SPECTRa-T Project

Alan Tonge

Semantic Web Data Repositories from Chemistry e-Thesis Data Mining

Open Repositories 2008
Southampton University
2 April 2008

Project Overview

Submission, Preservation and Exposure of Chemistry Teaching and Research Data – in Theses

- **12-month project between University of Cambridge and Imperial College London to develop text- and data-mining tools to extract chemical data from e-theses**
- **Part of the JISC Digital Repositories programme**

UNIVERSITY OF CAMBRIDGE

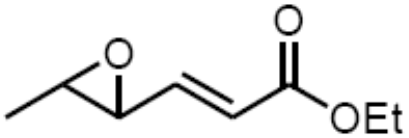
Imperial College London

SPECTRa-T

Background

Chemistry is an experimental science

Synthetic Organic Chemistry is the basis of Pharmaceutical and Agrochemical industries



Where does the information to make this molecule come from?

Systematic Name : **Ethyl 4,5-epoxy-hex-2-enolate**

Molecular Formula : **C8H12O3**

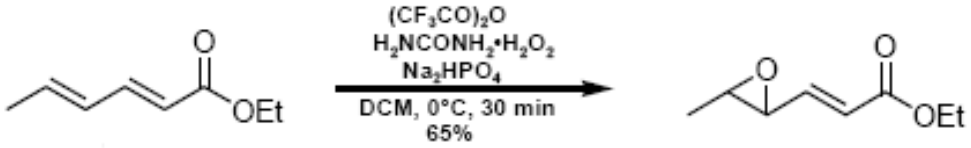
Search Chemical patent & journal abstracting services – *e.g.*

Chemical Abstracts (*9000+ journals - 12,000 structures/day*)
Beilstein (180 core journals)
Patents (CAS, Derwent, MDL) (400,000 /annum)

Academic chemistry publications largely derived from PhD Theses

Perhaps ~10K published per year worldwide

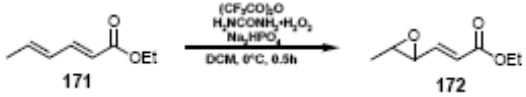
Synthetic : contains 50-60 preparations – only 20% published in detail



- **List of Starting Materials & Reagents**
- **Recipe: Reactions Conditions & Work-up**
- **Product Characterization – spectroscopic & physical properties**

Sample preparation from synthetic chemistry thesis

7.3.1 Preparation of (2*E*,4*R**,5*R**)-ethyl-4,5-epoxy-hex-2-enoate (**172**)



171 172

Trifluoroacetic anhydride (14.8 ml, 104 mmol) was added slowly to a suspension of (2*E*,4*E*) ethyl hexa-2,4-dienoate **171** (2.44 g, 17.4 mmol), urea hydrogen peroxide addition compound (37.7 g, 0.39 mol) and disodium hydrogenphosphate (27.6 g, 195 mmol) in DCM (250 ml) at 0°C. After removing from the ice bath, the reaction mixture was stirred at rt for 30 min and then cautiously poured into a vigorously stirred and precooled (0°C) solution of $NaHCO_3$ (800 ml). After effervescence had ceased, the phases were separated and the organic phase washed sequentially with $NaHCO_3$ solution (3 x 300 ml) and NaCl solution (300 ml), dried ($MgSO_4$) and filtered. Concentration *in vacuo* followed by flash column chromatography (eluent PE:Et_2O 7:1) provided the epoxide **172** (1.09 g, 7 mmol, 41%) as a colourless oil; **ν_{max} (film)/cm^{-1}:** 2981, 1716 (C=O), 1655 (C=C), 1446, 1378, 1367, 1340, 1302, 1258, 1185, 1140, 1096, 1031, 1005, 975; **δ_H (400 MHz, $CDCl_3$):** 1.15 (3H, t, *J* 7.1, OCH_2CH_3), 1.24 (3H, d, *J* 5.2, 6-H x 3), 2.84 (1H, qd, *J* 5.2, 2.0, 5-H), 3.05 (1H, dd, *J* 7.0, 2.0, 4-H), 4.07 (2H, q, *J* 7.1, OCH_2CH_3), 5.99 (1H, dd, *J* 15.7, 0.6, 2-H), 6.54 (1H, dd, *J* 15.7, 7.0, 3-H); **δ_C (100 MHz, $CDCl_3$):** 165.5, 144.5, 123.6, 60.4, 57.3, 57.1, 17.4, 14.1; ***m/z* (+EI):** 179 ([MNa]$^+$, 100%). Found: [*M*Na]$^+$, 179.060. $[C_8H_{12}O_3Na]^+$ requires 179.0684. Data was consistent with those reported in the literature.

