

The **SPECTRa** Project :

*Generating & Depositing Chemistry Research
Data*

Alan Tonge

University of Cambridge

*Digital Repositories :
Dealing with the Data Deluge*

*Manchester University
5 June 2007*



Project Overview

- 18-month project between University of Cambridge and Imperial College London to develop customized tools to deposit chemistry data in digital repositories
- Part of the **JISC** Digital Repositories programme
- Closely integrated with eBank and eCrystals (Bath and Soton)

The Problem

**Experimental chemistry data is a resource / asset ...
almost always omitted from traditional publishing**

- PDF image files (supplementary data) : *not machine readable*
- Proprietary spectra formats (NMR, IR, UV) : *~5-year shelf life*
- CIF xray : *80% remain unpublished*

**Cambridge / Imperial: 100,000+ NMR Spectra / year
300 xray
...much of which will become lost or unreadable**

Most of the problems are social, not technical

Requirements & use determined by survey

Chemistry is multi-disciplinary : experimental & theoretical studies on small macromolecular and polymeric structures. Requirements in selected user disciplines

- ***synthetic organic chemistry***
- ***departmental crystallography services***
- ***computational chemistry***

Determined by general voluntary questionnaire of all researchers.

Specific needs identified by one-to-one interviews

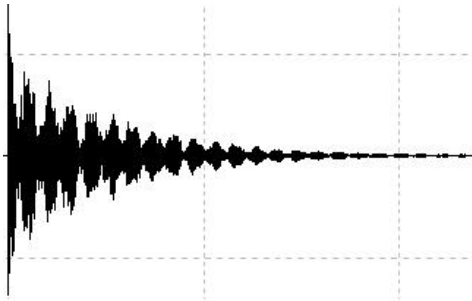


Survey Results

The main conclusions were:

- *A complex list of data file formats (particularly proprietary binary formats) being used*
- *Much data is not stored electronically (e.g. lab books, paper copies of spectra)*
- *A significant ignorance of digital repositories*
- *A requirement for **restricted access** to deposited experimental data*

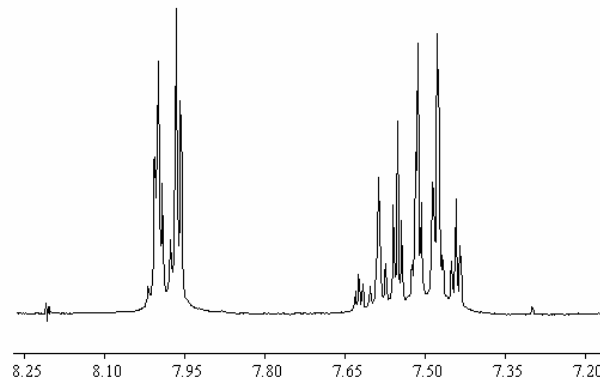
Selective NMR Data Capture



Raw binary data

```
##XYDATA= (X++(Y..Y))
16383 -105239 -129156 -22382 -207 80957 77779 48208
16376 55139 38551 89509 37860 18919 106418 150251
16369 37279 -13052 4733 -37056 -982 -43426 26769
16362 104325 162266 40340 127499 152950 129357 40124
16355 16159 -65067 -58307 -96110 -73637 -48226 -110550
16348 4580 90453 66464 5427 12699 -1258 -38892
16341 -81147 -12906 30188 106241 68684 1701 46013
16334 9932 2981 16045 50599 79390 177277 73054
16327 31941 -137585 -27444 29287 87577 78078 87685
16320 59198 52303 111211 28899 -7161 31638 33837
16313 -1731 87344 193683 178168 130570 48655 42330
16306 16517 46196 100796 30545 53651 37331 144520
16299 71291 -37141 -24169 -89071 -180654 -201961 -45301
16292 -23441 -13387 -5293 -8660 -31767 -110985 7173
16285 39252 8528 -27980 69996 7775 39407 31368
16278 -54159 -44298 -58669 57208 85859 4135 33141
```

Transformed non-binary



Displayed Image

Non-binary formats which are **accepted data standards** within the various chemistry disciplines :

- **Crystallography: CIF files**
- **NMR: JCAMP-DX and MDL molfiles**
- **Computational Chemistry: Gaussian Archive files**

Chemical Markup Language (CML) can provide machine-based validation of marked-up chemistry data through the use of XSD schemas. All four file types identified above converted to the appropriate CML subtype and validated before deposition.

