Deposition and use of raw diffraction images

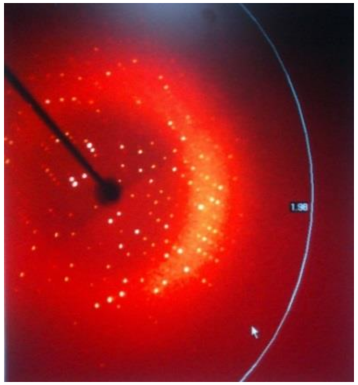
By

John.R. Helliwell

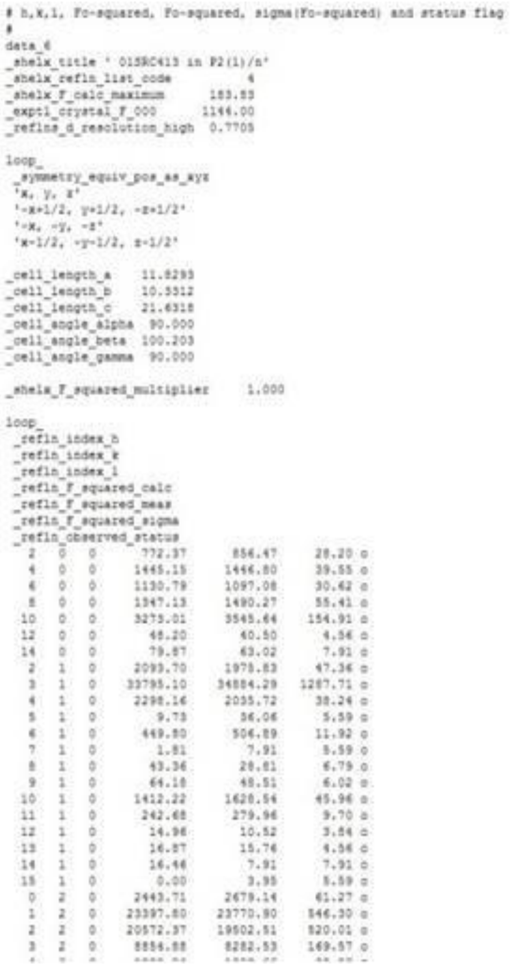
Crystallographic Information and Data Management
A Satellite Symposium to the 28th European Crystallographic Meeting
University of Warwick, Coventry, England
Sunday, August 25, 2013
Organised by
COMCIFS

Key terminology

- Raw data
  (few MB – few GB)

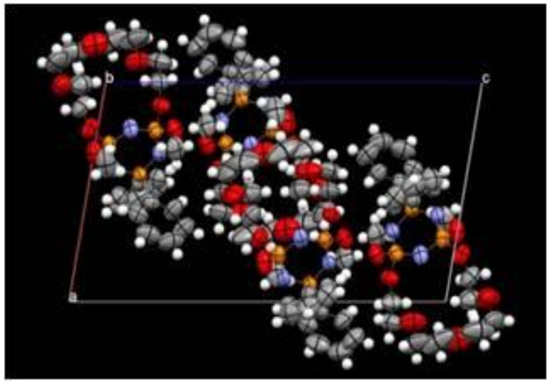


- Processed data
  (tens of kB – few MB)



```
# h,k,l, Fc-squared, Fo-squared, sigma(Fo-squared) and status flag
#
data_6
_shelx_title ' 01SRC413 in P2(1)/n'
_shelx_refln_list_code          4
_shelx_F_calc_maximum      183.53
_exptl_crystal_F_000      1144.00
_reflns_d_resolution_high  0.7705

loop_
 _symmetry_equiv_pos_as_xyz
 'x, y, z'
 '-x+1/2, y+1/2, -z+1/2'
 '-x, -y, -z'
 'x-1/2, -y-1/2, z-1/2'

_cell_length_a     11.8293
_cell_length_b     10.5312
_cell_length_c     21.6318
_cell_angle_alpha  90.000
_cell_angle_beta  100.203
_cell_angle_gamma  90.000

_shelx_F_squared_multiplier     1.000

loop_
 _refln_index_h
 _refln_index_k
 _refln_index_l
 _refln_F_squared_calc
 _refln_F_squared_meas
 _refln_F_squared_sigma
 _refln_observed_status
   2   0   0     772.37     856.47    28.20 o
   4   0   0    1445.15    1446.80    39.55 o
   6   0   0    1130.79    1097.08    30.62 o
   8   0   0    1347.13    1490.27    55.41 o
  10   0   0    3273.01    3545.64   154.91 o
  12   0   0      48.20      40.50     4.56 o
  14   0   0      79.87      63.02     7.91 o
   2   1   0    2093.70    1975.83    47.36 o
   3   1   0   33795.10   34884.29  1287.71 o
   4   1   0    2298.16    2035.72    38.24 o
   5   1   0       9.73      36.06     5.59 o
   6   1   0     449.80     506.89    11.92 o
   7   1   0       1.81       7.91     5.59 o
   8   1   0      43.36      28.81     6.79 o
   9   1   0      64.18      48.51     6.02 o
  10   1   0    1412.22    1628.54    45.96 o
  11   1   0     242.68     279.96     9.70 o
  12   1   0      14.98      10.52     3.84 o
  13   1   0      16.87      15.76     4.56 o
  14   1   0      16.48       7.91     7.91 o
  15   1   0       0.00       3.95     5.59 o
   0   2   0    2443.71    2679.14    61.27 o
   1   2   0   23397.80   23770.90   546.30 o
   2   2   0   20572.37   19602.51   520.01 o
   3   2   0    8854.88    8282.53   169.57 o
```

- Derived data
  (few kB – ~1 MB)