The Problem

- ~80% of (academic) synthetic preparations remain locked in theses
- Manual abstraction (*cf* journals/patents) not an option

The Solution

- **OSCAR3 : *Automatic*** high-throughput chemical name and chemical term recognition

  Open Source Chemistry Analysis Routines is an extensible Open Source framework which can identify much of the chemical terminology in electronic articles
- **Semantic Web :** Deposit extracted terms in searchable RDF triplestore

OSCAR Name recognition:

1. Dictionary of chemical names/terms (ChEBI Ontology)
2. Rules; chemical suffix filters
3. Regular expressions to recognise: data, formulae

7.3.1 Preparation of (2E,4R*,5R*)-ethyl-4,5-epoxy-hex-2-enoate (172)

Trifluoroacetic anhydride (14.8 ml, 104 mmol) was added slowly to a suspension of (2E,4E) ethyl hexa-2,4-dienoate 171 (2.44 g, 17.4 mmol), ureahydrogen peroxideaddition compound (37.7 g, 0.39 mol) and disodium hydrogenphosphate (27.6 g, 195 mmol) in DCM (250 ml) at 0° C. After removing from the ice bath, the reaction mixture was stirred at rt for 30 min and then cautiously poured into a vigorously stirred and precooled (0°C) solution of NaHCO3 (800 ml). After effervescence had ceased, the phases were separated and the organic phase washed sequentially with NaHCO3 solution (3 x 300 ml) and NaCl solution (300 ml), dried (MgSO4) and filtered. Concentration in vacuo followed by flash column chromatography (eluent PE:Et2O 7:1) provided the epoxide 172 (1.09 g, 7 mmol, 41%) as a colourless oil; □max (film)/cm-1: 2981, 1716 (C=O), 1655 (C=C), 1446, 1378, 1367, 1340, 1302, 1258, 1185, 1140, 1096, 1031, 1005, 975; □H (400 MHz, CDCl3): 1.15 (3H, t, J 7.1, OCH2CH3), 1.24 (3H, d, J 5.2, 6-H x 3), 2.84 (1H, qd, J 5.2, 2.0, 5-H), 3.05 (1H, dd, J 7.0, 2.0, 4-H), 4.07 (2H, q, J 7.1, OCH2CH3), 5.99 (1H, dd, J 15.7, 0.6, 2-H), 6.54 (1H, dd, J 15.7, 7.0, 3-H); □C (100 MHz, CDCl3): 165.5, 144.5, 123.6, 60.4, 57.3, 57.1, 17.4, 14.1; m/z (+EI): 179 ([MNa]+, 100%). Found: [MNa]+, 179.060. [C8H12O3Na]+ requires 179.0684. Data was consistent with those reported in the literature.16

- Experimental data
- Ontology term
- Chemical (etc.) with structure
- Chemical (etc.), without structure
- Reaction
- Chemical adjective
- enzyme -ase word
- Chemical prefix

Input: PDF Legacy Format

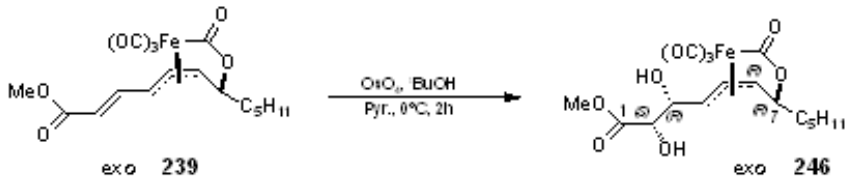
PDF is the *de facto* format for electronic document deposition in digital repositories

Problem:

PDF text is a Page Description Format –
optimized for *human*, not *machine*, readability

- irregular word order
- line-breaks: loss of continuous text; paragraphs difficult to identify
- loss of subscripts and superscripts
- non-printing characters
- erroneous character assignment with OCR.

