ChemicalTagger

Lezan Hawizy,
David Jessop, Daniel Lowe and Peter Murray-Rust

UNIVERSITY OF CAMBRIDGE

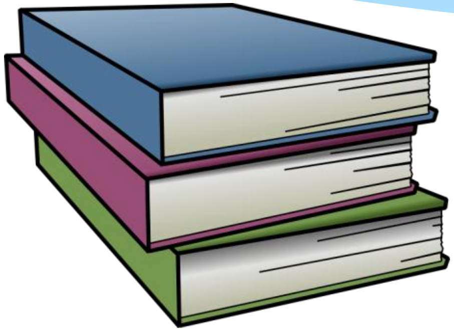
Unilever Cambridge
Centre For Molecular Science Informatics

Outline

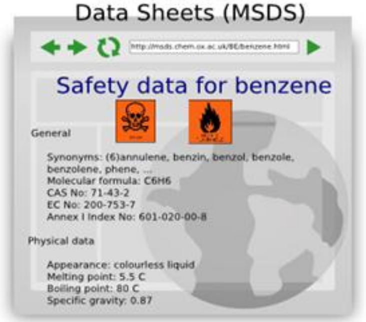
- Aim
- ChemicalTagger Components
- Evaluation
- Applications
- Demo

Current State of Chemical Information

- Large amounts of data produced annually
- Theses, journals, patents and reports
- Unstructured free-flowing text



Materials Safety Data Sheets (MSDS)



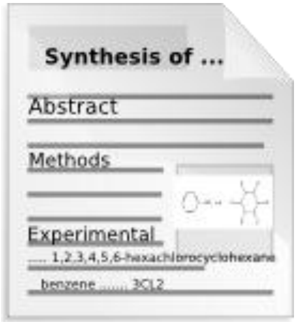
Safety data for benzene

General

Synonyms: (6)annulene, benzin, benzol, benzole, benzolene, phene, ...
Molecular formula: C6H6
CAS No: 71-43-2
EC No: 200-753-7
Annex I Index No: 601-020-00-8

Physical data

Appearance: colourless liquid
Melting point: 5.5 C
Boiling point: 80 C
Specific gravity: 0.87



Synthesis of ...