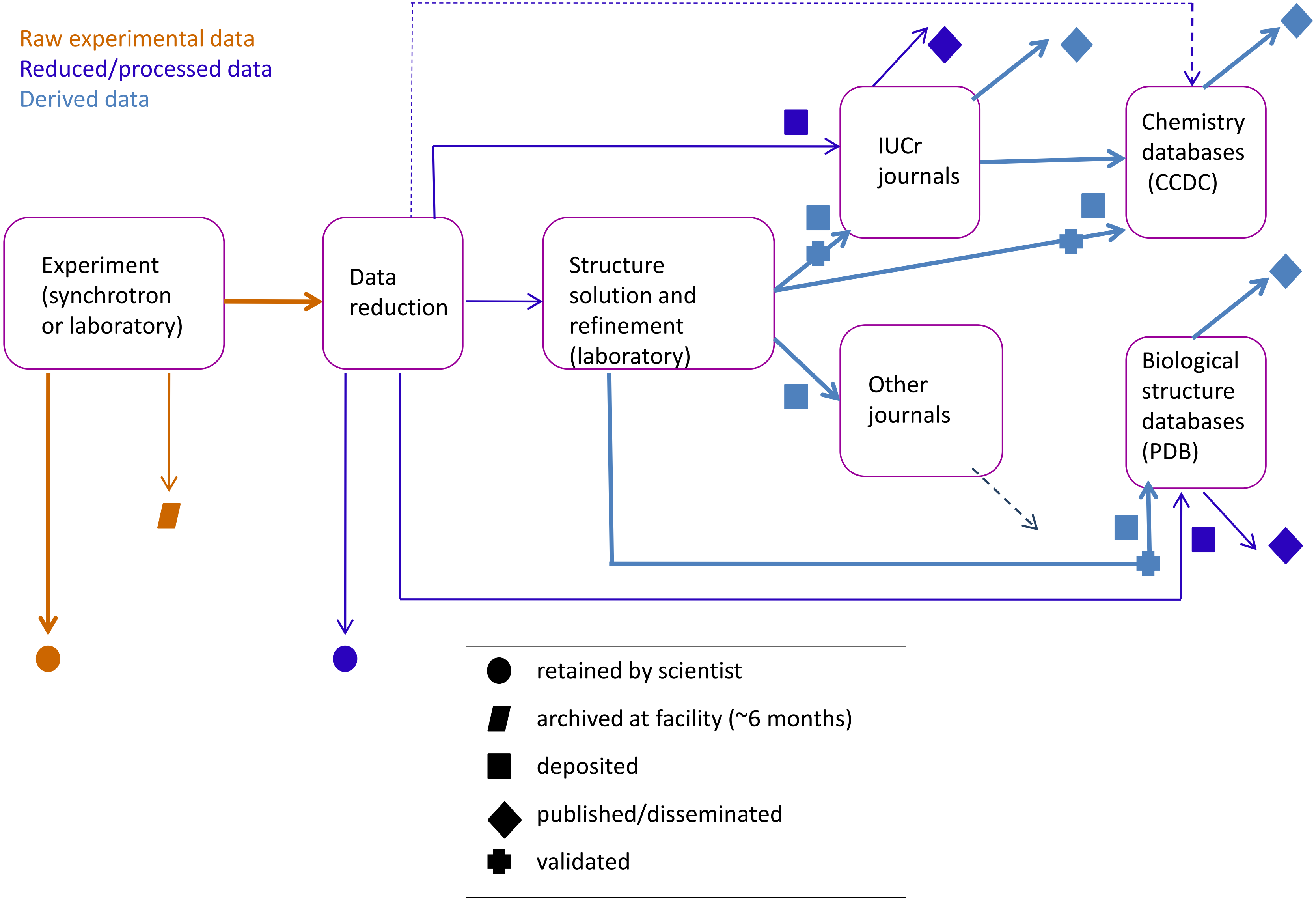


Why publish data?

Some reasons:

- Verify or support the validity of deductions from an experiment
- Safeguard against error or fraud
- Allow other scholars to conduct further research based on experiments already conducted
- Allow reanalysis at a later date, especially to extract 'new' science as new techniques are developed
- Provide example materials for teaching and learning
- As a mechanism for long-term preservation of experimental results

Data flow in crystallography



Raw experimental data

Reduced/processed data

Derived data

Experiment (synchrotron or laboratory)

Data reduction

Structure solution and refinement (laboratory)

IUCr journals

Other journals

Chemistry databases (CCDC)

Biological structure databases (PDB)

● retained by scientist

▰ archived at facility (~6 months)

