolysis in the presence of Amberlyst-15 resin then furnishes α,β-unsaturated acylgermanes **5** (Scheme 1).[3]

Scheme 1 Acylgermanes by Hydrolysis of Germyl Enol Ethers[3]



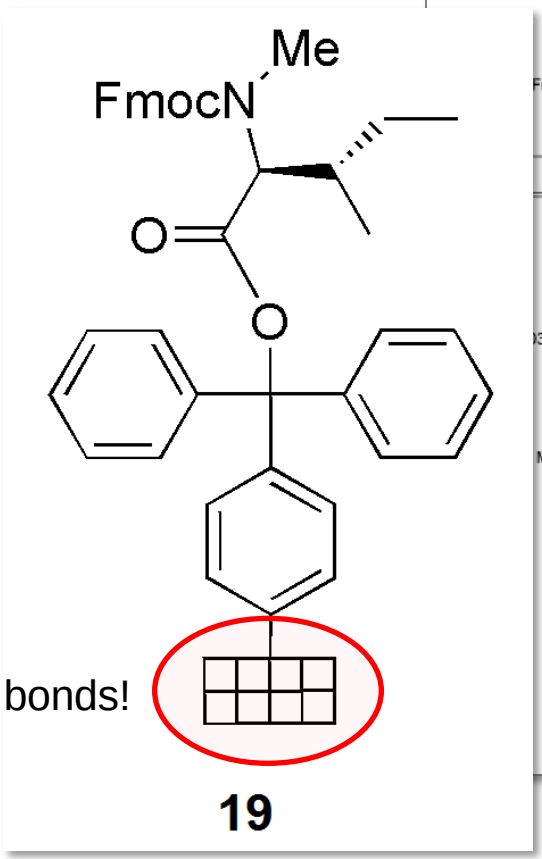
1 2 5

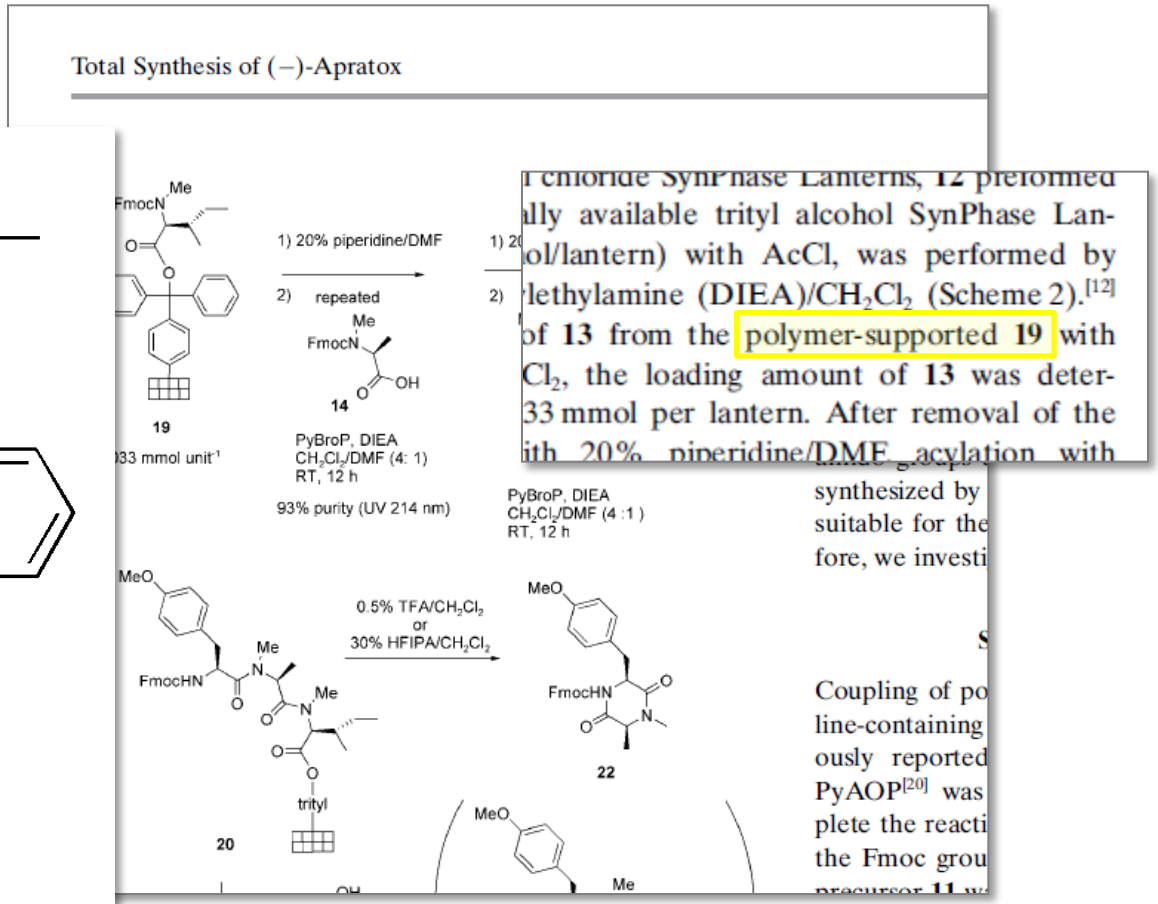
```
<sdoc name="07EBF-DOC-6" key="S0105100700300000006">
 <pmme>5.1.16.6.5 Method 5: Application of Acylgermanes as Amide Precursors</pmme>
 <sosc name="07EBF-SCH-6.1" key="S0105100700300000006 01">
<posc>Scheme 5 Synthesis of Amides from Acylgermanes[6]</posc>
<pisc><file name="S_102EBF005" key="S0105100700300000006 0101"><dsub>16,17</dsub></file></pisc>
<tabl>
<thdr>
<trow><tcel>R<ctfs>1</ctfs></tcel><tcel>R<ctfs>2</ctfs></tcel><tcel>R<ctfs>3</ctfs></tcel><tcel>Yield
(%)</tcel><tcel>Ref</tcel></trow>
</thdr>
<trow><tcel><ctfi>t</ctfi>-Bu</tcel><tcel>Bn</tcel><tcel>Bn</tcel><tcel>79</tcel><tcel><cbib>[6]</cbib></tcel></trow>
<trow><tcel>Ph</tcel><tcel>Bn</tcel><tcel>Bn</tcel><tcel>90</tcel><tcel><cbib>[6]</cbib></tcel></trow>
</tabl></sosc>
</sdoc>
</sman>
```