- **'Low hanging Fruit'**
- **No raw experimental data (e.g. x-ray diffraction patterns, nmr FID's)**

Conversion of MDL molfile structure format to CML

Data validation with XSD Schemas (data type, data range)

```
-ISIS- 09080615252D
9 9 0 0 0 0 0 0 0 0 0999 V2000
-0.3806 -0.7208 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-0.3818 -1.5482 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.3331 -1.9610 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.0495 -1.5477 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.0466 -0.7172 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0.3313 -0.3080 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.7596 -0.3020 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2.4756 -0.7118 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1.7564 0.5230 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4 5 1 0 0 0 0
2 3 1 0 0 0 0
5 6 2 0 0 0 0
6 1 1 0 0 0 0
1 2 2 0 0 0 0
5 7 1 0 0 0 0
3 4 2 0 0 0 0
7 8 1 0 0 0 0
7 9 2 0 0 0 0
M END
```

```
<molecule xmlns="http://www.xml-cml.org/schema/cml2/core">
  <atomArray>
    <atom id="a1" elementType="C" x2="-0.380600" y2="-0.720800" />
    <atom id="a2" elementType="C" x2="-0.381800" y2="-1.548200" />
    <atom id="a3" elementType="C" x2="0.333100" y2="-1.961000" />
    <atom id="a4" elementType="C" x2="1.049500" y2="-1.547700" />
    <atom id="a5" elementType="C" x2="1.046600" y2="-0.717200" />
    <atom id="a6" elementType="C" x2="0.331300" y2="-0.308000" />
    <atom id="a7" elementType="C" x2="1.759600" y2="-0.302000" />
    <atom id="a8" elementType="C" x2="2.475600" y2="-0.711800" />
    <atom id="a9" elementType="O" x2="1.756400" y2="0.523000" />
  </atomArray>
  <bondArray>
    <bond atomRefs2="a4 a5" order="1" />
    <bond atomRefs2="a2 a3" order="1" />
    <bond atomRefs2="a5 a6" order="2" />
    <bond atomRefs2="a6 a1" order="1" />
    <bond atomRefs2="a1 a2" order="2" />
    <bond atomRefs2="a5 a7" order="1" />
    <bond atomRefs2="a3 a4" order="2" />
    <bond atomRefs2="a7 a8" order="1" />
    <bond atomRefs2="a7 a9" order="2" />
  </bondArray>
</molecule>
```

The Solution

Capture selected data from chemistry workflows in open format
(JCAMP, MOL, CIF)