■ deposited

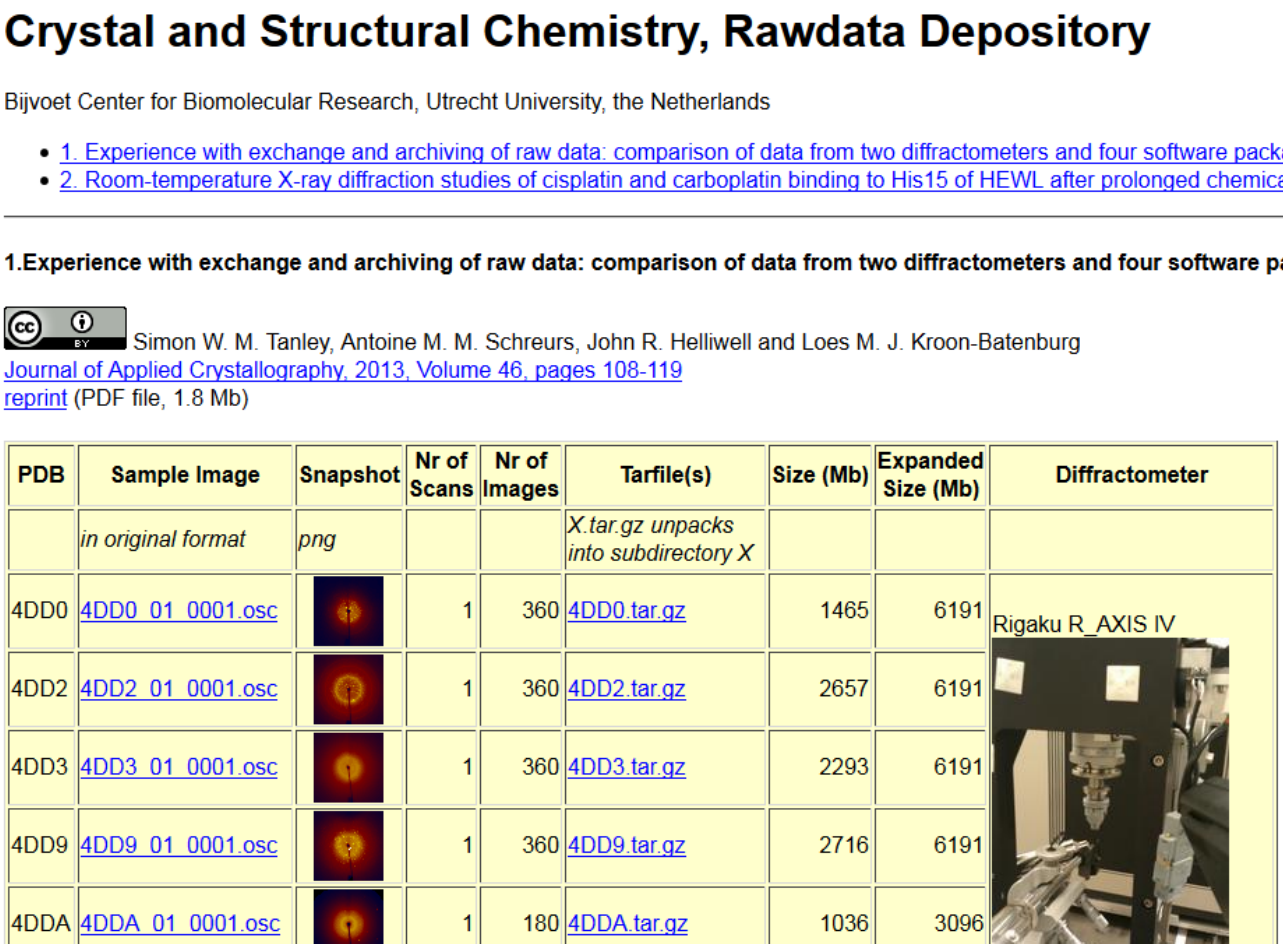
◆ published/disseminated

✚ validated

Why publish raw crystallographic diffraction data?

- Just what is the symmetry layout of a crystal namely its space-group symmetry?
- Just what is the diffraction resolution limit?
- The diffuse scattering may be significant and yield details of conformational mobility and or flexibility;
- Raw data availability can be used by developers to improve software *ie* our data processing tools;
- Raw data being an obligatory requirement for publication could serve to prevent fraud;
- Structure determination cannot proceed and needs a wider community effort *eg* if the diffraction is from an awkward composite of crystals.

Two of our publications have the raw data thus far both concern medical related studies:
J Appl Cryst 2013



Crystal and Structural Chemistry, Rawdata Depository

Bijvoet Center for Biomolecular Research, Utrecht University, the Netherlands

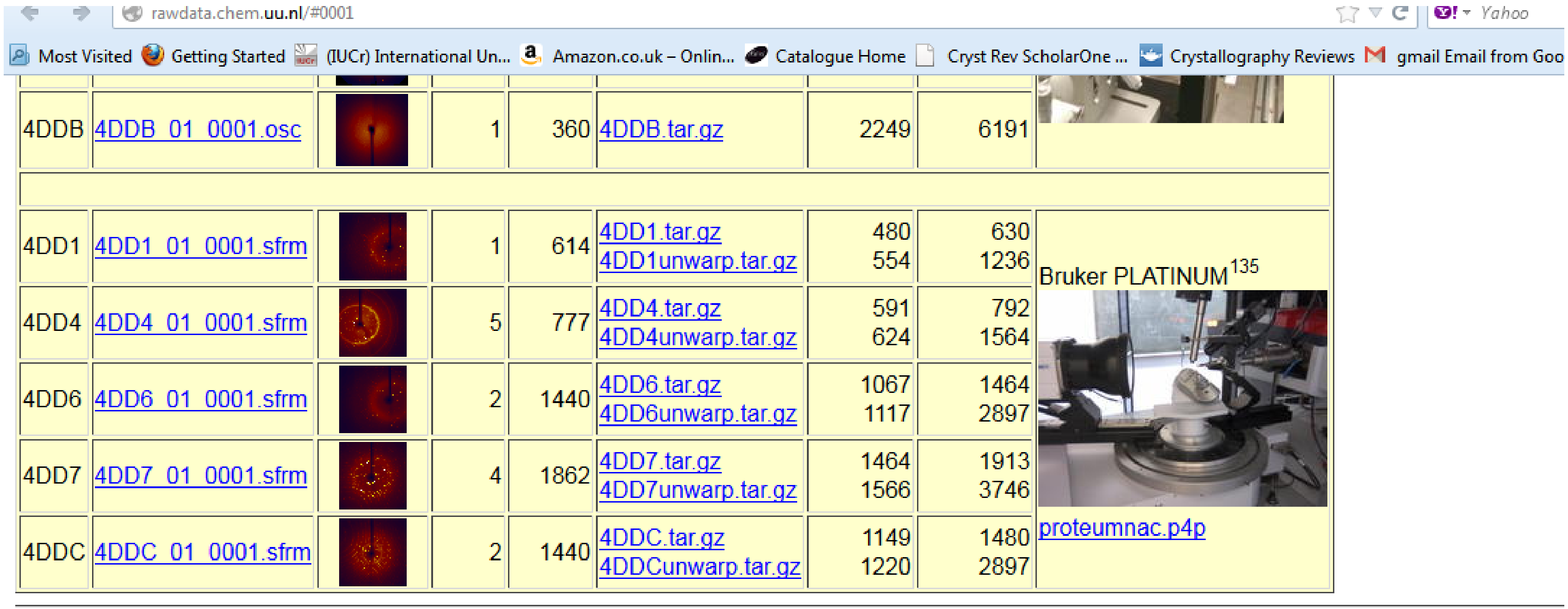
- 1. Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals
- 2. Room-temperature X-ray diffraction studies of cisplatin and carboplatin binding to His15 of HEWL after prolonged chemical exposure