						ield[a] (%) of **5**	Ref
						5	[3]
						.r.	[3]
						5	[3]
						7	[3]
						.r.	[3]
						.r.	[3]
H	H	$CH_2CMe_2CH_2$	$CH_2CMe_2CH_2OH$	Bu	90	n.r.	[3]

CDX Work-up: Nothing is as it Seems!

» ‘Non chemical’ objects

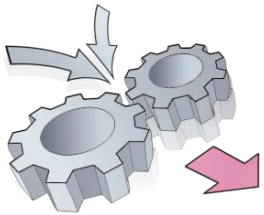




CDX Work-up: Solution/Approach ICSchemeProcessor

» **Algorithmic detection of features**

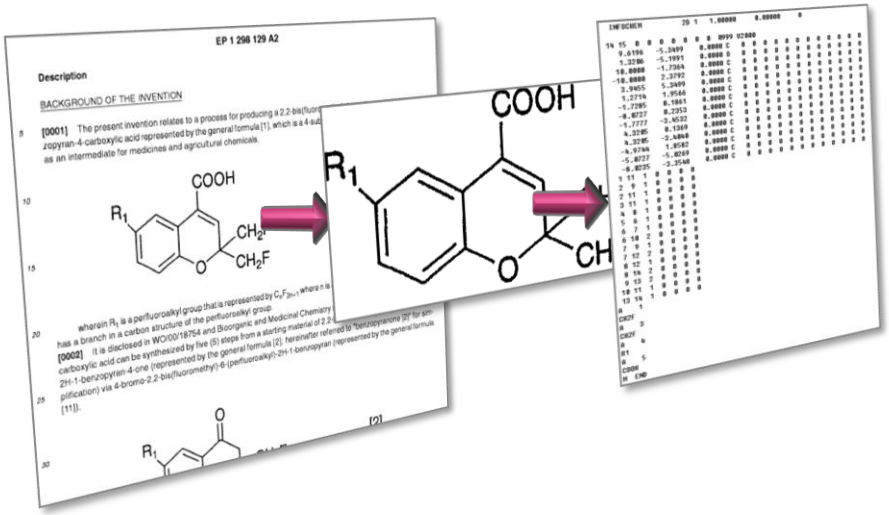
- Repeating groups
- Aliases, R-groups within CDX files



