Unlocking chemical information from tables and legacy articles

Daniel Lowe and Roger Sayle
NextMove Software

Aileen Day and Antony Williams
Royal Society of Chemistry
Topics

• Chemical property extraction

• Application of chemical property extraction to tables

• RSC back-archive mining
Chemical property extraction

- Melting points
- Boiling points
- Mass spectrum
- Textual NMR spectra
- Specific rotation
- Chromatography retention times
- IR/UV spectra
- Activity data e.g. IC$_{50}$, EC$_{50}$, Ki
- Etc.
Simple grammar and corresponding state machine

Isotope: ‘1H’ | ‘13C’ | ‘19F’
Nmr: ‘-NMR’
NmrPrelog: Isotope Nmr
Melting point recognition

M.p.: 230°C (dec.)

<table>
<thead>
<tr>
<th>Term</th>
<th>Examples of text matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>FromLiterature</td>
<td>“lit.”</td>
</tr>
<tr>
<td>MeltingPoint</td>
<td>“mpt”, “melting point”, “m.p.”</td>
</tr>
<tr>
<td>Qualifier</td>
<td>“&gt;”; “approximately”</td>
</tr>
<tr>
<td>Value</td>
<td>“75° C”, “200° F”, “one hundred degrees Celsius”</td>
</tr>
<tr>
<td>Range</td>
<td>“184-186° C”, “191.5 to 192.4° C”</td>
</tr>
<tr>
<td>MeasurementError</td>
<td>“50±° C”</td>
</tr>
<tr>
<td>OutcomeQualifier</td>
<td>“decomp.”, “with decomposition”, “subl.”</td>
</tr>
</tbody>
</table>
NMR recognition

Isotope NMR NmrMethod? Peak PeakAnnotation? (Delimiter Peak PeakAnnotation?)*

\[ ^1H \text{ NMR (300 MHz, DMSO): 7.5-7.8 (m, 5H), 7.9 (d, J=8Hz, 2H), 8.33 (d, J=5Hz, 2H)} \]

<table>
<thead>
<tr>
<th>Term</th>
<th>Examples of text matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isotope</td>
<td>“1H”, “13C”, “19F”</td>
</tr>
<tr>
<td>NMR</td>
<td>“NMR”, “RMN”</td>
</tr>
<tr>
<td>NmrMethod</td>
<td>“400 MHz, CDCl3”</td>
</tr>
<tr>
<td>Peak</td>
<td>“3.7”</td>
</tr>
<tr>
<td>PeakAnnotation</td>
<td>“s, 3H”</td>
</tr>
</tbody>
</table>
Recognition and parsing

• Grammar distinguishes parts of an entity of interest e.g. 25°C → 25 (value) °C (unit)

• Can groups constructs together e.g. 25 to 30 (range)
Example parse Tree serialised to XML

Mp: 131.9-132.6 °C

<parse>
  <quantityType quantityType="MeltingPoint">Mp</quantityType>
  <measurement>
    <range>
      <valueOptUnit>
        <decimalValue>131.9</decimalValue>
      </valueOptUnit>
      <rangeDelimiter>-</rangeDelimiter>
      <valueOptUnit>
        <decimalValue>132.6</decimalValue>
      </valueOptUnit>
      <unitContainer>
        <unit unitType="Temperature" normalizationFactor="1">°C</unit>
      </unitContainer>
    </range>
  </measurement>
</parse>
Recognition and parsing

• Grammar distinguishes parts of an entity of interest e.g. 25°C → 25 (value) °C (unit)

• Can groups constructs together e.g. 25 to 30 (range)

• However this introduces non-determinism e.g. after seeing “25” both the possibility of being in and not being in a range, need to be considered
• Same grammar can be used to generate:
  – Single state machine representation
    • Parts of entity not distinguished
    • Extremely fast recognition
    • Allows spelling correction of input that is close to being a match

  – Multi state machine parser representation
    • Slower... but only needs to be run on a small amount of text
    • Distinguishes parts of entity
    • Can group parts into a parse tree
Grammar implementation details

<regex name="\$Number\$" regex="(\[1-9]\[0-9\]*|0)\((0-9)+\)?"/>
<regexToken name="PeakValue" tagName="peakValue" regex="[\+\-]?(\$Number\$\([0-9]+\))"/>