Add context-specific and embargo metadata +

Persistent identifiers



Deposit as METS package in DSpace Digital Repository



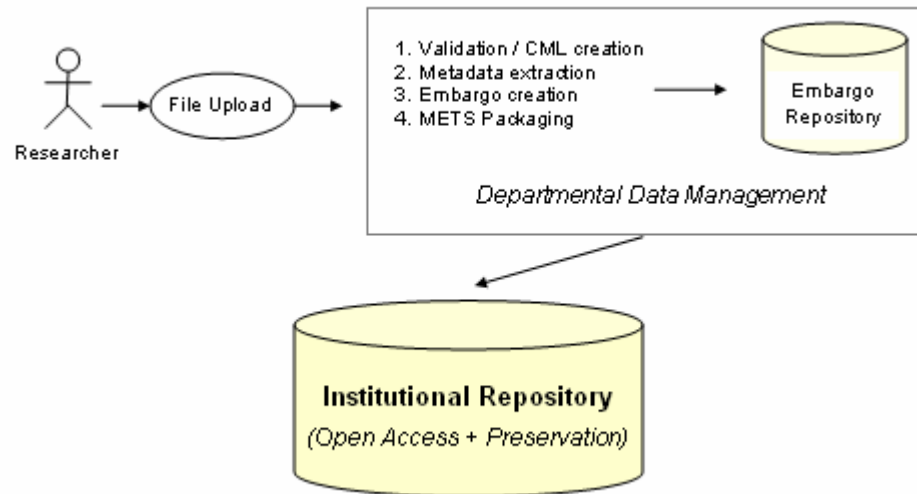
New feature (Controlled) public release



User search tools

**OAI-PMH Metadata
Harvesting**

Repository Deposition



Adding Metadata to NMR file package

Embargo


Embargo Period No embargo - publish immediately ▾

When period has elapsed

Publish automatically

Review

Experiment data

Experiment Date 24/07/1997 

Experiment Ref.

Spectrum Type

1D

Nucleus 1H ▾

2D

Nucleus [Select at least one]

- 1H
- 13C
- 15N
- 19F
- 31P
- Other

Pulse Sequence [Select one] ▾

Solvent CDCB ▾

Temperature Ambient ▾

Chemical data

Chemical Formula C8H8O1

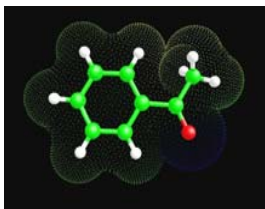
Systematic Chemical Name

Compound Class Organic ▾

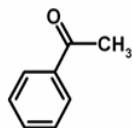
Authors

Chemist Initials (e.g. W.H.) W.H. LastName Perkin

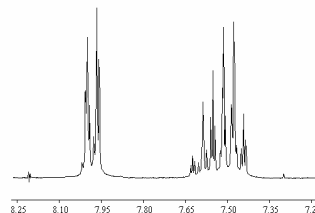
Supervisor



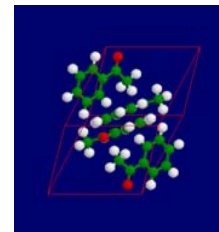
Computational
Chemistry Calculations



2D Chemical
Structures



NMR
Spectra



3D X-ray
Structures

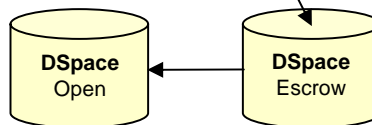
SPECTRa Deposit Tools
Create CML, InChI, metadata

InChI :

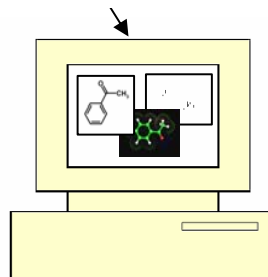
InChI=1/C8H8O/c1-7(9)8-5-3-2-4-6-8/h2-6H,1H3

CML :