» **Guidelines for authors and typesetters**

- Syntax definitions for tables, R-groups etc.
- Syntax rules for captions
- Reaction arrangement, forked arrows
- Location and colour rules for reaction conditions (reactants, catalysts, solvents)

Image Recognition: the Supreme Challenge



» **In many cases only raster images are available**

» **Multiple step process:**

- page segmentation
- image classification
- vectorization, OCR, reconstruction

» **Part of ChemProspector: *ICImg2Struct***

- InfoChem proprietary development, started mid 2010

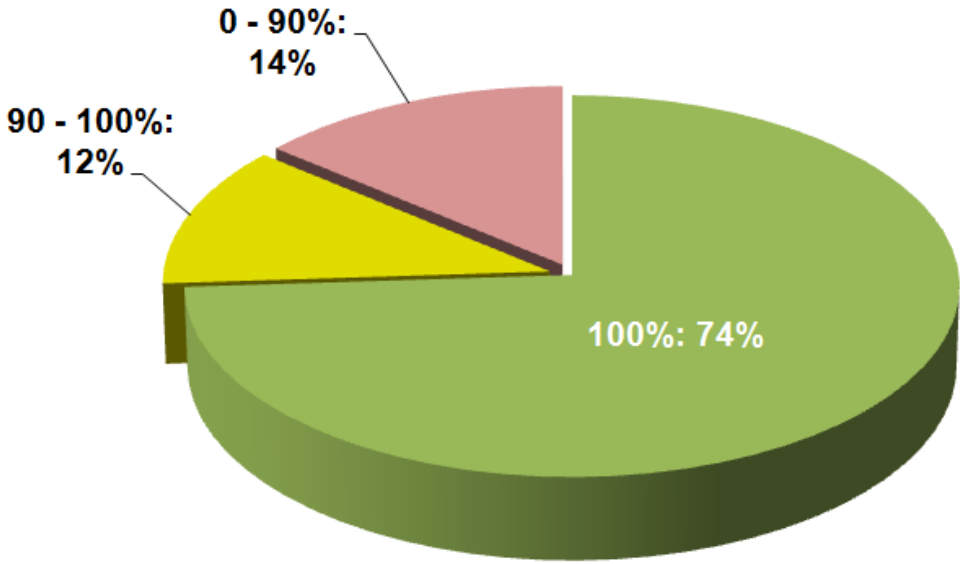
Image Recognition ICImg2Struct: Status and Evaluation

» **Used benchmark set: OSRA validation set (USPTO)**

- 5,735 chemical structures (images and associated MOL files) (http://cactus.nci.nih.gov/osra/uspto-validation.zip)

» **Degree of match: InfoChem SimilarityMCD**

- 100% correct: 4,234 structures (74%)
- 90 - 100%: 722 structures (12%)
- 0 – 90%: 779 structures (14%)



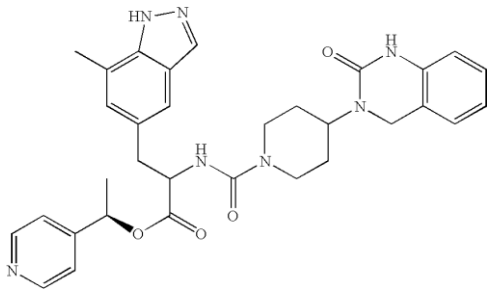
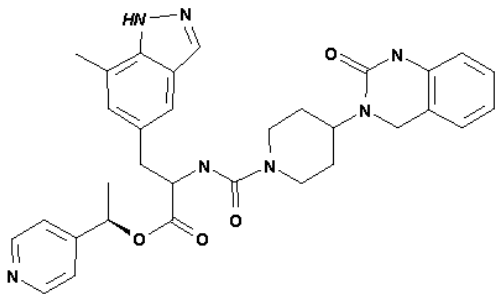
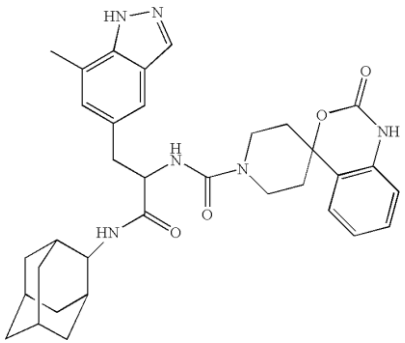
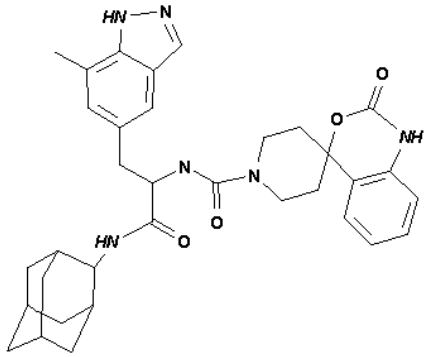
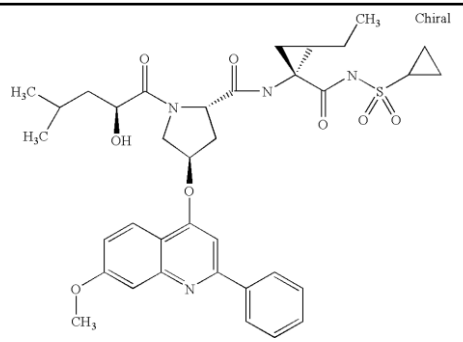
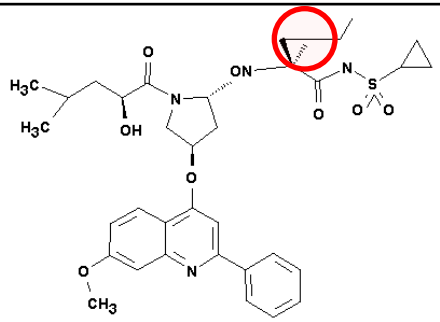
Original Image	ICImg2Struct	SimilarityMCD
		100%
		100%
		96.3%