1.Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals

Simon W. M. Tanley, Antoine M. M. Schreurs, John R. Helliwell and Loes M. J. Kroon-Batenburg
Journal of Applied Crystallography, 2013, Volume 46, pages 108-119
reprint (PDF file, 1.8 Mb)

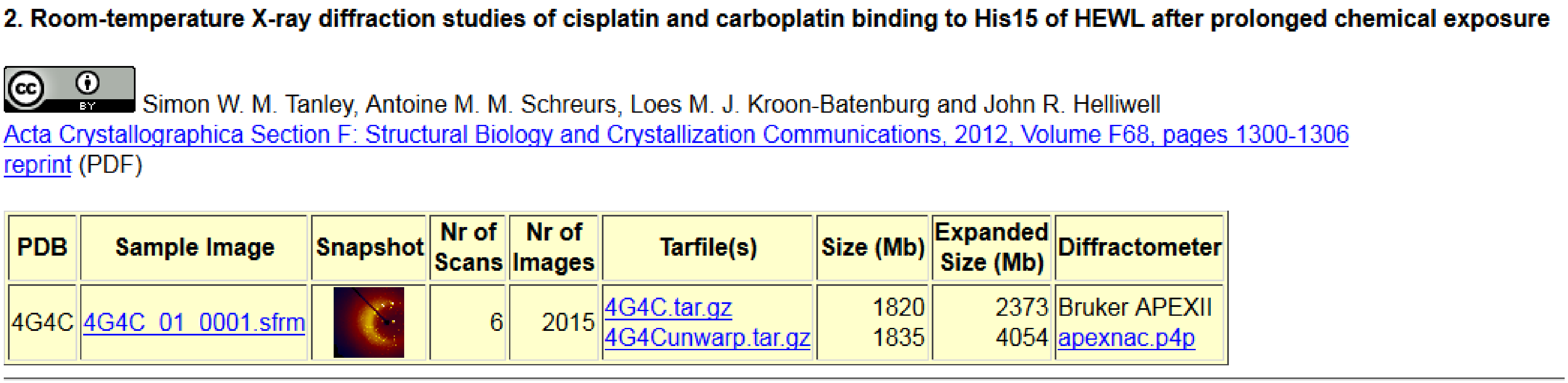
PDB	Sample Image	Snapshot	Nr of Scans	Nr of Images	Tarfile(s)	Size (Mb)	Expanded Size (Mb)	Diffractometer
	in original format	*png*			*X.tar.gz unpacks into subdirectory X*			
4DD0	4DD0_01_0001.osc		1	360	4DD0.tar.gz	1465	6191	Rigaku R_AXIS IV
4DD2	4DD2_01_0001.osc		1	360	4DD2.tar.gz	2657	6191	
4DD3	4DD3_01_0001.osc		1	360	4DD3.tar.gz	2293	6191	
4DD9	4DD9_01_0001.osc		1	360	4DD9.tar.gz	2716	6191	
4DDA	4DDA_01_0001.osc		1	180	4DDA.tar.gz	1036	3096	

J Appl Cryst and our raw data continued:-



4DDB	4DDB_01_0001.osc		1	360	4DDB.tar.gz	2249	6191	
4DD1	4DD1_01_0001.sfrm		1	614	4DD1.tar.gz 4DD1unwarp.tar.gz	480 554	630 1236	Bruker PLATINUM^{135} proteumnac.p4p
4DD4	4DD4_01_0001.sfrm		5	777	4DD4.tar.gz 4DD4unwarp.tar.gz	591 624	792 1564	
4DD6	4DD6_01_0001.sfrm		2	1440	4DD6.tar.gz 4DD6unwarp.tar.gz	1067 1117	1464 2897	
4DD7	4DD7_01_0001.sfrm		4	1862	4DD7.tar.gz 4DD7unwarp.tar.gz	1464 1566	1913 3746	
4DDC	4DDC_01_0001.sfrm		2	1440	4DDC.tar.gz 4DDCunwarp.tar.gz	1149 1220	1480 2897	

And the raw data for our recent article in *Acta Cryst F* 2013



2. Room-temperature X-ray diffraction studies of cisplatin and carboplatin binding to His15 of HEWL after prolonged chemical exposure