Abstract

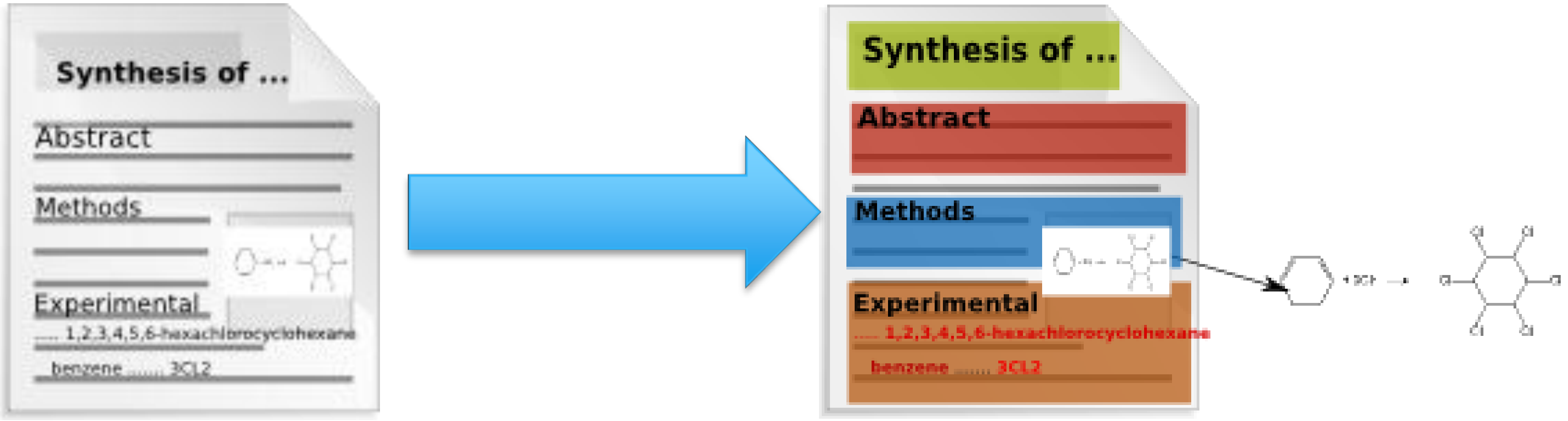
Methods

Experimental

..... 1,2,3,4,5,6-hexachlorocyclohexane

benzene 3Cl2

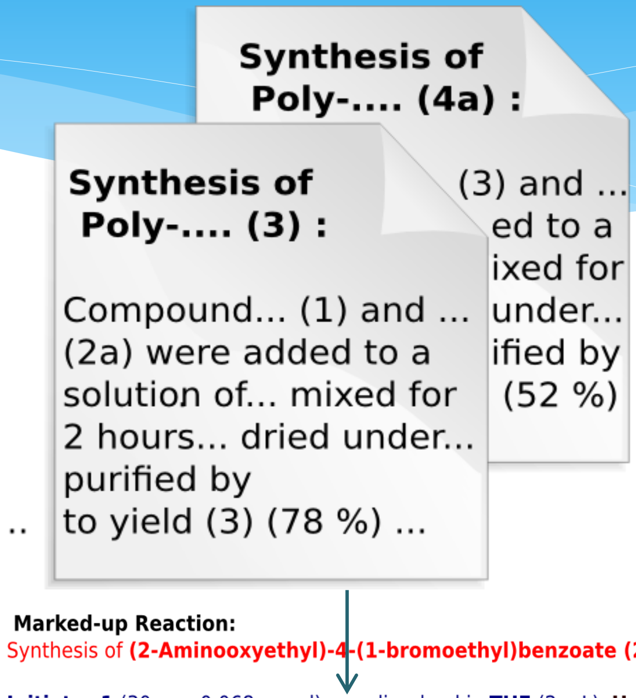
Aim: Enhance Data



ChemicalTagger

ChemicalTagger

- Open Source tool for text-mining chemistry
- OSCAR + NLP tools
- Extendible and adaptable tagging and parsing components
- Converts free flowing text to structured text



Marked-up Reaction:

Synthesis of **(2-Aminooxyethyl)-4-(1-bromoethyl)benzoate (2)**:

Initiator 1 (30 mg, 0.068 mmol) was dissolved in **THF** (2 mL). **Hydrazine hydrate** (10 uL, 0.34 mmol) was added via syringe, and the solution was refluxed for 2.5 h. The flask was cooled to room temperature and then filtered through a 20 m syringe filter. The filtrate was condensed in vacuo and the crude product was purified by column chromatography using **dichloromethane/ethyl acetate** (5/1) as the mobile phase to yield 42 mg (80%) of **2** as a white solid.

Legend:
Dissolve-Phrase
Add-Phrase
Degass-Phrase
Cool-Phrase
Filter-Phrase
Condense-Phrase
Purify-Phrase
Yield-Phrase

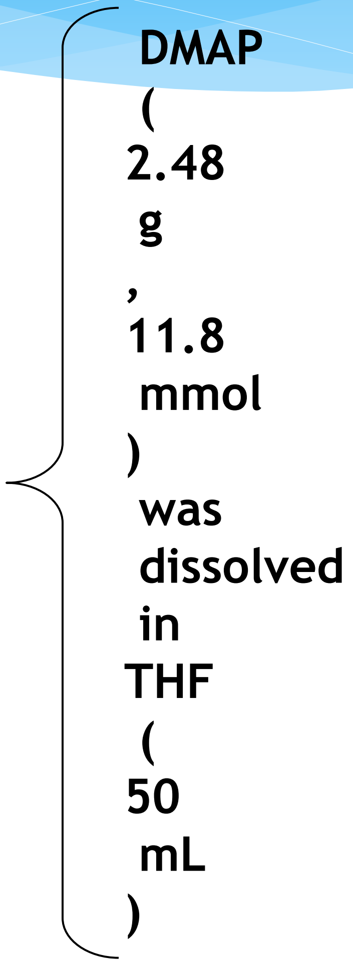
ChemicalTagger Components

- Tokenisers:
  - Split sequence into individual tokens
- Taggers:
  - Assign parts-of-speech to each token
- Parser:
  - Groups tagged tokens into phrases
- Role Identifier:
  - Assigns roles to the parsed phrases

Tokenisers

- Split a phrase into individual tokens:
- OSCAR-Tokeniser and WhiteSpaceTokeniser

DMAP (2.48 g, 11.8 mmol) was dissolved in THF (50 mL)



DMAP
(
2.48
g
,
11.8
mmol
)
was
dissolved
in
THF
(
50
mL
)

Taggers

- Assign parts of speech to a token
- Three step process:
  - OSCAR: Chemical Entities
  - Regex: Chemistry-related entities
  - OpenNLP: English entities

RegEx

Regular Expression

```
/h[a4@](([c<]((k)|(\|<)))|((k)|(\|<))|(x))\s+\
((d)|([t\+]h))[3ea4@]\s+p[l1][a4@]n[3e][t\+]/i
```

(C)2006 FTS Conventures - www.ftsconventures.com

openNLP

1. http://www.notaconmedia.com

Taggers

DMAP (2.48 g, 11.8 mmol) was dissolved in THF (50 mL)

Taggers: OSCAR

DMAP (2.48 g , 11.8 mmol) was dissolved in THF (50 mL)

Taggers: Regex

DMAP (2.48 g , 11.8 mmol) was dissolved in THF (50 mL)