Image Recognition: Variable Points of Attachment

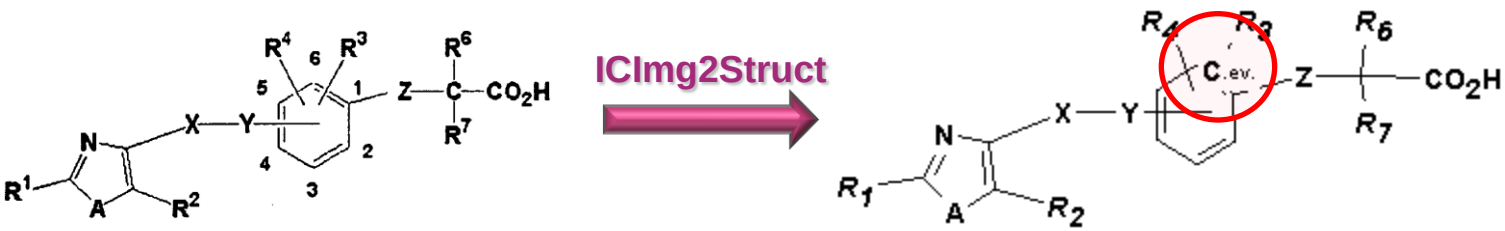
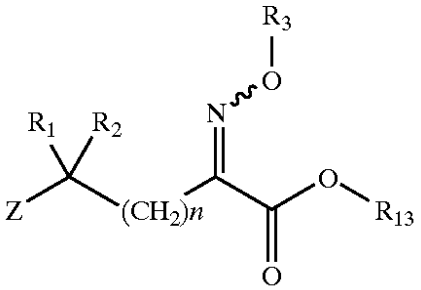
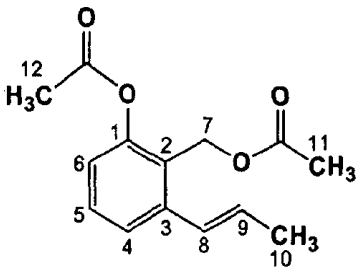


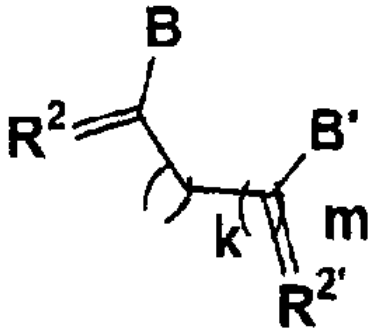
Image Recognition: More Challenges

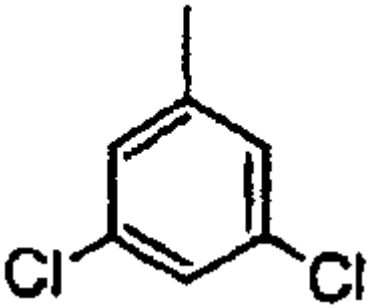
- » Wavy bonds
- » Atom numbers
- » Brackets
- » Chlorine, Iodine
- » Circles
- » Fused characters
- » Charges
- » Crossing bonds
- » Variable bonds
- » …