Preparation of [(4*E*,2*S,3*R**,6*R**,7*R**)-7-(carbonyloxy-κC)-2,3-dihydroxy-2-methoxyoxycarbonyl-(4,5,6-η)-dodeca-4-en-6-yl]tricarbonyliron (246)**



exo 239 exo 246

Word processed document

7.5.11 Preparation [(4E,2S*,3R*,6R*,7R*)-7-(carbonyloxy-κC)-2,3-dihydroxy-2-methoxyoxycarbonyl-(4,5,6-η)-dodeca-4-en-6-yl]tricarbonyliron (246)
C5H11
O
(OC)3Fe
O
239exo
O
MeO
C5H11(R)
O
(OC)3Fe
O
246exo
(S)
O
MeO
OH
HOOsO
4,
tBuOH
Pyr., 0°C, 2h
(R)
(R)1 7

Disconnected text fragments output from processed PDF document

Programmatic modifications to:

- **Remove linebreaks from extended chemical names**
- **Remove text fragments derived from Figures and Tables**
- **Correct whitespace in chemical names**

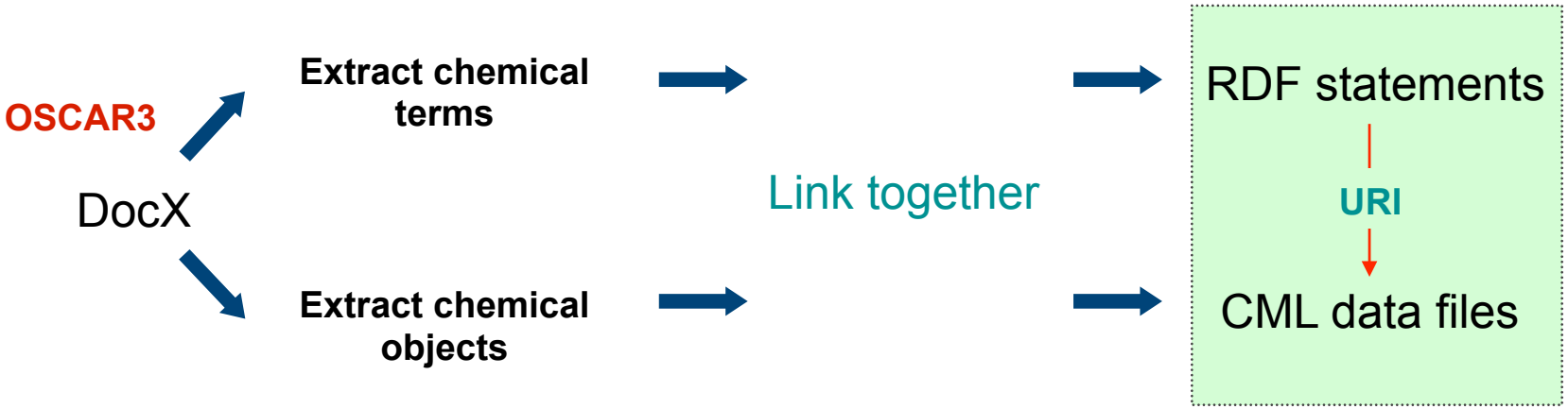
OSCAR used ‘as is’ on PDF e-theses :

Gives 5000 terms / thesis (80% duplicates)

Cannot identify chemical objects (spectra assignments; properties)

Input: MS Office Open XML – 'docx'

- No information loss from student's deposited thesis (written with MS software)
- Identification of experimental sections no longer a problem -> Chemical Objects
- Conversion of CO's into Chemical Markup Language



Data Repository

Sample preparation from chemistry thesis

7.4.3 Preparation of 5-chloro-4-hydroxy-3-methylene-6-phenyl-hex-2-one (**192**)

191 182 192

Freshly distilled methylvinylketone **182** (0.31 g, 0.37 ml, 4.4 mmol) was added dropwise to a stirred solution of the α-chlorophenylaldehyde **191** (0.74 g, 4.4 mmol) and DABCO catalyst (0.112 g, 1 mmol) in dry THF (6 ml) at rt. The solution was stirred for one week and then diluted with Et_2O (5 ml). The mixture was then washed with 3N-HCl (2 x 5 ml) and the aqueous layer extracted with Et_2O (2 x 10 ml). The combined organic fractions were washed with Na_2CO_3 solution (10 ml), dried ($MgSO_4$), filtered and then concentrated *in vacuo*. Purification was readily achieved by flash column chromatography on silica gel (eluent Et_2O:PE 1:2) and afforded an inseparable mixture of *cis* and *trans* diastereoisomeric (ratio 2:1) chloro-hydroxy-phenyl-enones **192** (0.30 g, 29%) as a dark yellow oil; **ν_{max} (film)/cm^{-1}:** 3446 br (OH), 3062, 3029, 2922, 1672 (C=O), 1604 (C=C), 1496, 1454, 1366, 1299, 1135, 1078, 974, 670; **δ_H (600 MHz, $CDCl_3$):** 2.35 (3H´, *s*, 1-H´ x 3), 2.37 (3H, *s*, 1-H x 3), 2.51 (1H´, d, *J* 9.0, OH´ x 1), 2.57 (1H´, m , 6-H_b´ x 1), 2.90 (1H, dd, *J* 15.0, 9.5, 6-H_b x 1), 3.21 (1H´, m, 6-H_a´ x 1), 3.33 (1H, dd, *J* 15.0, 3.0, 6-H_a x 1), 3.36 (1H, d, *J* 8.0, OH x 1), 4.37 (1H, td, *J* 11.0, 4.1, 5-H x 1), 4.45 (1H´, td, *J* 8.0, 2.0, 5-H´ x 1), 4.48 (1H, app. t, *J* 7.0, 4-H x 1), 4.71 (1H´, app. d, *J* 9.0, 4-H´ x 1), 6.14 (1H´, s, 3-H_a´ x 1), 6.16 (1H, s, 3-H_a x 1), 6.25 (1H, s, 3-H_b x 1), 6.29 (1H´, s, 3-H_b´ x 1), 7.24-7.38 (5H+5H´, m, aryl-H x 5); **δ_C (50 MHz, $CDCl_3$):** 198.6 (C=O), 198.5,