ChemicalTagger

Taggers: OpenNLP

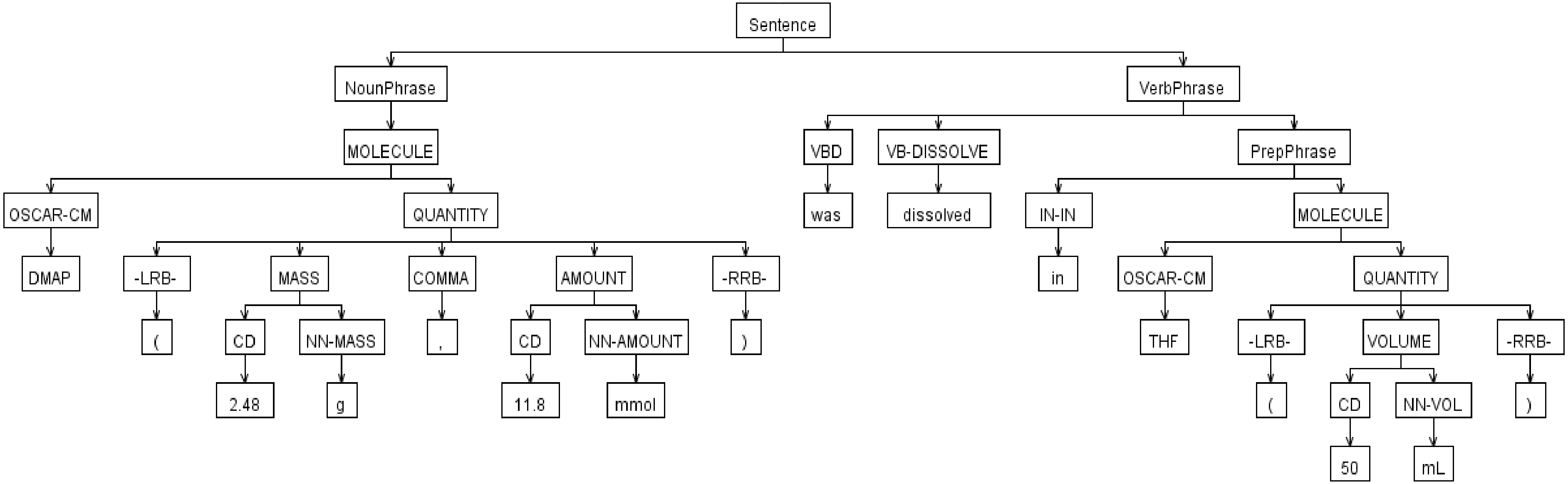
DMAP (2.48 g , 11.8 mmol) was dissolved in THF (50 mL)

Parser

- Converts tagged text into formal representations.
- Rule-Based Grammar

  - Sentence : Noun-Phrase Verb-Phrase
  - Noun-Phrase : dt? adj? NOUN+
  - Verb-Phrase : verb+ Prep-Phrase
  - Prep-Phrase : prep Noun-Phrase
  - NOUN : MOLECULE, nn, nns …
  - MOLECULE : oscar-cm and AMOUNT
  - AMOUNT : numbers, units (e.g. ml, grams, mols) between brackets

Parser

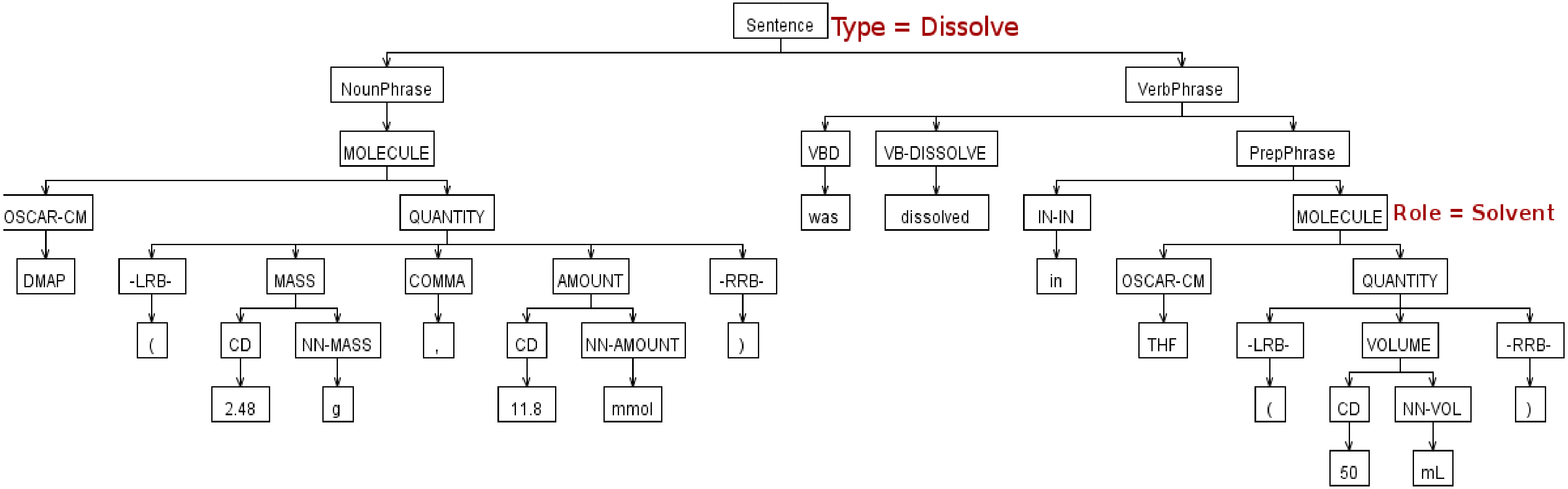


Sentence
NounPhrase
VerbPhrase
MOLECULE
VBD
VB-DISSOLVE
PrepPhrase
OSCAR-CM
QUANTITY
was
dissolved
IN-IN
MOLECULE
DMAP
-LRB-
MASS
COMMA
AMOUNT
-RRB-
in
OSCAR-CM
QUANTITY
(
CD
NN-MASS
,
CD
NN-AMOUNT
)
THF
-LRB-
VOLUME
-RRB-
2.48
g
11.8
mmol
(
CD
NN-VOL
)
50
mL