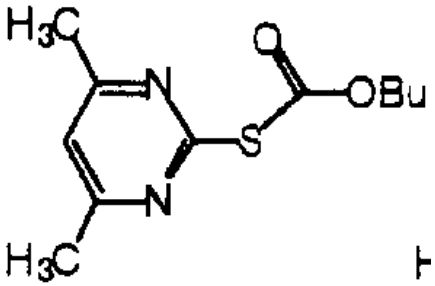




(III)







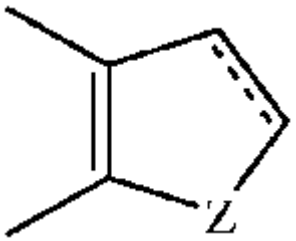
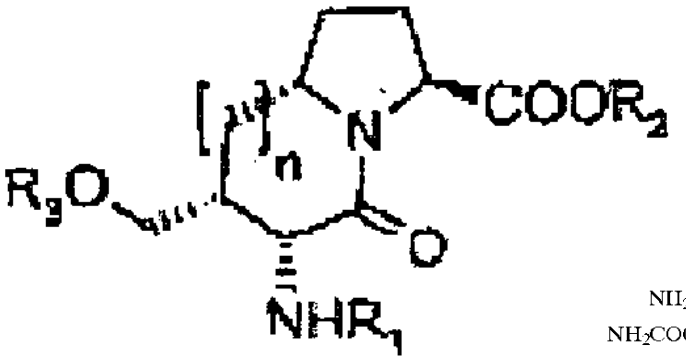
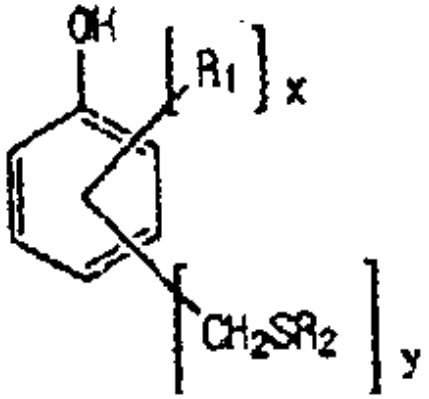
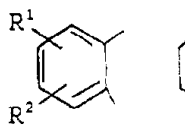


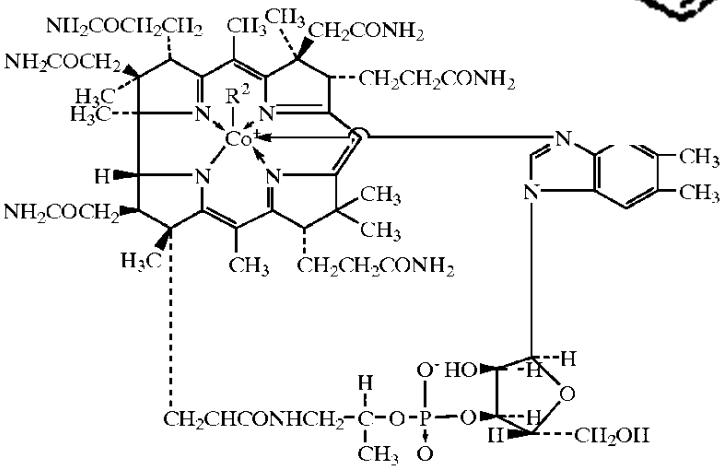
Image Recognition: Even More Challenges

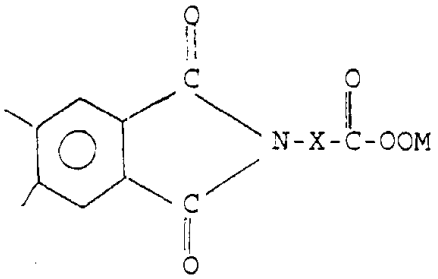


(Ia)









[0021] R^2=CN: Cyanocobalamin (I)

[0022] R^2=OH: Hydroxocobalamin (II)

[0023] R^2=CH_3: Methylcobalamin (V)

Combining the Techniques: ChemProspector

» Main emphasis:

'The automatic extraction of Markush Structures from patent documents'