<tokenList name="IsotopicallyLabelledElement" tagName="nmrElement">
  <token isootope="1" element="H">1H</token>
  <token isootope="1" element="H">1-H</token>
  <token isootope="1" element="H">1 H</token>
  <token isootope="2" element="H">2H</token>
  <token isootope="2" element="H">2-H</token>
  <token isootope="2" element="H">2 H</token>
  <token isootope="3" element="H">3H</token>
  <token isootope="3" element="H">3-H</token>
  <token isootope="3" element="H">3 H</token>
  ...
</tokenList>

<rule name="PeakValueRange" nestingTagName="peakRange" regex="About? (PeakValue Space* (Hyphen|PeakRangeIndicator)"
Table Extraction
# Melting point table

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>M.p. [° C.]</th>
<th>MS-FAB [M + H]+</th>
</tr>
</thead>
<tbody>
<tr>
<td>“A2”</td>
<td>3-(3-Hydroxybenzylamino)-4-(4-hydroxy-3-phenylphenylamino)cyclobut-3-ene-1,2-dione</td>
<td>246-247</td>
<td></td>
</tr>
<tr>
<td>“A3”</td>
<td>3-(4-Hydroxy-3-pyridin-2-ylphenylamino)-4-[(R)-1-phenylethylamino]cyclobut-3-ene-1,2-dione</td>
<td>298-290</td>
<td></td>
</tr>
<tr>
<td>“A4”</td>
<td>3-(4-Hydroxy-3-pyridin-2-ylphenylamino)-4-[(R)-1-(3-methoxyphenyl)ethylamino]cyclobut-3-ene-1,2-dione</td>
<td>263-264</td>
<td></td>
</tr>
<tr>
<td>“A5”</td>
<td>3-(4-Hydroxy-3-pyridin-2-ylphenylamino)-4-[(R)-1-(3-hydroxyphenyl)ethylamino]cyclobut-3-ene-1,2-dione</td>
<td>284-285</td>
<td></td>
</tr>
<tr>
<td>“A6”</td>
<td>3-(3-Pyridin-2-ylphenylamino)-4-[(R)-1-(3-hydroxyphenyl)ethylamino]cyclobut-3-ene-1,2-dione</td>
<td>264-265</td>
<td></td>
</tr>
</tbody>
</table>

250th ACS National Meeting, Boston MA, USA 17th August 2015
# NMR table

<table>
<thead>
<tr>
<th>Ex. n°</th>
<th>Structure</th>
<th>Name</th>
<th>(^1)H-NMR δ ppm</th>
<th>m.p. °C.</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td><img src="image" alt="Structure" /></td>
<td>N-(1-(3-chlorophenyl)-1H-pyrazol-3-yl)-2-morpholino acetamide oxalate</td>
<td>DMSO-d_6: 10.6 (bs, 1H), 8.5 (d, J = 2.6 Hz, 1H), 7.85 (t, J = 1.9 Hz, 1H), 7.75 (dd, J = 1.8, 8.0 Hz, 1H), 7.5 (t, J = 8.0 Hz, 1H), 7.3 (dd, J = 1.7, 7.9 Hz, 1H), 6.8 (d, J = 2.6 Hz, 1H), 3.6 (t, J = 4.6 Hz, 4H), 3.35 (bs, 2H), 2.65 (m, 4H).</td>
<td>219-220</td>
<td>320</td>
</tr>
</tbody>
</table>
More difficult…

<table>
<thead>
<tr>
<th>Example Number</th>
<th>IC50 (nM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>14</td>
<td>59</td>
</tr>
<tr>
<td>49</td>
<td>710</td>
</tr>
<tr>
<td>52</td>
<td>320</td>
</tr>
<tr>
<td>53</td>
<td>93</td>
</tr>
</tbody>
</table>

Need to be looked up else where in document. Could be in text, might be in images.

Against what?
Even More difficult...