CML Infra-Red ASSIGNMENTS

```
<cml:spectrum type="cml:ir">
 <cml:conditionList>
  <cml:condition title="the form of the IR spectrum“ dictRef="cml:irform">film</cml:condition>
  </cml:conditionList>
 <cml:peakList>
  <cml:peak id="p1" xValue="3446" title="OH" />
  <cml:peak id="p2" xValue="3062" title="unassigned" />
  <cml:peak id="p3" xValue="3029" title="unassigned" />
  <cml:peak id="p4" xValue="2922" title="unassigned" />
  <cml:peak id="p5" xVa
  <cml:peak id="p6" xVa
  <cml:peak id="p7" xVa
  <cml:peak id="p8" xVa
  <cml:peak id="p9" xVa
  <cml:peak id="p10" xV
  <cml:peak id="p11" xV
  <cml:peak id="p12" xV
  <cml:peak id="p13" xV
   </cml:peakList>
  </cml:spectrum>
```

CML C-13 NMR ASSIGNMENTS

```
<cml:spectrum type="cml:cnmr">
 <cml:parameterList>
  <cml:parameter dictRef="cml:frequency" units="units:MHz">50</cml:parameter>
  </cml:parameterList>
 <cml:substanceList>
  <cml:substance ref="" />
  </cml:substanceList>
 <cml:peakList>
  <cml:peak xValue="198.6" integral="" peakMultiplicity="" title="C=O" />
  <cml:peak xValue="198.5" integral="" peakMultiplicity="" title="" />
  <cml:peak xValue="145.0" integral="" peakMultiplicity="" title="C" />
  <cml:peak xValue="142.7" integral="" peakMultiplicity="" title="C" />
  <cml:peak xValue="137.3" integral="" peakMultiplicity="" title="CH2" />
  <cml:peak xValue="136.7" integral="" peakMultiplicity="" title="CH2" />
  <cml:peak xValue="129.1" integral="" peakMultiplicity="" title="" />
  <cml:peak xValue="128.6" integral="" peakMultiplicity="" title="" />
  <cml:peak xValue="126.7" integral="" peakMultiplicity="" title="" />
  <cml:peak xValue="124.0" integral="" peakMultiplicity="" title="aryl-C" />
  <cml:peak xValue="62.5" integral="" peakMultiplicity="" title="CH" />
  <cml:peak xValue="59.0" integral="" peakMultiplicity="" title="CH" />
  <cml:peak xValue="55.2" integral="" peakMultiplicity="" title="CH" />
  <cml:peak xValue="54.9" integral="" peakMultiplicity="" title="CH" />
  <cml:peak xValue="38.5" integral="" peakMultiplicity="" title="CH2" />
  <cml:peak xValue="32.8" integral="" peakMultiplicity="" title="CH2" />
  <cml:peak xValue="26.1" integral="" peakMultiplicity="" title="CH3" />
  <cml:peak xValue="26.0" integral="" peakMultiplicity="" title="CH3" />
  </cml:peakList>
  </cml:spectrum>
```

RDF - Resource Description Framework.

A component of the Semantic Web, it is based upon the idea of making statements about resources/data in the form of a

subject-predicate-object (*or* **resource-property-value**)

expression (called a ***triple***) *e.g.* :

My_thesis **has_chemical_entity** **2,4-dinitrobenzene**

The value of one property can in turn be used as the resource for another.