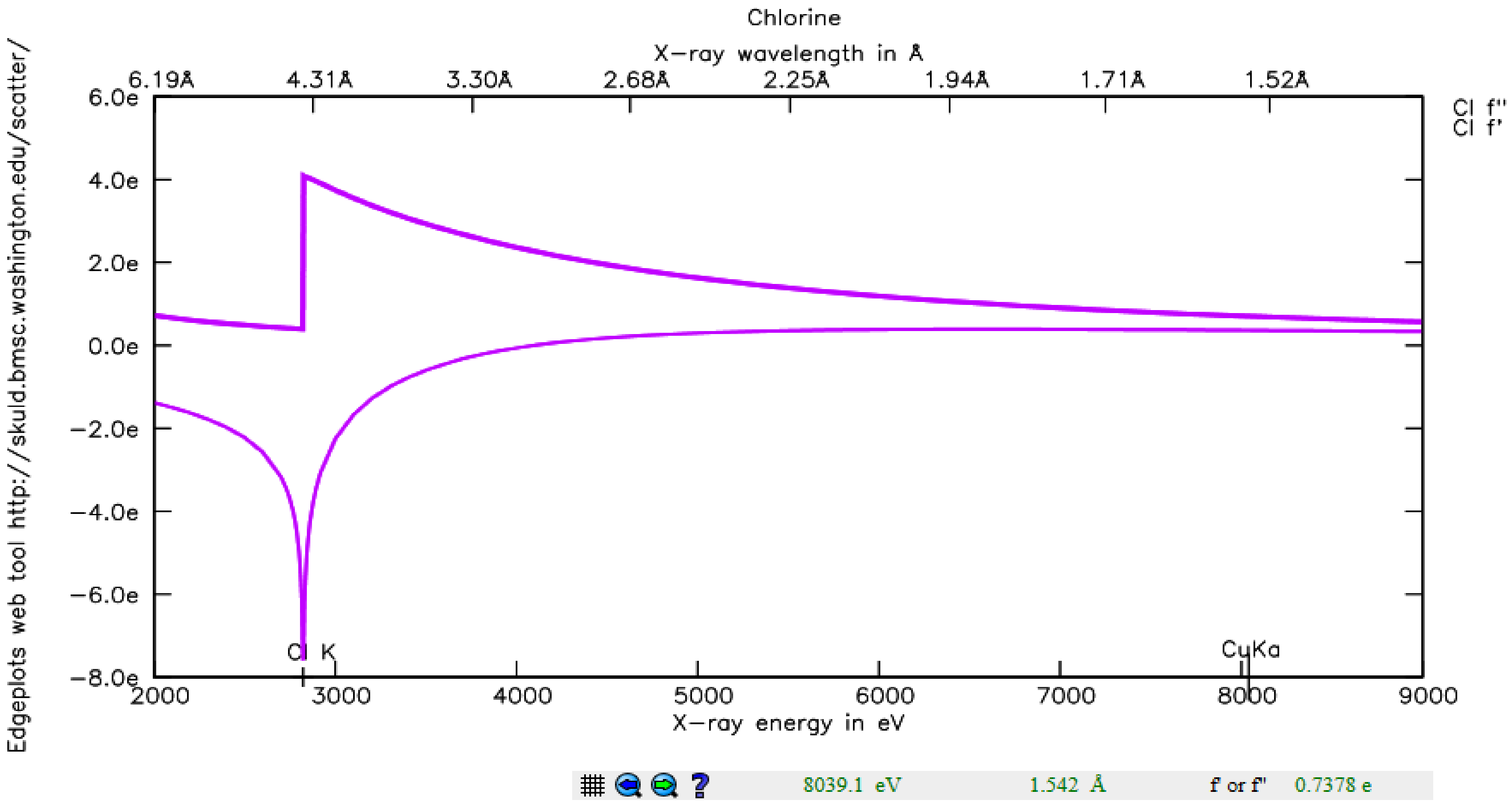
Simon W. M. Tanley, Antoine M. M. Schreurs, Loes M. J. Kroon-Batenburg and John R. Helliwell
Acta Crystallographica Section F: Structural Biology and Crystallization Communications, 2012, Volume F68, pages 1300-1306
reprint (PDF)

PDB	Sample Image	Snapshot	Nr of Scans	Nr of Images	Tarfile(s)	Size (Mb)	Expanded Size (Mb)	Diffractometer
4G4C	4G4C_01_0001.sfrm		6	2015	4G4C.tar.gz 4G4Cunwarp.tar.gz	1820 1835	2373 4054	Bruker APEXII apexnac.p4p

Why publish raw crystallographic diffraction data?