![Chemical structure](image)

<table>
<thead>
<tr>
<th>Preparations</th>
<th>R\textsubscript{6}, R\textsubscript{7}, R\textsubscript{8}</th>
<th>Intermediate (VII) or (VIII)</th>
<th>Characterization</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3</td>
<td>2,4-diCl</td>
<td>(VII): Alk = Et</td>
<td>NMR CDCl\textsubscript{3} (300 MHz): 1.25 ppm: t: 3H; 4.25 ppm: q: 2H; 7.13–7.44 ppm: m: 4H; 12.03 ppm: m: 1H. m.p. = 73° C.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(VIII): Alk = Me</td>
<td></td>
</tr>
</tbody>
</table>

250\textsuperscript{th} ACS National Meeting, Boston MA, USA 17\textsuperscript{th} August 2015
### TABLE 3

<table>
<thead>
<tr>
<th>Entry Name</th>
<th>$^1$H-NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  $N^1$-hydroxy-$N^2$-{[4-(phenyloxy)phenyl]sulfonyl}-D-lysinamide</td>
<td>H-NMR; $\delta$ (CD3OD): 7.79 (d, 2H), 7.42 (t, 2H), 7.22 (t, 1H), 7.09 (d, 2H), 7.05 (d, 2H), 3.63 (t, 1H), 2.87 (t, 2H), 1.57-1.68 (m, 4H), 1.44 (m, 1H), 1.37 (m, 1H)</td>
</tr>
<tr>
<td>2  $N^1$-hydroxy-$N^2$-{[4-(phenyloxy)phenyl]sulfonyl}-3-piperidin-3-ylalaninamide</td>
<td>H-NMR; $\delta$ (CD3OD): 7.79 (q, 2H), 7.42 (t, 2H), 7.22 (t, 1H), 7.10 (d, 2H), 7.05 (d, 2H), 3.72 (m, 1H), 3.23-3.47 (m, 2H), 2.87 (m, 1H), 2.64 (t, 1H), 1.83-2.01 (m, 3H), 1.17-1.74 (m, 4H)</td>
</tr>
</tbody>
</table>
...and the xml provided

<row>
<entry>1</entry>
<entry>N<sup>1</sup>-hydroxy-N<sup>2</sup>-{[4-(phenyloxy)phenyl]sulfonyl}-</entry>
<entry>H-NMR; &#x3b4; (CD3OD): 7.79 (d, 2H),</entry>
</row>
<row>
<entry/> <entry>D-lysinamide</entry>
<entry>7.42 (t, 2H), 7.22 (t, 1H), 7.09 (d, 2H),</entry>
</row>
<row>
<entry/> <entry/> <entry>7.05 (d, 2H), 3.63 (t, 1H), 2.87 (t, 2H),</entry>
</row>
<row>
<entry/> <entry/> <entry/> <entry>1.57-1.68 (m, 4H), 1.44 (m, 1H),</entry>
</row>
<row>
<entry/> <entry/> <entry/> <entry>1.37 (m, 1H)</entry>
</row>
Naïve interpretation (Google patents)

TABLE 3

<table>
<thead>
<tr>
<th>Entry Name</th>
<th>$^1$H-NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. N₁-hydroxy-N₂-[[4-(phenyloxy)phenyl]sulfonyl]-D-lysinamide</td>
<td>H-NMR; $\delta$ (CD3OD): 7.79 (d, 2H), 7.42 (t, 2H), 7.22 (t, 1H), 7.09 (d, 2H), 7.05 (d, 2H), 3.63 (t, 1H), 2.87 (t, 2H), 1.57-1.68 (m, 4H), 1.44 (m, 1H), 1.37 (m, 1H)</td>
</tr>
<tr>
<td>2. N₁-hydroxy-N₂-[[4-(phenyloxy)phenyl]sulfonyl]-3-piperidin-3-ylalaninamide</td>
<td>H-NMR; $\delta$ (CD3OD): 7.79 (q, 2H), 7.42 (t, 2H), 7.22 (t, 1H), 7.10 (d, 2H), 7.05 (d, 2H), 3.72 (m, 1H), 3.23-3.47 (m, 2H), 2.87 (m, 1H), 2.64 (t, 1H), 1.83-2.01 (m, 3H), 1.17-1.74 (m, 4H)</td>
</tr>
</tbody>
</table>