Role Identifier

- Post-process parse trees
- Adds ‘Action Roles’ to phrases:
  - 21 types of ‘Action Roles’
  - E.g.: adding, dissolving, purifying, yielding etc…
- Adds ‘roles’ to molecules (e.g: solvents):
  - Dissolve-Phrase:
    - DMAP (2.48 g, 11.8 mmol) was dissolved in THF (50 mL)
  - Wash-Phrase:
    - The organic extracts were washed with brine

Role Identifier



ChemicalTagger: Input

Methyl(2S,3S,αS)-2-hydroxy-3-[N-benzyl-N-(α-methylbenzyl)amino]-4-tri-iso-propylsilyloxy-butanoate 33

BuLi (2.5 M in hexanes , 8.14 mL , 11.4 mmol) was added dropwise via syringe to a stirred solution of (S)-N-benzyl-N-(α-methylbenzyl)amine (2.48 g , 11.8 mmol) in THF (50 mL) at -78 °C . After stirring for 30 min a solution of 28 (2.0 g , 7.35 mmol) in THF (20 mL) at -78 °C was added dropwise via cannula . After stirring for a further 2 h at -78 °C the reaction mixture was quenched with (+)-CSO (3.37 g , 14.7 mmol) and allowed to warm to rt over 12 h . Sat aq NH4Cl (5 mL) was added and the mixture was stirred for 5 min before being concentrated in vacuo . The residue was partitioned between DCM (50 mL) and 10 % aq citric acid (10 mL) . The organic layer was separated and the aqueous layer was extracted with DCM (2 × 50 mL) . The combined organic extracts were washed sequentially with sat aq NaHCO3 (50 mL) and brine (50 mL) , dried and concentrated in vacuo . The residue was dissolved in Et2O (50 mL) , the insoluble CSO residues were filtered off , and the filter cake was washed with Et2O (2 x 20 mL) . The filtrate was concentrated in vacuo and the process was repeated . Purification via flash column chromatography (eluent 30-40 °C petrol- Et2O , 20 : 1) gave 33 as a colourless oil (2.75 g , 75 % , > ; 98 % de)

ChemicalTagger: Output

Methyl(2S,3S,αS)-2-hydroxy-3-[N-benzyl-N-(α-methylbenzyl)amino]-4-tri-iso-propylsilyloxy-butanoate 33

BuLi (2.5 M in hexanes , 8.14 mL , 11.4 mmol) was **added** dropwise via syringe to a **stirred** solution of (S)-N-benzyl-N-(α-methylbenzyl)amine (2.48 g , 11.8 mmol) **in** THF (50 mL) at -78 °C . After **stirring** for 30 min a solution of 28 (2.0 g , 7.35 mmol) **in** THF (20 mL) at -78 °C was **added** dropwise via cannula . After **stirring** for a further 2 h at -78 °C the reaction mixture was **quenched** with (+)-CSO (3.37 g , 14.7 mmol) and allowed to **warm** to rt over 12 h . Sat aq NH4Cl (5 mL) was **added** and the mixture was **stirred** for 5 min before being **concentrated** in vacuo . The residue was **partitioned** between DCM (50 mL) and 10 % aq citric acid (10 mL) . The organic layer was **separated** and the aqueous layer was **extracted** with DCM (2 × 50 mL) . The combined organic extracts were **washed** sequentially with sat aq NaHCO3 (50 mL) and brine (50 mL) , dried and **concentrated** in vacuo . The residue was **dissolved** in Et2O (50 mL) , the insoluble CSO residues were **filtered** off , and the filter cake was **washed** with Et2O (2 x 20 mL) . The filtrate was **concentrated** in vacuo and the process was repeated . **Purification** via flash column chromatography (eluent 30-40 °C petrol- Et2O , 20 : 1) **gave** 33 as a colourless oil (2.75 g , 75 % , > ; 98 % de)

Evaluation

- Corpus of 50 experimental paragraphs
- Four Annotators and ChemicalTagger
- Annotation guidelines
- Inter-Annotator agreement as well as Machine-Annotators agreement
- Three types of evaluation:
  - Action name agreement
  - Filtered Phrase agreement
  - Phrase agreement using Sequence Alignment
- Similarity measured by Dice Coefficient

Evaluation: Action Name Agreement (%)

Annotator	1	2	3	4	Chemical Tagger
1	-	91.4	94.0	94.3	92.1
2	91.4	-	92.2	92.5	91.5
3	94.0	92.2	-	94.0	92.0
4	94.3	92.5	94.0	-	92.2
Chemical Tagger	92.1	91.5	92.0	92.2	-
Machine-Annotator Agreement	91.9				
Inter-annotator Agreement	93.1				