SPARQL QUERY

```
PREFIX st: <http://wwmm.ch.cam.ac.uk/spectra-t#>
PREFIX dcrdf: <http://purl.org/metadata/dublin_core#>
CONSTRUCT { ?thesis st:hasBicycloMoleculeAndHNMR ?chemical .
?thesis dcrdf:author ?author
}
WHERE { ?thesis dcrdf:creator ?author .
 ?thesis st:hasChemicalName ?annot .
 ?annot st:chemicalName ?chemical .
 ?annot st:hasHNMRSpectrum ?hnmr .
FILTER regex(?chemical, ".*bicyclo.*") .
 }
```

RDF TRIPLESTORE ENTRY

RESULT

```
<rdf:Description rdf:about="file:/C:/spectra-t-articles/B207708F.docx">
        <st:hasBicycloMoleculeAndHNMR>5-Acetyl-7,8-bis(trimethylsilyl)bicyclo[4.2.1]nona-4,7-diene</st:hasBicycloMoleculeAndHNMR>
        <dcrdf:author>N.R.Champness</dcrdf:author>
        <st:hasBicycloMoleculeAndHNMR>5-Acetyl-bicyclo[4.2.1]nona-4,7-diene</st:hasBicycloMoleculeAndHNMR>
        <dcrdf:author>N.R.Champness</dcrdf:author>
        <st:hasBicycloMoleculeAndHNMR>5-Phenyl-bicyclo[4.2.1]nona-3,7-diene</st:hasBicycloMoleculeAndHNMR>
        <dcrdf:author>N.R.Champness</dcrdf:author>
        <st:hasBicycloMoleculeAndHNMR>5-Acetyl-7,8-bis(trimethylsilyl)bicyclo[4.2.1]nona-4,7-diene</st:hasBicycloMoleculeAndHNMR>
        <dcrdf:author>N.R.Champness</dcrdf:author>
        <st:hasBicycloMoleculeAndHNMR>5-Acetyl-bicyclo[4.2.1]nona-4,7-diene</st:hasBicycloMoleculeAndHNMR>
        <dcrdf:author>N.R.Champness</dcrdf:author>
        <st:hasBicycloMoleculeAndHNMR>5-Phenyl-bicyclo[4.2.1]nona-3,7-diene</st:hasBicycloMoleculeAndHNMR>
        <dcrdf:author>N.R.Champness</dcrdf:author>
</rdf:Description>
```

```
  <spectra-t:chemicalName>(3E,5S,6S)-8-(p-Methoxy-benzyloxy)-5,6-epoxy-6-methyl-oct-3-en-2-one</spectra-t:chemicalName>
  <spectra-t:hasHNMRSpectrum>http://fiwlt.ch.cam.ac.uk:8182/8f2d98b04/data-20.cml</spectra-t:hasHNMRSpectrum>
  <spectra-t:hasIRSpectrum>http://fiwlt.ch.cam.ac.uk:8182/8f2d98b04/data-20.cml</spectra-t:hasIRSpectrum>
  <spectra-t:hasMassSpectrum>http://fiwlt.ch.cam.ac.uk:8182/8f2d98b04/data-20.cml</spectra-t:hasMassSpectrum>
  <spectra-t:hasHRMSSpectrum>http://fiwlt.ch.cam.ac.uk:8182/8f2d98b04/data-20.cml</spectra-t:hasHRMSSpectrum>
  <spectra-t:hasPreparation>http://fiwlt.ch.cam.ac.uk:8182/8f2d98b04/preparation-20.sci.xml</spectra-t:hasPreparation>
  </rdf:Description>
 </spectra-t:hasChemicalName>

 </rdf:Description>
 <rdf:RDF>
```

Message to repository managers:

PDF is a limited format for data extraction from e-theses

Docx allows chemical data object extraction (~80% precision / recall)

Caveats (Proof-of-concept):

Single subject area (synthetic organic chemistry)

Single institution docx (limited variation in document structure)

Limited thesis availability

Solutions :

Domain ontology development

Make your e-theses public!

Acknowledgements

- Project Director: *Peter Morgan UL Cambridge*
- Chemistry leads: *Henry Rzepa, Peter Murray-Rust*
- Developers: *Jim Downing, Diana Stewart, Joe Townsend, Matt Harvey*
- Project Manager: *Alan Tonge*

http://www.lib.cam.ac.uk/spectra-t/

JISC

UNIVERSITY OF CAMBRIDGE

Imperial College London

SPECTRa Tools Workshop

Autumn 2008

Unilever Centre, Cambridge, UK

Contact: ***Peter Murray-Rust (pm286@cam.ac.uk)***

Peter Morgan (pbm2@cam.ac.uk)