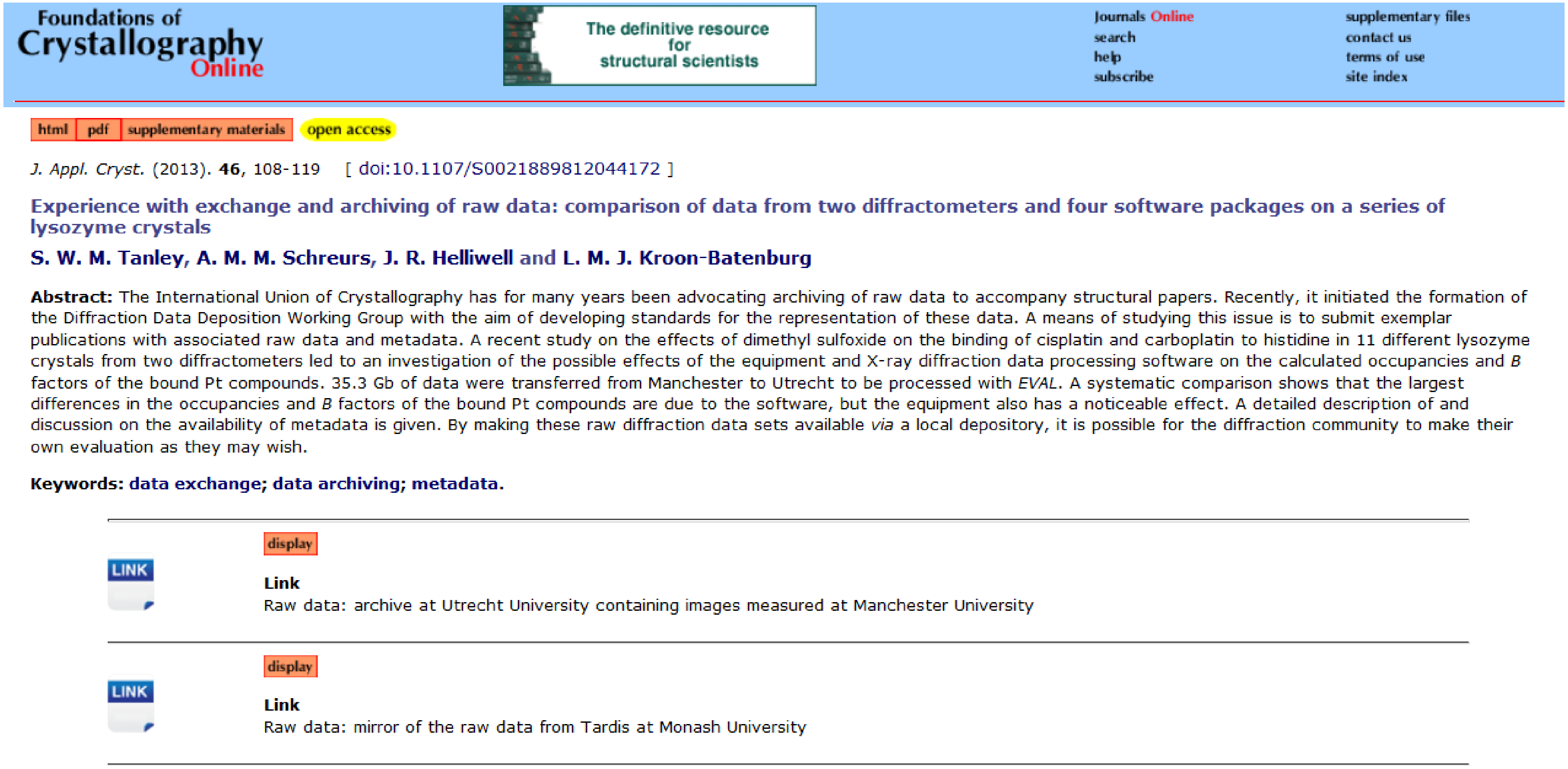
- There is an even simpler reason (see next slide); comparing and harnessing different software to chase down weak anomalous signals>>>>>

Weak structural signals such as anomalous dispersion far from resonance such as chlorine>>>>>



8039.1 eV 1.542 Å f or f'' 0.7378 e

Our *J. Appl Cryst* raw data were downloaded by Dr Kay Diederichs of XDS software



J. Appl. Cryst. (2013). **46**, 108-119 [doi:10.1107/S0021889812044172]

Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals

S. W. M. Tanley, A. M. M. Schreurs, J. R. Helliwell and L. M. J. Kroon-Batenburg

Abstract: The International Union of Crystallography has for many years been advocating archiving of raw data to accompany structural papers. Recently, it initiated the formation of the Diffraction Data Deposition Working Group with the aim of developing standards for the representation of these data. A means of studying this issue is to submit exemplar publications with associated raw data and metadata. A recent study on the effects of dimethyl sulfoxide on the binding of cisplatin and carboplatin to histidine in 11 different lysozyme crystals from two diffractometers led to an investigation of the possible effects of the equipment and X-ray diffraction data processing software on the calculated occupancies and *B* factors of the bound Pt compounds. 35.3 Gb of data were transferred from Manchester to Utrecht to be processed with *EVAL*. A systematic comparison shows that the largest differences in the occupancies and *B* factors of the bound Pt compounds are due to the software, but the equipment also has a noticeable effect. A detailed description of and discussion on the availability of metadata is given. By making these raw diffraction data sets available *via* a local depository, it is possible for the diffraction community to make their own evaluation as they may wish.

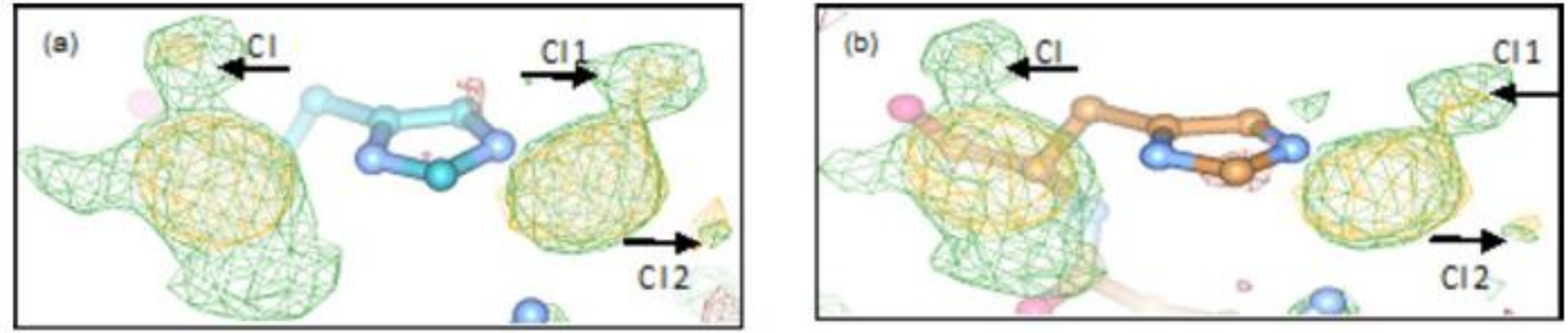
Keywords: data exchange; data archiving; metadata.