Evaluation : Filtered Phrase Agreement (%)

Annotator	1	2	3	4	Chemical Tagger
1	-	75.1	70.2	75.0	61.4
2	75.1	-	77.6	80.0	60.7
3	70.2	77.6	-	79	56.5
4	75.0	80.0	79.0	-	63.0
Chemical Tagger	61.4	60.7	56.5	63.0	-
Machine-Annotator Agreement	60.0				
Inter-annotator Agreement	76.2				

Evaluation: Phrase Alignment

Annotator A

1. to a 25 ml three-necked round-bottomed flask fitted with a dean-stark trap, a condenser, and a nitrogen inlet / outlet and magnetic stirrer

2. stirring the reaction mixture overnight at room temperature

3. evaporation of the eluate

4. afforded 8 as a white solid (2.63 g 57 % yield)

Annotator B

1. a 25 ml three-necked round-bottomed flask fitted with a dean-stark trap, a condenser, and a nitrogen inlet / outlet

2. after stirring the reaction mixture overnight at room temperature

3.which then afforded 8 as a white solid (2.63 g 57 % yield)

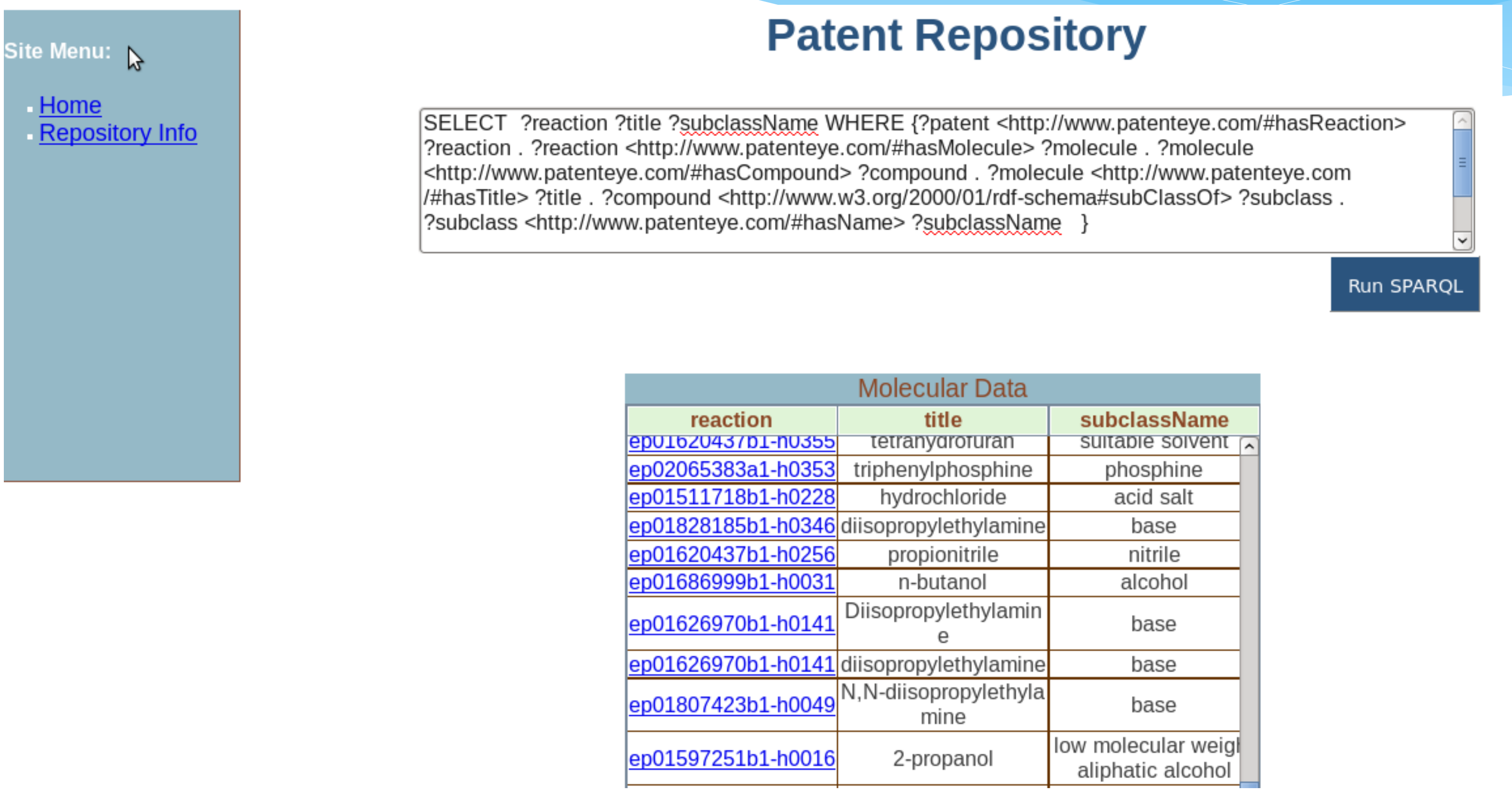
Evaluation: Sequence Alignment

- Used in Bioinformatics for protein and nucleotide alignment.
- Needleman-Wunsch algorithm
- Comparing pairs of sequences and computing a score measurement
- Example ABC and ABBC :
  - AB_C
  - ABBC