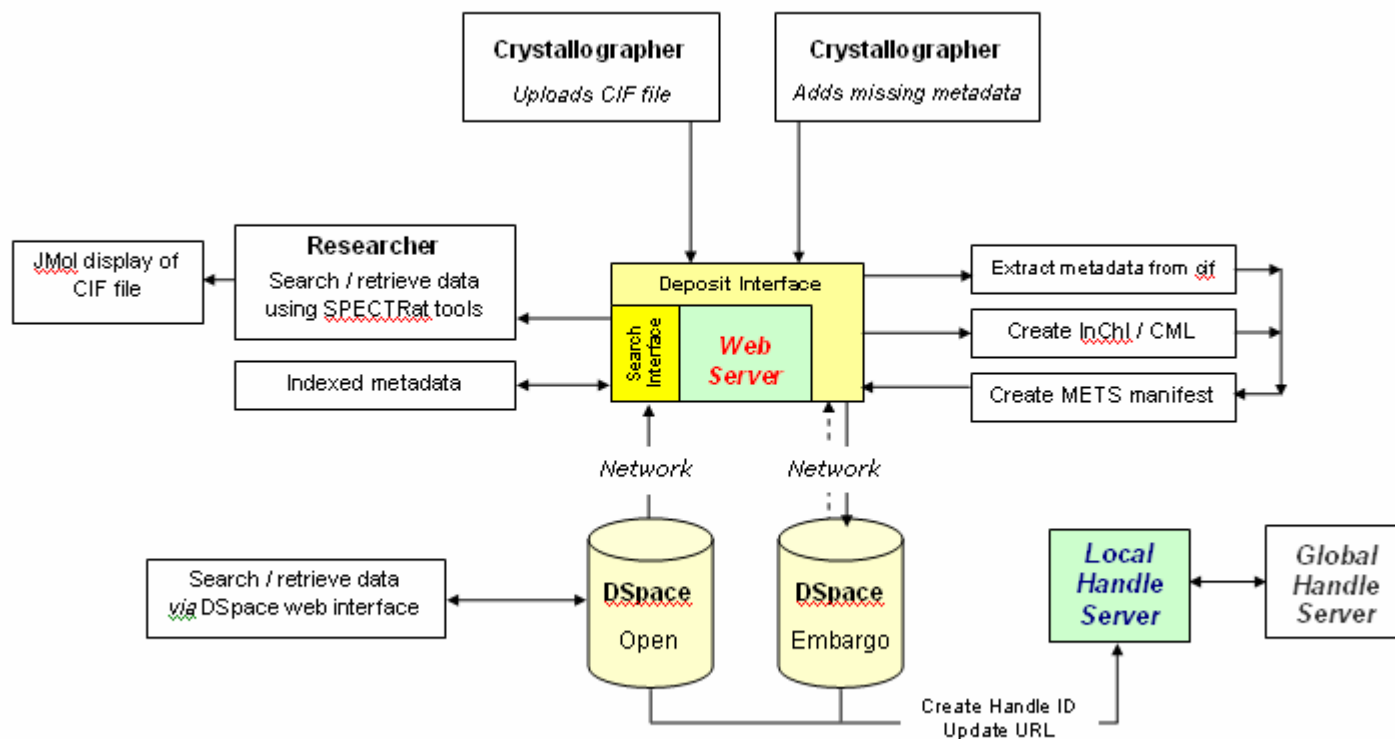
```
<molecule xmlns="http://www.xml.cml.org/schema">
<atomArray>
<atom id="a1" elementType="C" x2="-0.380600" y2="-0.720800"/>
<atom id="a2" elementType="C" x2="-0.381800" y2="-1.548200"/>
<atom id="a3" elementType="C" x2="0.333100" y2="-1.961000"/>
<atom id="a4" elementType="C" x2="1.049500" y2="-1.547700"/>
<atom id="a5" elementType="C" x2="1.046600" y2="-0.717200"/>
<atom id="a6" elementType="C" x2="0.331300" y2="-0.308000"/>
<atom id="a7" elementType="C" x2="1.759600" y2="-0.302000"/>
<atom id="a8" elementType="C" x2="2.475600" y2="-0.711800"/>
<atom id="a9" elementType="O" x2="1.756400" y2="0.523000"/>
</atomArray>
<bondArray>
<bond atomRefs2="a4 a5" order="1"/>
<bond atomRefs2="a2 a3" order="1"/>
<bond atomRefs2="a5 a6" order="2"/>
<bond atomRefs2="a6 a1" order="1"/>
<bond atomRefs2="a1 a2" order="2"/>
<bond atomRefs2="a5 a7" order="1"/>
<bond atomRefs2="a3 a4" order="2"/>
<bond atomRefs2="a7 a8" order="1"/>
<bond atomRefs2="a7 a9" order="2"/>
</bondArray>
</molecule>
```



SPECTRa Search Tools
OAI-PMH Harvesting



Crystallography Tool Architecture



Some Outcomes & Recommendations

- **Data Management** : No tradition amongst chemists (crystallographers apart) for organized deposition and re-use of experimental data.
- **Data re-use** : Additional analysis tools will be required to add value to large-scale data aggregates.
- **Legacy Data** : We did not appreciate the scale of non-conformance and changing standards for legacy file formats and data types.
- **IPR** : Who owns the deposited data? Guidelines for scientific data should be prepared by JISC in consultation with research funding bodies.
- **Data Management** : The project did not investigate the resource requirements for large-scale deposition and management of this experimental data

