OSCAR4

Architecture and API

Sam Adams, David Jessop, Lezan Hawizy, Egon Willighagen
Peter Murray-Rust

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Science Informatics

Refactoring OSCAR3

Improve maintainability
and extensibility

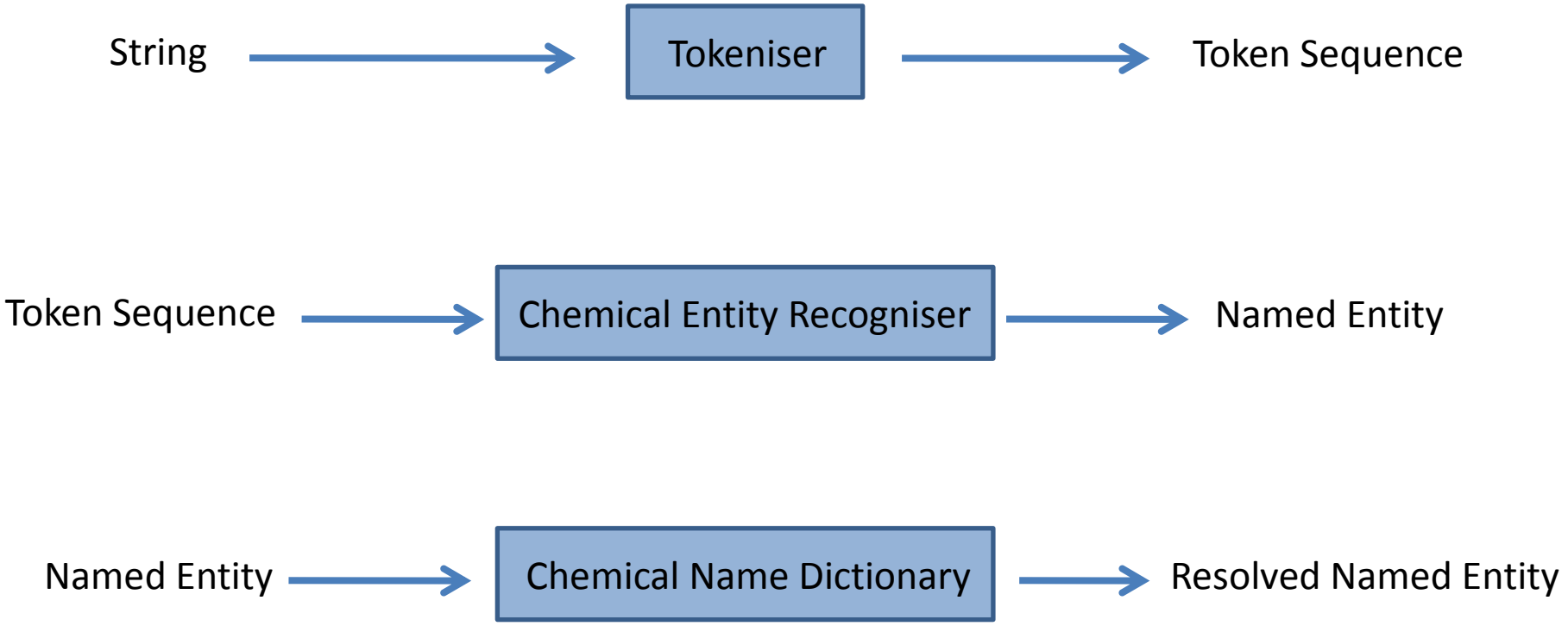
Modularisation, documentation
performance, testing

Straightforward to use, debug, extend, distribute

'convention over configuration'

OSCAR4

Architecture:
an extensible library


String
Tokeniser
Token Sequence
Token Sequence
Chemical Entity Recogniser
Named Entity
Named Entity
Chemical Name Dictionary
Resolved Named Entity