Evaluation : Phrase Alignment Agreement(%)

Annotator	1	2	3	4	Chemical Tagger
1	-	90.2	89.2	91.1	88.4
2	90.2	-	90.8	91.6	89.8
3	89.2	90.8	-	91.6	87.2
4	91.1	91.6	91.6	-	90.2
Chemical Tagger	88.4	89.8	87.2	90.2	-
Machine-Annotator Agreement	88.9				
Inter-annotator Agreement	90.8				

Applications: Reaction Repositories



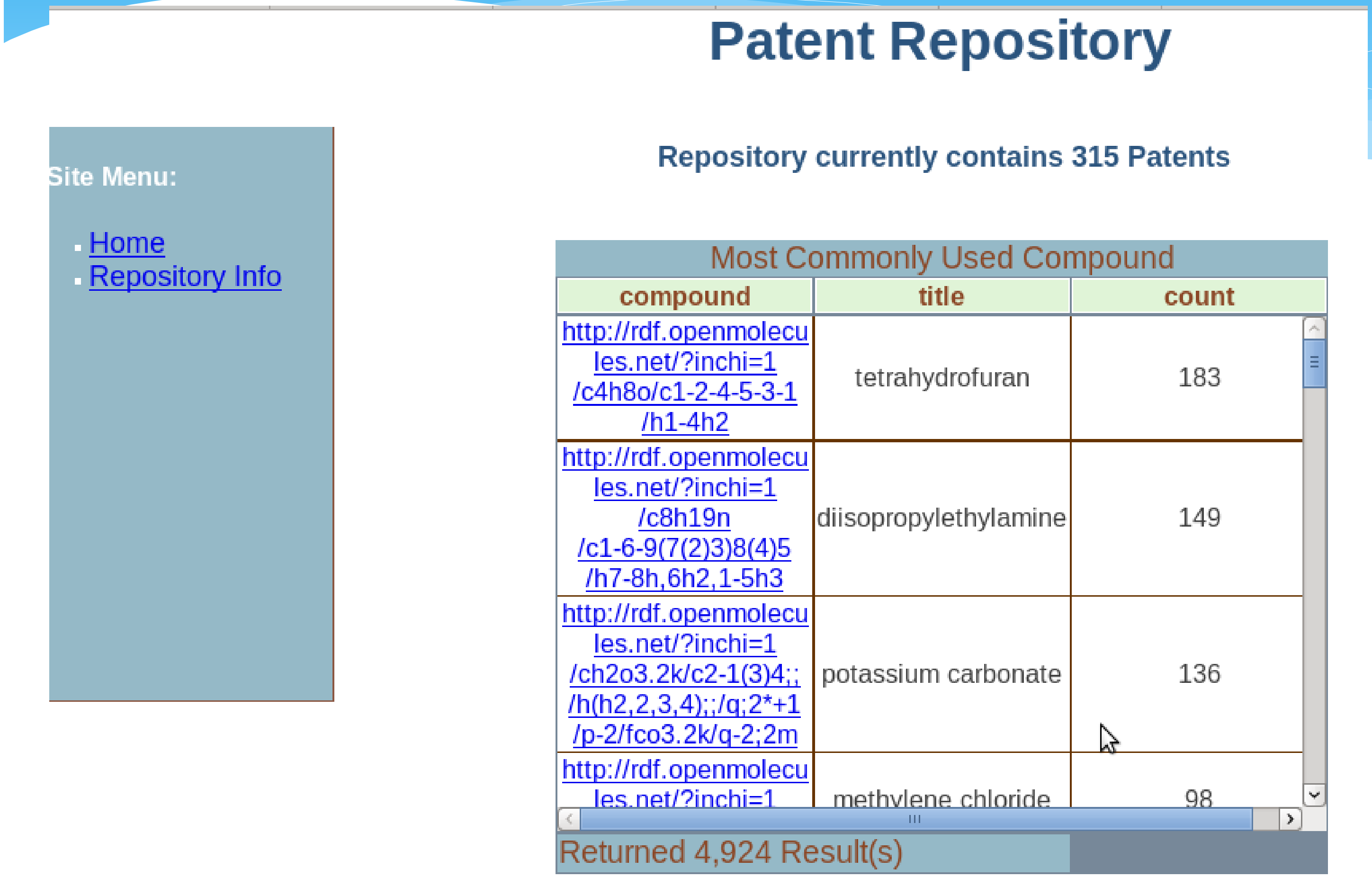
Patent Repository

```
SELECT  ?reaction ?title ?subclassName WHERE {?patent <http://www.patenteye.com/#hasReaction>
?reaction . ?reaction <http://www.patenteye.com/#hasMolecule> ?molecule . ?molecule
<http://www.patenteye.com/#hasCompound> ?compound . ?molecule <http://www.patenteye.com
/#hasTitle> ?title . ?compound <http://www.w3.org/2000/01/rdf-schema#subClassOf> ?subclass .
?subclass <http://www.patenteye.com/#hasName> ?subclassName   }
```