Acknowledgements

- **Project Director:** *Peter Morgan UL Cambridge*
- **Chemistry leads:** *Henry Rzepa, Peter Murray-Rust*
- **Project Officers:** *Fiona Cotterill, Jim Downing*
- **Project Manager:** *Alan Tonge*
- **Library Liaison:** *Janet Evans, Lorraine Windsor*

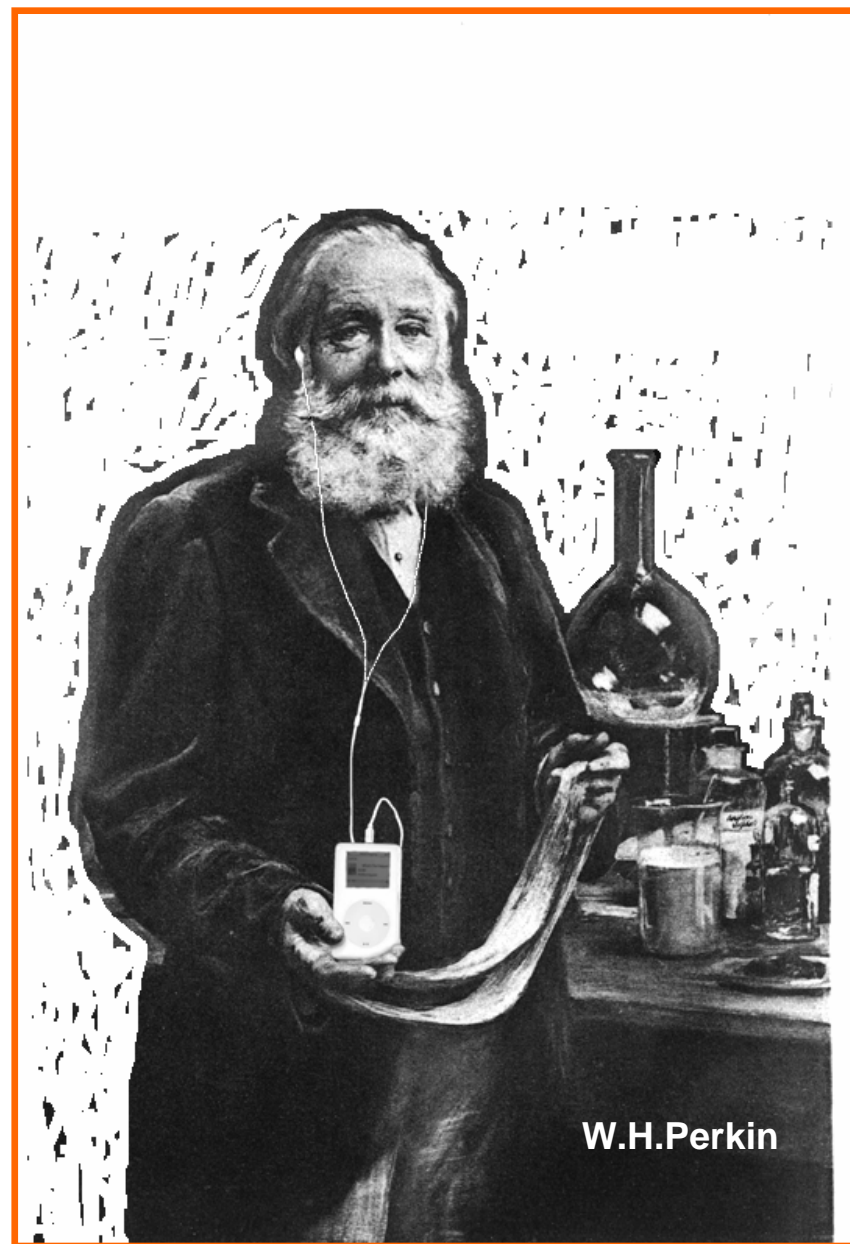
<http://www.lib.cam.ac.uk/spectra/>

JISC

 UNIVERSITY OF
CAMBRIDGE
Imperial College
London



Something completely different



W.H.Perkin

...a cool chemist