Recognisers

Pattern Recogniser

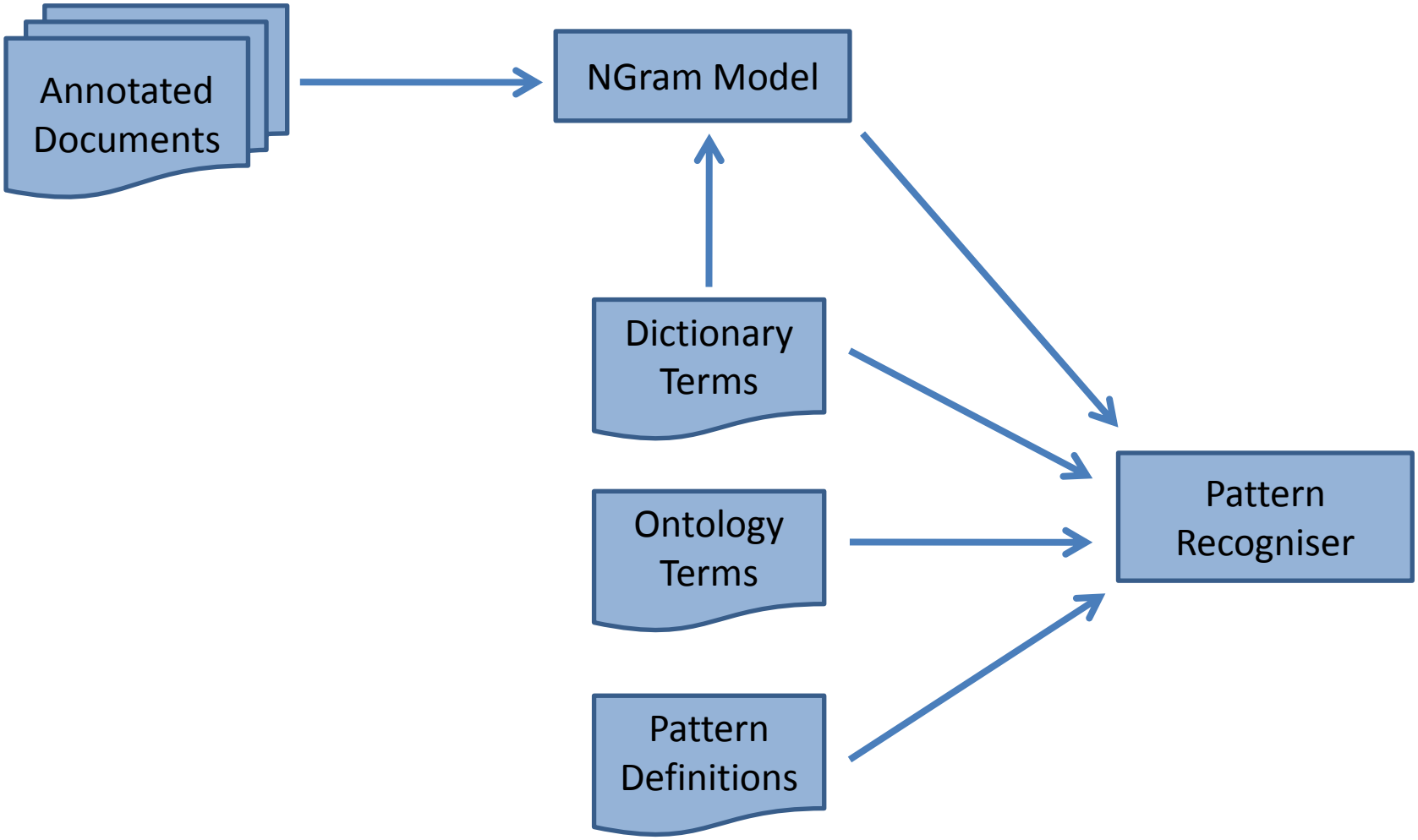
MEMM Recogniser

ChemPapers Model

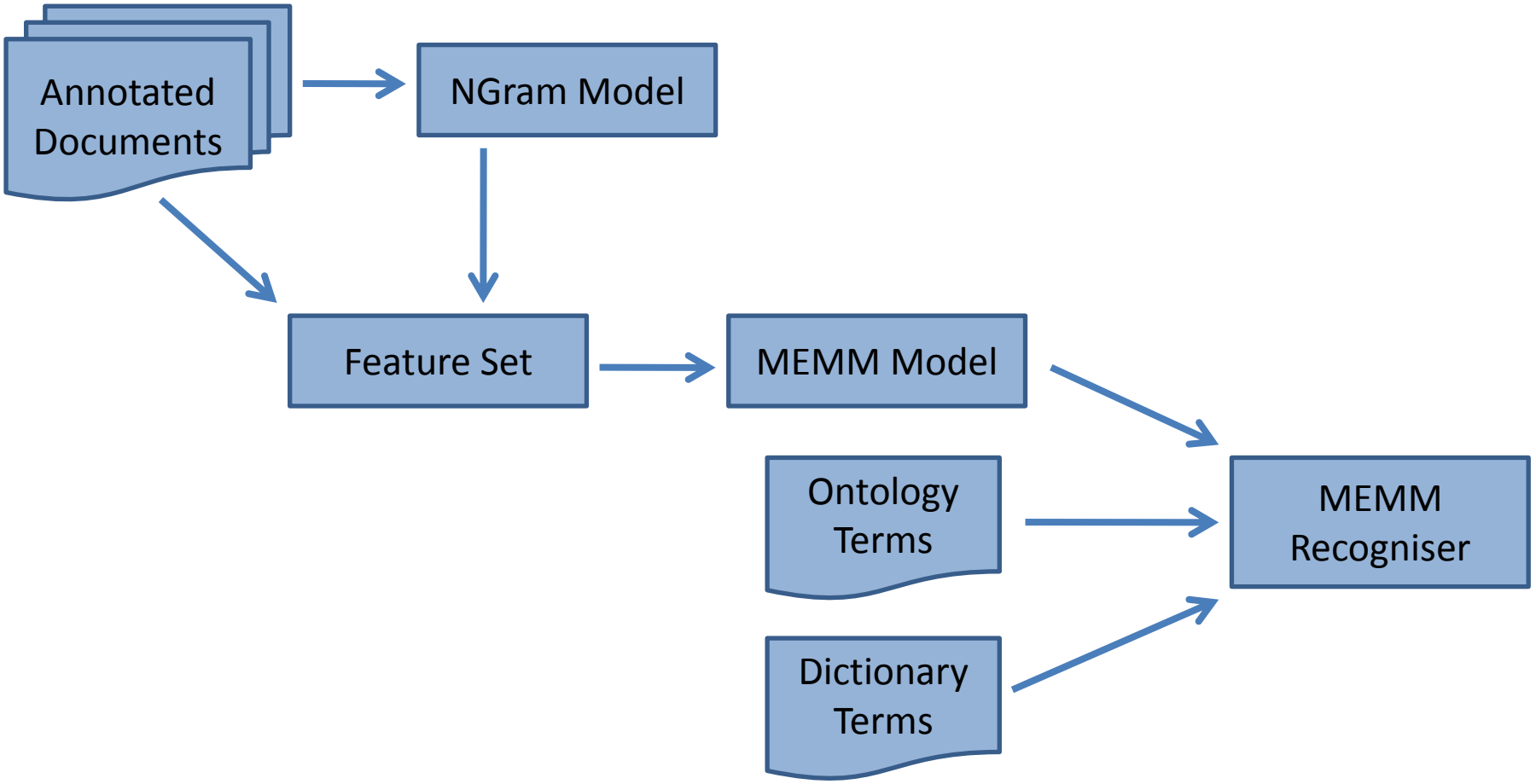
Pubmed Model

RegexRecogniser

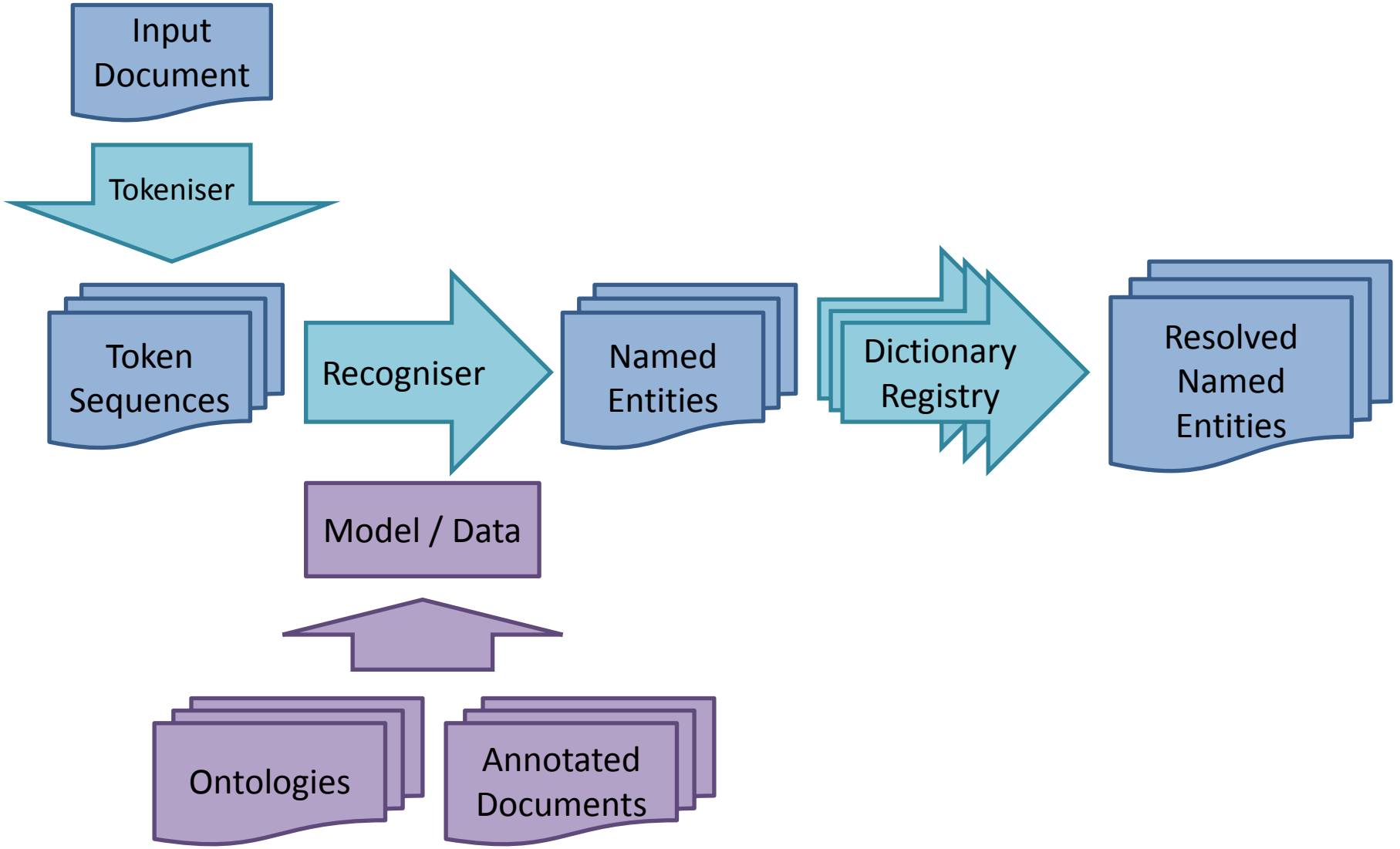
Pattern Recogniser



MEMM Recogniser