Molecular Data

reaction	title	subclassName
ep01620437b1-h0355	tetrahydrofuran	suitable solvent
ep02065383a1-h0353	triphenylphosphine	phosphine
ep01511718b1-h0228	hydrochloride	acid salt
ep01828185b1-h0346	diisopropylethylamine	base
ep01620437b1-h0256	propionitrile	nitrile
ep01686999b1-h0031	n-butanol	alcohol
ep01626970b1-h0141	Diisopropylethylamine	base
ep01626970b1-h0141	diisopropylethylamine	base
ep01807423b1-h0049	N,N-diisopropylethylamine	base
ep01597251b1-h0016	2-propanol	low molecular weigh aliphatic alcohol

Applications: Reaction Repositories



Patent Repository

Site Menu:

- Home
- Repository Info

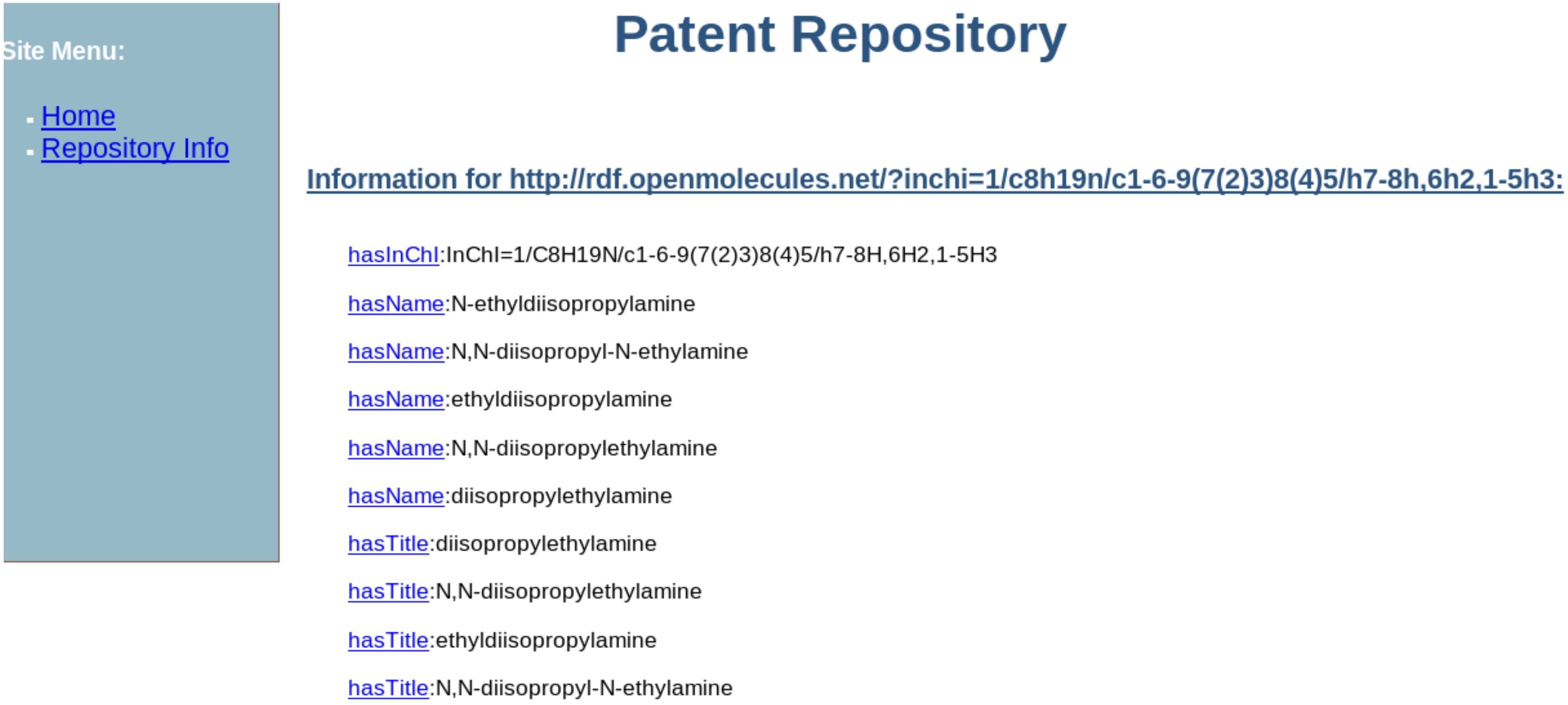
Repository currently contains 315 Patents

Most Commonly Used Compound

compound	title	count
http://rdf.openmolecules.net/?inchi=1/c4h8o/c1-2-4-5-3-1/h1-4h2	tetrahydrofuran	183
http://rdf.openmolecules.net/?inchi=1/c8h19n/c1-6-9(7(2)3)8(4)5/h7-8h,6h2,1-5h3	diisopropylethylamine	149
http://rdf.openmolecules.net/?inchi=1/ch2o3.2k/c2-1(3)4;;/h(h2,2,3,4);;/q;2*+1/p-2/fco3.2k/q-2;2m	potassium carbonate	136
http://rdf.openmolecules.net/?inchi=1	methylene chloride	98