Green: chemical substituent  
Purple: chemical molecule  
Blue: NMR
<table>
<thead>
<tr>
<th>Entry</th>
<th>Name</th>
<th>1H-NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>N1-hydroxy-N2-[[4-(phenyloxy)phenyl]sulfonyl]- H-NMR; δ (CD3OD): 7.79 (d, 2H), D-lysineamide</td>
<td>7.42 (t, 2H), 7.22 (t, 1H), 7.09 (d, 2H),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.05 (d, 2H), 3.63 (t, 1H), 2.87 (t, 2H),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.57-1.68 (m, 4H), 1.44 (m, 1H),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.37 (m, 1H)</td>
</tr>
<tr>
<td>2</td>
<td>N1-hydroxy-N2-[[4-(phenyloxy)phenyl]sulfonyl]- H-NMR; δ (CD3OD): 7.79 (q, 2H), 3-piperidin-3-ylalaninamide</td>
<td>7.42 (t, 2H), 7.22 (t, 1H), 7.10 (d, 2H),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.05 (d, 2H), 3.72 (m, 1H),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.23-3.47 (m, 2H), 2.87 (m, 1H), 2.64 (t,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1H), 1.83-2.01 (m, 3H), 1.17-1.74 (m,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4H)</td>
</tr>
</tbody>
</table>
After heuristically detecting which rows are the same row

**TABLE 3**

<table>
<thead>
<tr>
<th>Entry</th>
<th>Name</th>
<th>$^1$H-NMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$N^1$-hydroxy-$N^2$-{[4-(phenyloxy)phenyl]sulfonyl}-D-lysinamide</td>
<td>$\delta$ (CD3OD): 7.79 (d, 2H), 7.42 (t, 2H), 7.22 (t, 1H), 7.09 (d, 2H), 7.05 (d, 2H), 3.63 (t, 1H), 2.87 (t, 2H), 1.57-1.68 (m, 4H), 1.44 (m, 1H), 1.37 (m, 1H)</td>
</tr>
<tr>
<td>2</td>
<td>$N^1$-hydroxy-$N^2$-{[4-(phenyloxy)phenyl]sulfonyl}-3-piperidin-3-ylalaninamide</td>
<td>$\delta$ (CD3OD): 7.79 (q, 2H), 7.42 (t, 2H), 7.22 (t, 1H), 7.10 (d, 2H), 7.05 (d, 2H), 3.72 (m, 1H), 3.23-3.47 (m, 2H), 2.87 (m, 1H), 2.64 (t, 1H), 1.83-2.01 (m, 3H), 1.17-1.74 (m, 4H)</td>
</tr>
</tbody>
</table>

Purple: chemical molecule  
Blue: NMR
What could be extracted?

Name/identifier to property relationships

ISO, TSO, LigandBinding, MolecularWeight, MeltingPoint, Yield, Pic50, pH, RetardationFactor, BoilingPoint, Solubility, Tg, LogP, EnantiomericExcess, FlashPoint, SpecificRotation, Jka, KinematicViscosity, DecompositionPoint, LogD, CloudPoint, DDiasteromericExcess, ZetaPotential, CrudeYield, LogD, SublimationPoint.
Compound number determination

EXAMPLE-14

2-(2,4-difluorophenoxy)-5...

3. (4aS,8aR)-2-(1-Acetyl-pipe...

2-Chloro-5-iodo-1H-benzo[d]imidazole (1)
RSC-back archive mining
RSC back archive

• 1841-1999, 211k articles (available as XML derived from OCR and PDF)
• 2000 -, 230k articles (available as born digital XML and PDF)

• Also over 150k Electronic supporting information files (mostly PDF, but also Word docs, Excel files, videos etc.)
Legacy document handling

• Chemical properties are often implicitly associated with a compound by being in the same experimental section
Step 4: **Bis-(4-hydroxybenzoyl)-diethylgermanium**

![Chemical Structure](image)

Bis-(4-methoxybenzoyl)diethylgermanium (8.0 g, 20.0 mmol) was dissolved in anhydrous toluene (200 ml) under protective gas and had Celatom (10 g) and aluminium chloride (9.6 g, 72.0 mmol) added to it. The reaction mixture was heated for 2 h under reflux. After cooling, water (10 ml) was added and the suspension was stirred for 10 min at room temperature. The solvent was removed on the rotary evaporator. The residue had ethyl acetate (300 ml) added to it. The suspension was stirred for 16 h at room temperature and filtered over a thin layer of silica gel. The filtrate was concentrated on the rotary evaporator. The oily brown residue had chloroform (200 ml) added to it. The suspension was stirred for 16 h at room temperature and filtered. The filtration residue was washed with chloroform (80 ml) and dried in the vacuum drying oven (50°C, 125 mbar). 4.23 g (11.3 mmol, 57% yield) of a light yellow solid was obtained (mp: 167-168°C).