Full Workflow

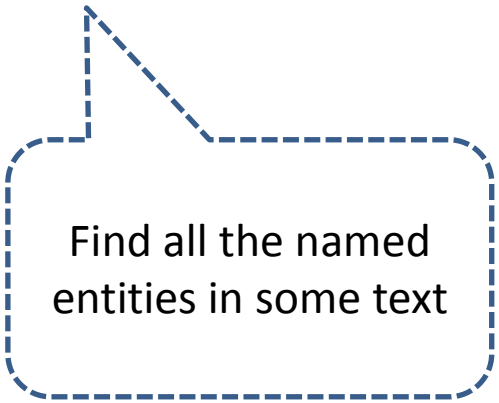


OSCAR4

API:

convention over configuration

```
Oscar oscar = new Oscar();
List<NamedEntity> namedEntities
      = oscar.findNamedEntities(s);
```

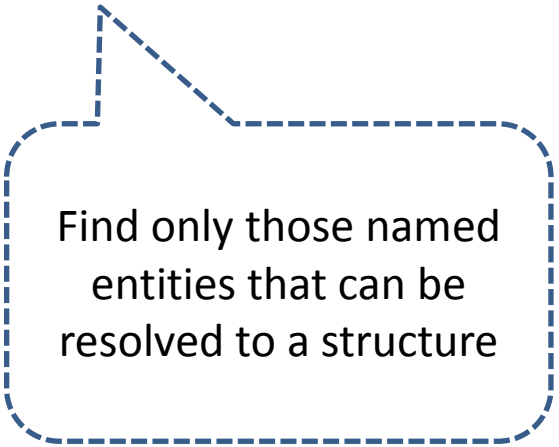


```
Oscar oscar = new Oscar();
List<ResolvedNamedEntity> entities
      = oscar.findAndResolveNamedEntities(s);

for (ResolvedNamedEntity entity : entities) {
      ChemicalStructure structure
            = entity.getFirstChemicalStructure(
                  FormatType.INCHI));
      ...
}
```