» Research SME-project within the THESEUS research program

» THESEUS: 'New Technologies for the Internet of Services'

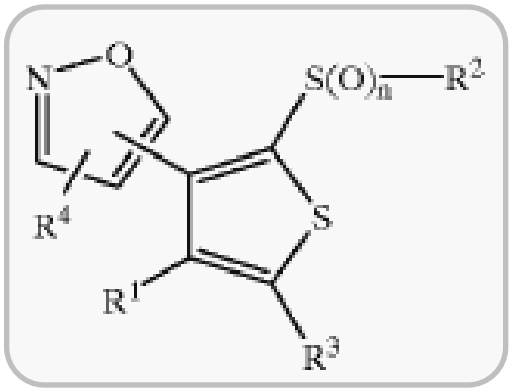
» Duration THESEUS: five years (2007 - 2011)

» Duration ChemProspector: July 2009 – end of 2011

Supported by:
Federal Ministry of Economics and Technology
on the basis of a decision by the German Bundestag

Main Challenges

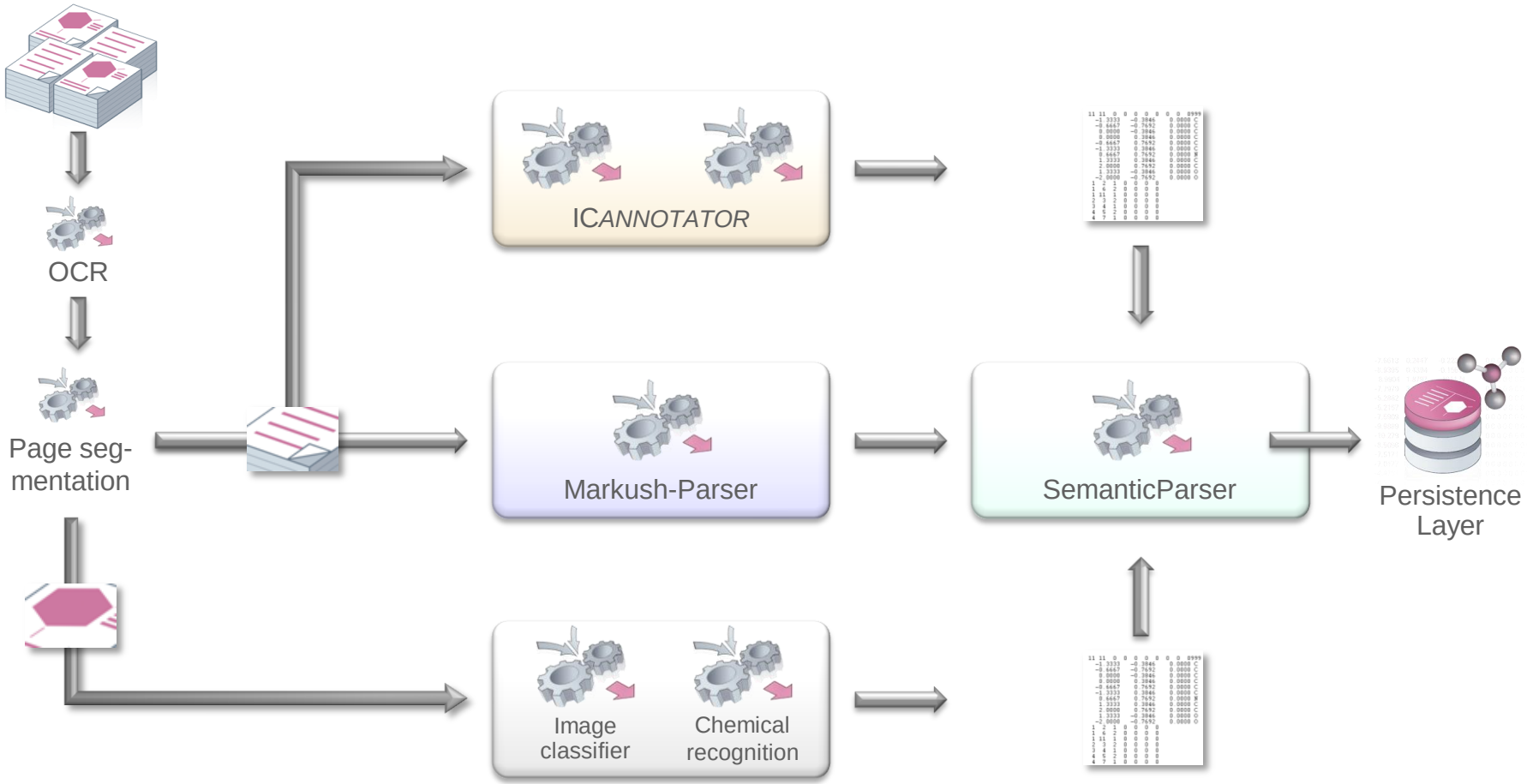
More specifically, the present invention is directed to a substituted isoxazolylthiophene compound represented by the formula (I)