**1H-NMR (DMSO-d₆, 400 MHz):** δ=1.04 (t, 6H, J=7.9 Hz, —CH₃), 1.38 (q, 4H, J=7.9 Hz, Ge—CH₂—), 6.88 (d, 4H, J=8.5 Hz, Ar—H³), 7.58 (d, 4H, J=8.5 Hz, Ar—H²), 10.53 (s, 2H, OH).

**13C-NMR (DMSO-d₆, 100.6 MHz):** δ=61 (—CH₃), 8.9 (Ge—CH₂—), 115.7 (Ar—C³), 130.3 (Ar—C²), 133.3 (Ar—C¹), 162.7 (Ar—C⁴), 225.5 (C=O).
Legacy document handling

• Chemical properties are often implicitly associated with a compound by being in the same experimental section

• This requires section detection e.g. a heading and/or a paragraph where a compound is being synthesised

• In the XML for pre-2000 papers all sections on a page run together (including page numbers!), and the text position information is lost.

• ...so back to the source PDF
Heading/Paragraph detection workflow
## Results (Melting points)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Compound-value associations</strong></td>
<td>2,155</td>
<td>29,996</td>
<td>172,886</td>
</tr>
<tr>
<td><strong>Suspicious Values</strong> (typically mistake in the document)</td>
<td>70 (3.2%)</td>
<td>39 (0.13%)</td>
<td>426 (0.25%)</td>
</tr>
<tr>
<td><strong>Unique Compounds</strong> (StdInChI)</td>
<td>1,830 (84.9%)</td>
<td>27,956 (93.2%)</td>
<td>95,140 (55.0%)</td>
</tr>
</tbody>
</table>
# SDF output

<table>
<thead>
<tr>
<th>#</th>
<th>Structure</th>
<th>SMILES</th>
<th>Experimental Point</th>
<th>Value</th>
<th>PubChem CID</th>
<th>WebPage</th>
<th>ChemistryStructureName</th>
<th>CAS Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>25371</td>
<td><img src="image1.png" alt="Structure Image" /></td>
<td>OC1=C(O)=C(=Nc1ccc(c2ccc(n3c(cn4c(=O)bc5ccc6cc5)c(=O)c6)c6c7c8c9cc1)c7c8c9c10c10)</td>
<td>mp 235.0±0.2°C</td>
<td>255 to 265°C</td>
<td>6453289</td>
<td><a href="http://dx.doi.org/10.1039">PDF</a></td>
<td>10.11-Dihydro-15a-bromo-15a-hydroxy-3a,13a-diazadynen-6(1H)-ene</td>
<td>N/A</td>
</tr>
<tr>
<td>25372</td>
<td><img src="image2.png" alt="Structure Image" /></td>
<td>C1=CC(C(=O)C(=O)c1ccc(c2ccc(n3c(cn4c(=O)bc5ccc6cc5)c(=O)c6)c6c7c8c9cc1)c7c8c9c10c10)</td>
<td>mp 305.000°C</td>
<td>305 to 306°C</td>
<td>6453288</td>
<td><a href="http://dx.doi.org/10.1039">PDF</a></td>
<td>10.11-Dihydro-2,3-dioxo-13a-chlorobenzoxypiperiderin-6(1H)-ene</td>
<td>N/A</td>
</tr>
</tbody>
</table>
2,2-Difluoro-4-methyl-5-(o-tolyl)-5,10-dihydro-2H-[1,3,2]dioxaborinino[4,5-b]quinolin-1-ium-2-uide (2j). Yellow solid: mp 251–253 °C; $^1$H NMR (300 MHz, DMSO): $\delta$ 1.92 (s, 3H), 2.34 (s, 3H), 5.52 (s, 1H), 7.03–7.08 (m, 3H), 7.12–7.18 (m, 3H), 7.20–7.26 (m, 2H), 12.08 (s, 1H); $^{13}$C NMR (100 MHz, DMSO): $\delta$ 19.0, 20.8, 96.5, 117.0, 125.5, 125.8, 126.6, 126.9, 127.8, 129.3, 131.4, 132.3, 134.2, 144.4, 163.6, 179.0; IR (KBr, cm$^{-1}$): 3354, 1622, 1591, 1521, 1493, 1047, 764; Anal. calcd for C$_{18}$H$_{16}$BF$_2$NO$_2$: C, 66.09; H, 4.93; N, 4.28. Found: C, 66.41; H, 4.89; N, 4.32.
1,3-Dichloro-2,2,4,4-tetrafluoro-1λ^4,3λ^4-ditellurabicyclo[1.1.0]-butane (12)