Link
Raw data: archive at Utrecht University containing images measured at Manchester University

Link
Raw data: mirror of the raw data from Tardis at Monash University

Example of a low *Z* anom scatterer in pharmaceutical chemistry: partial chemical conversion of carboplatin to cisplatin under a high (10%) NaCl concentration

This means we have the challenge to see a partially occupied chlorine at λ = 1.54 Å (f " 0.7)



The two binding sites on the His-15 residue of HEWL. Fo-Fc density OMIT maps (green) and anomalous difference density (orange) maps at 3σ cut off; 4dd7 processed by software **(a) XDS and (b) EVAL**

Tanley *et al*; 2013 *JSR* in press

Our open archiving of our raw diffraction data images allowed wider software use and led to new results at the limits of visibility in terms of Signal to Noise

Now a short summary of policy matters>>>

The IUCr Diffraction Data Deposition Working Group

Established by IUCr Summer 2011

Terms of Reference

- It is becoming increasingly important to deposit the raw data from scattering experiments;
- A lot of valuable information gets lost when only structure factors are deposited.
- A number of research centres, *e.g.* synchrotron and neutron facilities, are fully aware of the need and have established detector working groups addressing this issue.

Membership

Full members	Steve Androulakis *Australia*	Brian McMahon *UK*
	Sol Gruner *USA*	Tom Terwilliger *USA*
	John R. Helliwell, Chair *UK*	John Westbrook *USA*
	Loes Kroon-Batenburg*The Netherlands*	Heinz-Josef Weyer *Switzerland*
By invitation	*Chairs and delegates of IUCr Commissions*	
Consultants	*Currently five specialists in data archiving, software development and macromolecular crystallography*	

Initial options studied

Centralised discipline-specific archives

- Curated databases (CCDC, PDB)
  - Additional storage and curation not affordable with current funding models
  - Currently handle mostly published structures
- IUCr journals
  - Insufficient storage and network capacity
  - Incomplete coverage of the literature

Centralised experimental facilities (synchrotron, neutron)

- Diverse policies and technical capabilities
- Diverse types of scientific information
- Do not cater for 'home' laboratory experiments

Working towards a federation of localised repositories (near to where data measured)

- Commission on Synchrotron Radiation has started a survey of SR Facilities (8 reported so far) suggesting that this is promising as an option; but each SR facility emphasised that they are not to be regarded as an archive. Neither
    - instantaneous delivery of data
    - provision of data sets certified to be 100% 'free of data corruption'

  could be guaranteed.
- Universities Data Repositories/Archives, even at the most advanced in their planning (*e.g.* University of Manchester), have yet to be proven in practice, *e.g.* with respect to the two issues above.

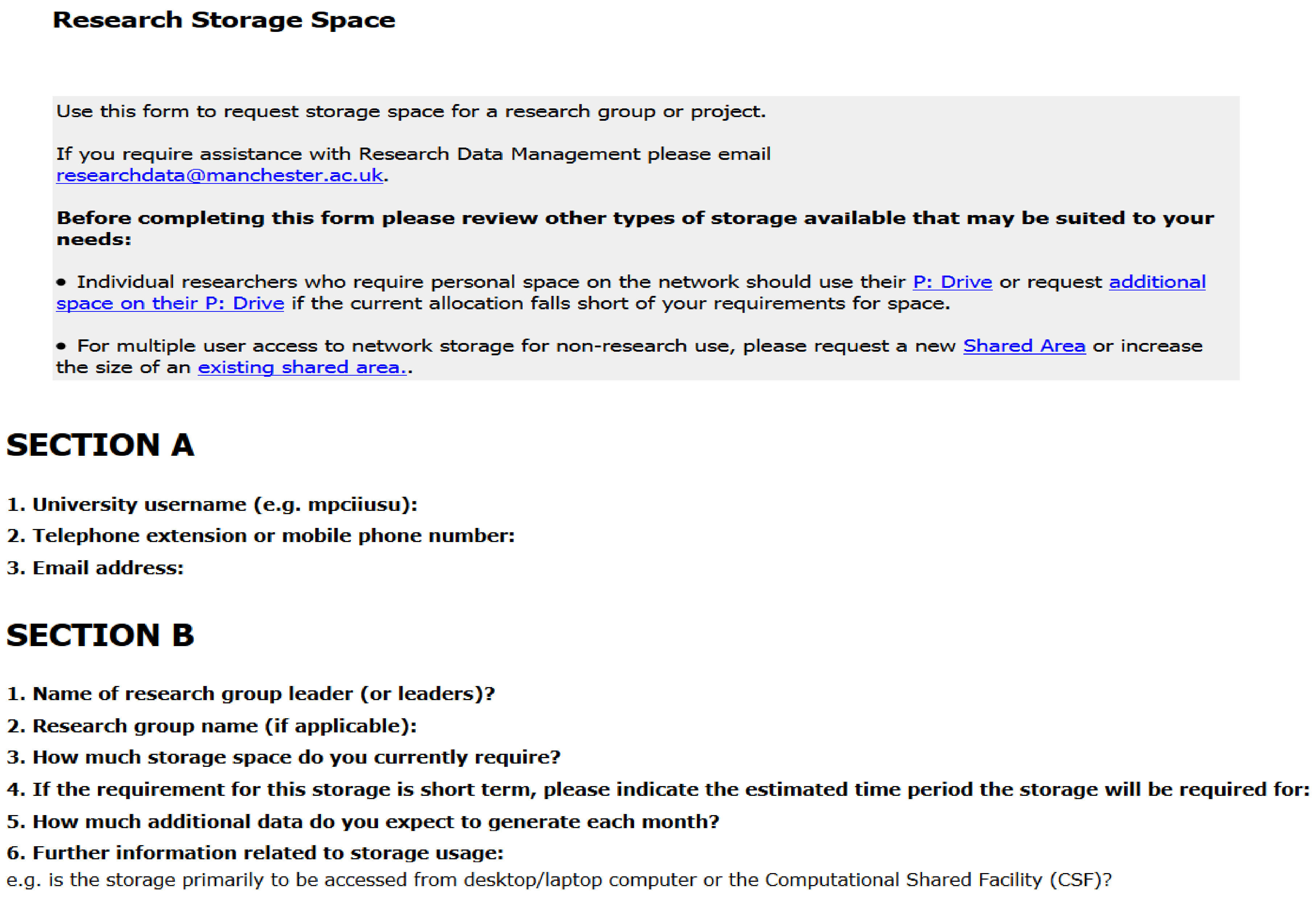
University of Manchester Research Data Archive launch September 2013