Returned 4,924 Result(s)

Applications: Reaction Repositories



Patent Repository

Information for http://rdf.openmolecules.net/?inchi=1/c8h19n/c1-6-9(7(2)3)8(4)5/h7-8h,6h2,1-5h3:

hasInChI:InChI=1/C8H19N/c1-6-9(7(2)3)8(4)5/h7-8H,6H2,1-5H3

hasName:N-ethyldiisopropylamine

hasName:N,N-diisopropyl-N-ethylamine

hasName:ethyldiisopropylamine

hasName:N,N-diisopropylethylamine

hasName:diisopropylethylamine

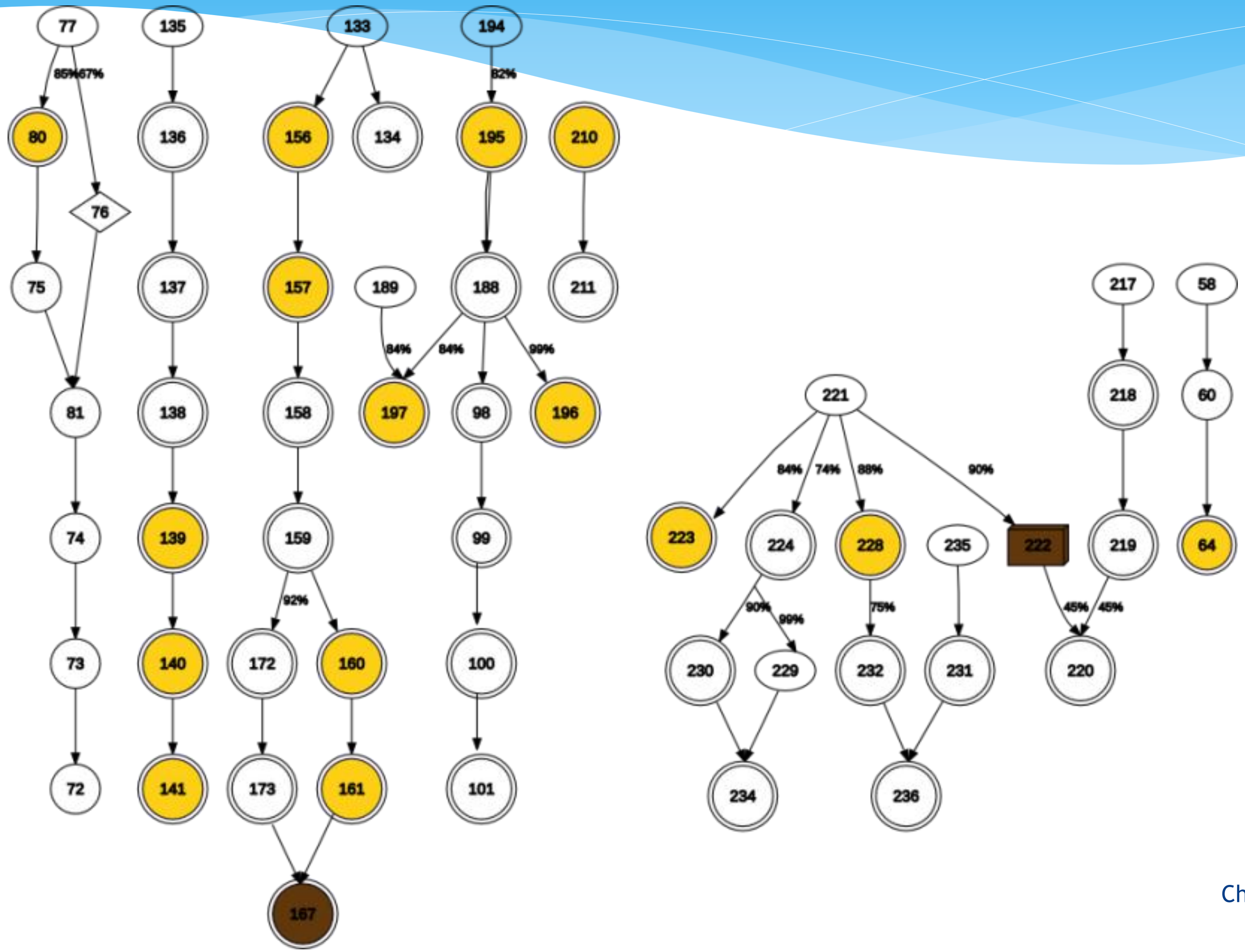
hasTitle:diisopropylethylamine

hasTitle:N,N-diisopropylethylamine

hasTitle:ethyldiisopropylamine

hasTitle:N,N-diisopropyl-N-ethylamine

Applications: Visualisations



ChemicalTagger

Acknowledgements

- Dr. Nico Adams
- Nicholas England
- Dr. Colin Batchelor
- Dr. Egon Willighagen
- Unilever
- JISC

UNIVERSITY OF
CAMBRIDGE

Unilever
Cambridge
Centre For Molecular Science Informatics

Thank you…

Questions and Comments?

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre For Molecular Science Informatics