wherein R^1 and R^2 individually represent an alkyl group of 1–5 carbon atoms, R^3 represents a cyano group or a group $CONR^5R^6$ (in which R^5 and R^6 individually represent a hydrogen atom or an alkyl group of 1–10 carbon atoms), R^4 represents an alkyl group of 1–5 carbon atoms or a phenyl group, and n is an integer of 0–2, or a salt thereof.

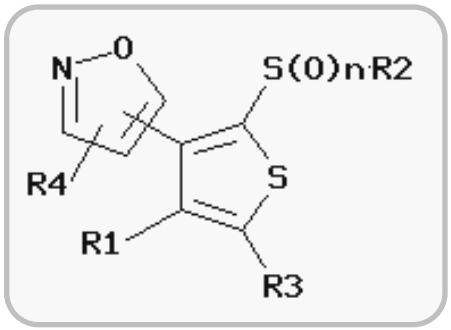
- Image classifier
- Chemical recognition
- ICANNOTATOR
- Markush-Parser
- SemanticParser

ChemProspector: Approach

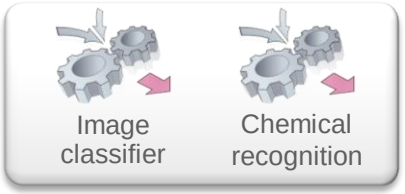


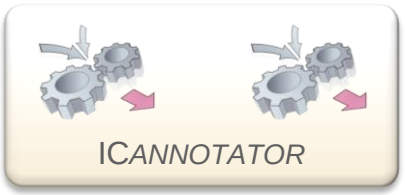
Results: Example

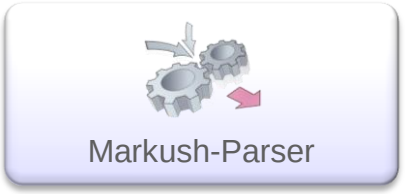
```
######## start markush description ########

US06355663-20020312-C00001
```




```
######## end markush description ########

R1 = "alkyl[Size(1-5)]"
R2 = "alkyl[Size(1-5)]"

R3 = "cyano" OR "CONR5R6"

R5 = "H" OR "alkyl[Size(1-10)]"
R6 = "H" OR "alkyl[Size(1-10)]"

R4 = "alkyl[Size(1-5)]" OR "phenyl"

n = 0-2
```



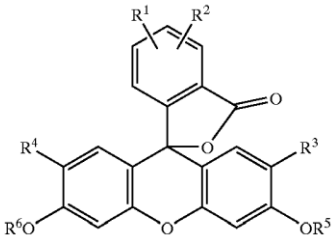




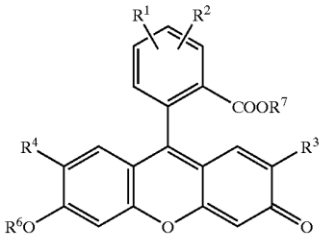
Results: US2003162298 A1

1. A compound represented by general formula (IA) or (IB) or a salt thereof:

(IA)

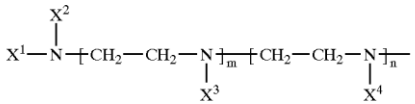


(IB)



wherein R^1 and R^2 independently represent a hydrogen atom or a group represented by formula (A):

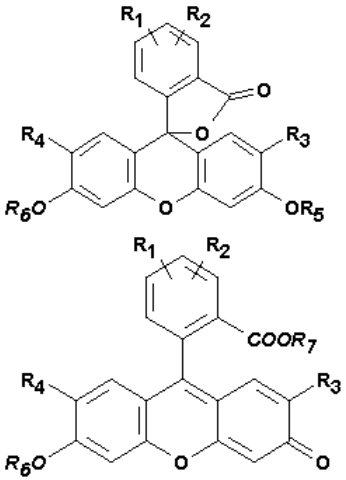
(A)



wherein X^1, X^2, X^3, and X^4 independently represent a hydrogen atom, an alkyl group, a 2-pyridylmethyl group, or a protective group for an amino group, and m and n independently represent 0 or 1, provided that R^1 and R^2 do not simultaneously represent hydrogen atoms; R^3 and R^4 independently represent a hydrogen atom or a halogen atom; R^5 and R^6 independently represent a hydrogen atom, an alkylcarbonyl group, or an alkylcarbonyloxymethyl group, and R^7 represents a hydrogen atom or an alkyl group.