Find all the named entities in some text, and get their InChIs when available

```
Oscar oscar = new Oscar();
List<ResolvedNamedEntity> entities
      = oscar.findResolvableEntities(s);
```



```
ChemicalEntityRecogniser myRecogniser
      = new PatternRecogniser()

Oscar oscar = new Oscar();
oscar.setRecogniser(myRecogniser);
oscar.setDictionaryRegistry(
            myDictionaryRegistry);
List<ResolvedNamedEntity> entities
      = oscar.findResolvableEntities(s);
```

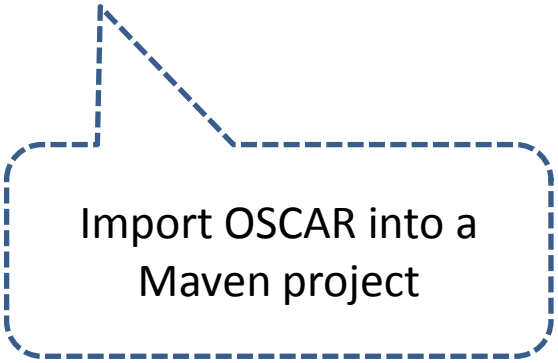
Use a different recogniser & dictionary

```
<project xmlns="http://maven.apache.org/POM/4.0.0"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    [ ... ]

    <repositories>
        <repository>
            <id>ucc-repo</id>
            <url>http://maven.ch.cam.ac.uk/m2repo</url>
        </repository>
    </repositories>

    <dependencies>
        <dependency>
            <groupId>uk.ac.cam.ch.wwmm.oscar</groupId>
            <artifactId>oscar4-api</artifactId>
            <version>4.0.1</version>
        </dependency>
    </dependencies>

</project>
```



OSCAR4

API:
reading data

Experimental Data Checker v2.6.1

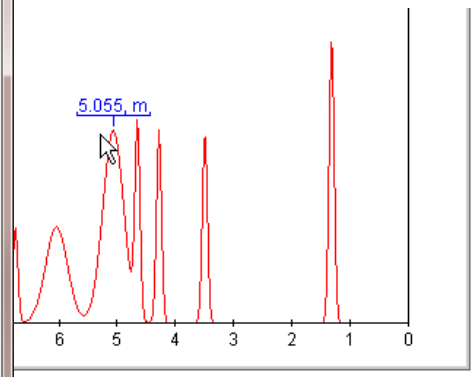
File Edit View Help

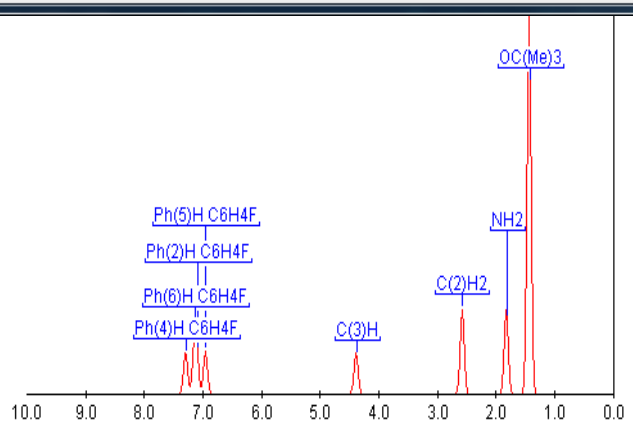
Preparation of tert-butyl (E)-3-(3-fluorophenyl)prop-2-enoate 2
Following general procedure 1, tert-butyl diethylphosphonoacetate (2.73 g, 10.83 mmol), n-BuLi (2.5 M, 4.2 ml, 10.3 mmol) in THF (10 ml) and 3-fluorobenzaldehyde (1.22 g, 9.85 mmol) in THF (10 ml) gave, after purification by column chromatography on silica gel (hexane-Et2O 40 : 1), 2 (2.01 g, 92%) as a colourless oil; vmax (film) 2980 (CH), 1712 (Cdouble bond, length as m-dashO); δH (400 MHz, CDCl3) 1.56 [9H, s, OC(Me)3], 6.38 [1H, d, J 16.0, C(2)H], 7.07 [1H, m, Ph(5)H C6H4F], 7.18-7.38 [3H, m, Ph(2)H, Ph(4)H and Ph(6)H C6H4F], 7.52 [1H, d, J 16.0, C(3)H]; δC (100 MHz, CDCl3) 28.6, 81.2, 114.6, 117.2, 122.1, 124.3, 130.8, 137.4, 142.5, 163.3, 166.5; m/z (CI+) 223 (MH+, 50%), 166 (MH+ – C4H8 100%); HRMS (CI+) C13H16FO2 requires 223.1134, found 223.1133.