In a Carius tube (50 ml) equipped with a Teflon-stemmed Young valve and a magnetic stirring bar 2,2,4,4-tetrafluoro-1,3-ditelluretane (43.4 mg, 0.122 mmol) was deposited and dissolved in CH₂Cl₂ (1 cm³). The dark blue solution was cooled to −196 °C and evacuated. Thereafter Cl₂ (26 mg, 0.733 mmol) was condensed in. The mixture was warmed to −80 °C and stirred for 1 h. Solvent and excess Cl₂ were removed at −50 °C \textit{in vacuo} providing temperature sensitive red-orange 12 (51.0 mg, 98%) soluble in CH₃CN. Decomposition temperature −15 °C. See SUP data for IR, ^13C, ^19F, ^125Te NMR and MS data for 12.
## Results (NMR)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound-value associations</td>
<td>4,972</td>
<td>94,610</td>
<td>1,295,325</td>
</tr>
<tr>
<td>Suspicious Values (typically mistake in the document)</td>
<td>561 (11.3%)</td>
<td>2,001 (2.11%)</td>
<td>29,775 (2.30%)</td>
</tr>
<tr>
<td>Unique Compounds (StdInChI)</td>
<td>2,899</td>
<td>48,137</td>
<td>655,295</td>
</tr>
</tbody>
</table>
Legacy text issues

• OCR errors in important compound names or data
  - chemical names in italics problematic... key compounds often in italics!
  - ° is more often than not misinterpreted e.g. ' o

• Tools prefer experimental sections where one compound is being synthesised, qualitatively older documents are less formalised
4-Chloro-6-hydroxy-2-methylaminopyrimidine.—4-Chloro-6-methoxy-2-methylaminopyrimidine (10 g.) was heated on the steam-bath for 30 min. with concentrated hydrochloric acid (50 c.c.). The hydroxy-compound which separated on cooling was collected and purified by dissoluction in alkali, etc., as above, and had m. p. 265° (decomp.) (5.5 g.) (Found: C, 38.3; H, 4.1; N, 26.2. C₆H₆ON₃Cl requires C, 37.6; H, 3.8; N, 26.3%).

4-ChZoro-6-hydroxy-2-methylZamino~yrimidine.-4-Chloro-6-methoxy-2-methylaminopyrim-idine (10g.) was heated on the steam-bath for 30 min. with concentrated hydrochloric acid (60 c.c.). The hydvoxy- cmfiound which separated on cooling was collected and purified by dis- solution in alkali,etc. as above and had m. p. 266" (decomp.) (6.6 g.) (Found c 38.3 ; H 4.1; N 26-2. C,H,0N3C1 requires C 37.6; H 3.8; N 26.3%).
Conclusions

• Grammars facilitate rapid extraction and interpretation of chemical properties

• Table extraction is vital to extracting large quantities of certain data e.g. activity data

• Large amounts of high quality data can be extracted from journal articles

• ...but extraction from older documents remains very challenging, and over time represents a smaller and smaller percentage of the scientific literature
Acknowledgements

• Igor Tetko (Melting point quality feedback)
• Carlos Cobas (NMR quality feedback)

Funding provided by:
Chemistry Enabling Chinese, Japanese and Korean Patents

6-aminopyrimidine-2,4,5-triol

Chinese (Hanzi used for each morpheme)

6-氨基嘧啶-2,4,5-三醇

ammonia radical pyrimidine three alcohol

Japanese (Phonetic translation to Katakana)

6-アミノピリミジン-2,4,5-トリオール

amino pyrimidine tri ol

Korean (Phonetic translation to Hangul)

6-아미노피리미딘-2,4,5-트리올

amino pyrimidine tri ol

Sci-Mix
8:00pm – 10:00pm Today
Hall C – Boston Convention & Exhibition Center
Thank you for your time!

http://nextmovesoftware.com
http://nextmovesoftware.com/blog
daniel@nextmovesoftware.com