######## start markush description ########

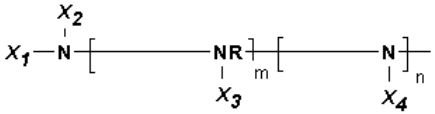
US20030162298A1-20030828-C00018



######## end markush description ########

R1 = "H" OR "STRUCT_IN_FORMULA(A)"
R2 = "H" OR "STRUCT_IN_FORMULA(A)"

US20030162298A1-20030828-C00019



X1 = "H" OR "alkyl" OR "2-pyridylmethyl" OR "Prt(amino)"
X2 = "H" OR "alkyl" OR "2-pyridylmethyl" OR "Prt(amino)"
X3 = "H" OR "alkyl" OR "2-pyridylmethyl" OR "Prt(amino)"
X4 = "H" OR "alkyl" OR "2-pyridylmethyl" OR "Prt(amino)"

m = 0-1
n = 0-1

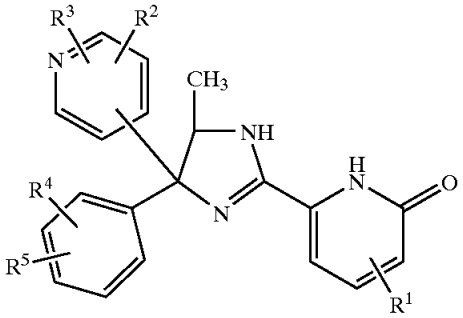
IF R1 AND R2 NOT SIMULTANEOUSLY = "H"

R3 = "H" OR "Hal"
R4 = "H" OR "Hal"
R5 = "H" OR "alkylcarbonyl" OR "alkylcarbonyloxymethyl"
R6 = "H" OR "alkylcarbonyl" OR "alkylcarbonyloxymethyl"
R7 = "H" OR "alkyl"

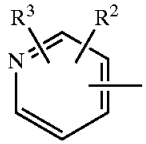
Results: US 20050154025 A1

1. A compound of the formula (I):

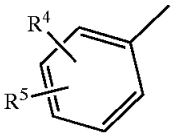
(I)



wherein R^1 is hydrogen, halogen, cyano, lower alkyl, halo-lower alkyl, hydroxy, lower alkoxy or aralkyloxy; R^2 and R^3 are each independently hydrogen, halogen or halo-lower alkyl; and R^4 and R^5 are each independently hydrogen or halogen, provided that when R^1 is hydrogen, a group of the formula:

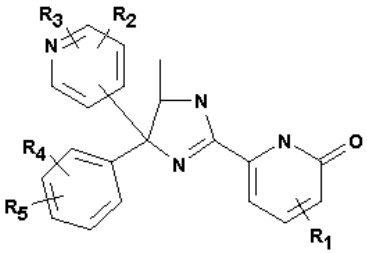


and a group of the formula:



######## start markush description ########

US20050154025A1-20050714-C00031



######## end markush description ########

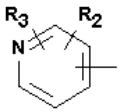
R1 = "H" OR "Hal" OR "cyano" OR "alkyl[Size(1-6)]" OR "alkyl" OR "hydroxy" OR "alkoxy[Size(1-6)]" OR "aralkyloxy"

R2 = "H" OR "Hal" OR "alkyl"
R3 = "H" OR "Hal" OR "alkyl"
R2 AND R3 INDEPENDENTLY
R4 = "H" OR "Hal"
R5 = "H" OR "Hal"
R4 AND R5 INDEPENDENTLY

IF R1 = "H" OR "Struct_E" OR "Struct_F"

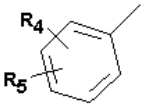
Struct_E

US20050154025A1-20050714-C00032



Struct_F

US20050154025A1-20050714-C00033



Acknowledgements

» The InfoChem Team

» The German Federal Ministry of Economy and Technology (BMWi)

Federal Ministry of Economics and Technology

Thank you!

Questions?