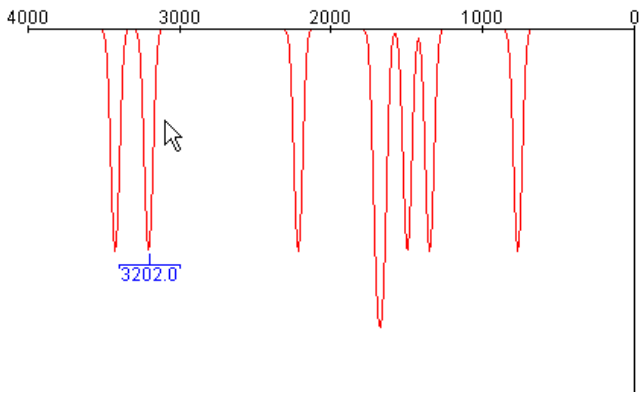
Preparation of tert-butyl (E)-3-(2-iodophenyl)prop-2-enoate 3
Following general procedure 1, tert-butyl diethylphosphonoacetate (5.0 g, 19.8 mmol), n-BuLi (1.6 M, 11.85 ml, 19.0 mmol) in THF (20 ml) and 2-iodobenzaldehye (4.0 g, 17.2 mmol) in THF (10 ml) gave, after purification by column chromatography on silica gel (hexane-Et2O 40 : 1), 3 (5.39 g, 93%) as a yellow oil; vmax (film) 2977 (CH), 1708 (Cdouble bond, length as m-dashO), 1637 (Cdouble bond, length as m-dashC); δH (400 MHz, CDCl3) 1.55 [9H, s, OC(Me)3], 6.25 [1H, d, J 15.7, C(2)H], 7.04 [1H, t, J 7.6, Ph(4)H C6H4I], 7.34 [1H, t, J 7.6, Ph(5)H C6H4I], 7.56 [1H, d, J 7.6, Ph(6)H C6H4I], 7.83 [1H, d, J 15.7, C(3)H], 7.90 [1H, d, J 7.6, Ph(3)H C6H4I]; δC (100 MHz, CDCl3) 28.6, 81.2, 101.7, 123.4, 127.7, 128.9, 130.7, 135.9, 140.4, 147.1, 166.0; m/z (CI+) 331 (MH+, 10%), 348 (MNH4+, 30%); HRMS (CI+) C13H16IO2 requires 331.0195, found 331.0194.

Preparation of tert-butyl (E)-3-(3-iodophenyl)prop-2-enoate 4
Following general procedure 1, tert-butyl diethylphosphonoacetate (4.1 g, 16.3 mmol), n-BuLi (2.5 M, 6.2 ml, 15.5 mmol) in THF (15 ml) and 3-iodobenzaldehye (3.43 g, 14.8 mmol) in THF (15 ml) gave, after purification by column chromatography on silica gel (hexane-Et2O 40 : 1), 4 (4.15 g, 85%) as a yellow oil; vmax (film) 1707 (Cdouble bond, length as m-dashO), 1638 (Cdouble bond, length as m-dashC); δH (400 MHz, CDCl3) 1.54 [9H, s, OC(Me)3], 6.36 [1H, d, J 16.0, C(2)H], 7.12 [1H, t, J 7.8, Ph(5)H C6H4I], 7.47 [1H, d, J 16.0, C(3)H], 7.48 [1H, d, J 8.0, Ph(6)H C6H4I], 7.69 [1H, d, J 8.3, Ph(4)H C6H4I], 7.87 [1H, s, Ph(2)H C6H4I]; δC (100 MHz, CDCl3) 28.6, 81.2, 95.1, 122.0, 127.6, 130.9, 137.0, 137.3, 139.1, 142.1, 166.2; m/z (CI+) 331 (MH+, 40%), 348 (MNH4+, 35%); HRMS (CI+) C13H16IO2 requires 331.0195, found 331.0197.

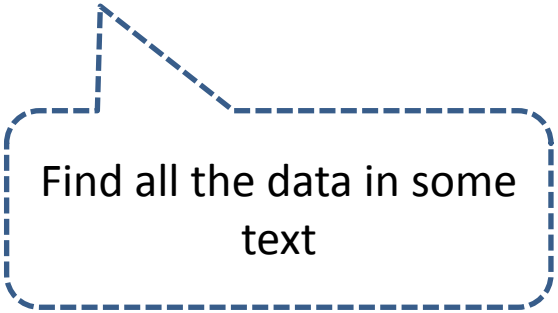
Preparation of tert-butyl (E)-3-(4-iodophenyl)prop-2-enoate115
Following general procedure 1, tert-butyl diethylphosphonoacetate (3.9 g, 15.5 mmol), n-BuLi (2.5 M, 9.25 ml, 14.8 mmol) in THF (20 ml) and 4-iodobenzaldehye (3.1 g, 13.5 mmol) in THF (20 ml) gave, after purification by column chromatography on silica gel (hexane-Et2O 40 : 1) and recrystallisation (hexane-Et2O), 5 (4.2 g, 94%) as white needles; mp 65-66 °C (hexane-Et2O); δH (400 MHz, CDCl3) 1.54 [9H, s, OC(Me)3], 6.38 [1H, d, J 16.0, C(2)H], 7.24 [2H, m, Ph(2)H and Ph(6)H C6H4I], 7.51 [1H, d, J 16.0, C(3)H], 7.72 [2H, d, J 8.5, Ph(3)H and Ph(5)H C6H4I].







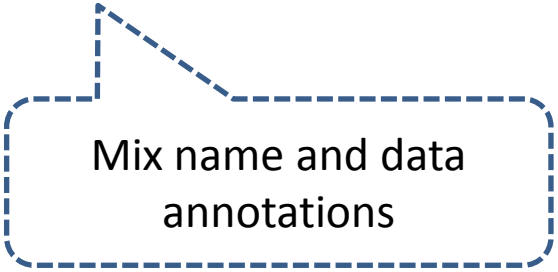
```
OscarData oscarData = new OscarData();
List<DataAnnotation> data
      = oscarData.findData(s);
```



```
Oscar oscar = new Oscar();
OscarData oscarData = new OscarData();

List<Annotation> annotations
      = new ArrayList<Annotation>();
annotations.addAll(oscarData.findData(s));
annotations.addAll(
      oscar.findAndResolveNamedEntities(s));

Annotation.getStart();
Annotation.getEnd();
Annotation.getSurface();
```



OSCAR4

CLI Application

```
$ java -jar oscar4-cli.jar "Then we mix benzene with toluene."
INFO - Initialising OPSIN...
INFO - OPSIN initialised
benzene: InChI=1/C6H6/c1-2-4-6-5-3-1/h1-6H
toluene: InChI=1/C7H8/c1-7-5-3-2-4-6-7/h2-6H,1H3
```

```
$ cat article.html | java -jar oscar4-cli.jar -stdin -html
INFO - Initialising OPSIN...
INFO - OPSIN initialised
carbon: InChI=1/C
HBr: InChI=1/BrH/h1H
acetic acid: InChI=1/C2H4O2/c1-2(3)4/h1H3,(H,3,4)/f/h3H
mercury(II) cyanide: InChI=1/2CN.Hg/c2*1-2;
nitromethane: InChI=1/CH3NO2/c1-2(3)4/h1H3
acetic acid: InChI=1/C2H4O2/c1-2(3)4/h1H3,(H,3,4)/f/h3H
...
```

OSCAR4

Conclusion

OSCAR4 compared to OSCAR3

Core Features

Find Named Entities

Resolve Structures

Find Data

Separate Applications

CLI Application

Workflow Components

ChemicalTagger

Upgrades

Straightforward to use

Easily extensible

Configurable

Performance & Reliability

Independent Tools

Annotation

Missing

XML Input & Output

Server / Workbench

Future Plans

XHTML Support

(customisable for alternative schema)

Web Application

Further improvements to API

More work on model generation

Additional recognisers

Bug fixing!

http://bitbucket.org/wwmm/oscar4

source code

documentation

bug reporting

feature requests