Two key points, as examples:-

The University endorses the **RCUK Common Principles on Data Policy** and requires all its staff and students to adhere to them, as well as taking into account any other research data management requirements that may apply.

Should the **Principal Investigator** leave the University or be unable to continue in the role before all his/her duties relating to the data have been discharged, it is the responsibility of his/her Head of School to appoint a replacement.

And a UniMan data storage request form!



Research Storage Space

Use this form to request storage space for a research group or project.

If you require assistance with Research Data Management please email researchdata@manchester.ac.uk.

Before completing this form please review other types of storage available that may be suited to your needs:

- Individual researchers who require personal space on the network should use their P: Drive or request additional space on their P: Drive if the current allocation falls short of your requirements for space.
- For multiple user access to network storage for non-research use, please request a new Shared Area or increase the size of an existing shared area..

SECTION A

1. University username (e.g. mpciiusu):

2. Telephone extension or mobile phone number:

3. Email address:

SECTION B

1. Name of research group leader (or leaders)?

2. Research group name (if applicable):

3. How much storage space do you currently require?

4. If the requirement for this storage is short term, please indicate the estimated time period the storage will be required for:

5. How much additional data do you expect to generate each month?

6. Further information related to storage usage:

e.g. is the storage primarily to be accessed from desktop/laptop computer or the Computational Shared Facility (CSF)?

Additional ‘fallback’ positions

- Corresponding authors set up web links to the data sets that underpin their publications.
- These may be or may be not DOI linked: such a requirement would be difficult to enforce although journals could ‘strongly recommend’.
- How would such a method for data archiving and access by readers be kept up to date, *e.g.* in the event of an author retiring (or what to do after their death?).

Current perspective

- There is enthusiasm and encouragement to archive more than derived or processed data in many areas of science besides our own.
- The crystallographic community prides itself in making its processed data accompany its publications; indeed it has been obligatory these last 10 years or so.
- We have three practical options in the near future to extend these principles to our raw data;
  - via the local Data Archive
  - via synchrotron data storage
  - or via the corresponding author setting up a personal link to datasets underpinning publications on their personal websites.

The future as seen by the particle physicists

- Use cloud storage;

- Our reaction as crystallographers:-
- Does this mean using commercial data storage suppliers like Google?
- So, do we feel comfortable trusting our data to a commercial agent?
- Cost issues also need to be evaluated carefully, but look promisingly, *ie* relatively, cheap;

Initial Proposals to IUCr Executive Committee

Initial recommendations to the IUCr Executive Committee in December 2012:

1) ***Authors should provide a permanent and prominent link from an article to the raw data sets underpinning a journal publication***

   (with a view to making this a formal requirement on authors at such time as the community has adopted raw data deposition as a routine procedure)

2) ***Commissions should be charged with the task of defining experimental metadata relevant to their scientific fields in order to harmonise raw data archiving at disparate facilities***

"should" changed to "may" by IUCr Exec at its meeting held Dec 2012 in Adelaide.

A role for CODATA

IUCr response to ICSU review of CODATA: *Q11. What, if anything, could CODATA do to serve you better?*

- Rationalise terminology of data descriptions: *raw, processed, derived, big, massive,* ...
- Promote guidelines for data archiving, building on best practice in different scientific disciplines
- Continue efforts to develop ontologies, metadata and interoperability standards

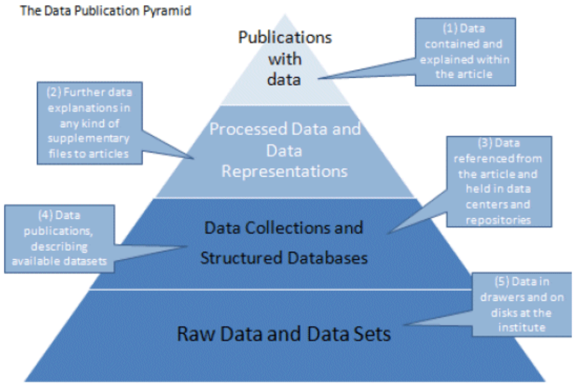
... through Task Groups and focused initiatives (*e.g.* nanomaterials)

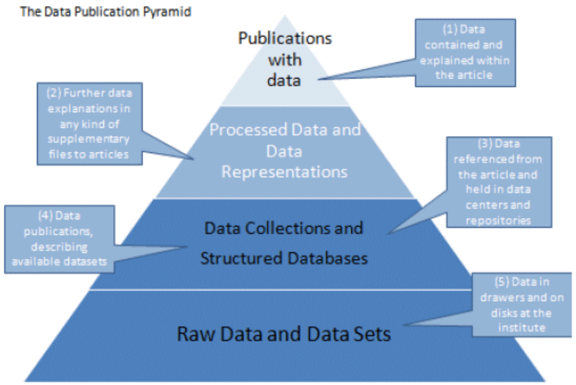
Acknowledgements

- I thank all members of the IUCr Working Group on Data and the IUCr Commission representatives and its Consultants
- I thank the participants at the major Workshop on these topics held at the European Crystallographic Meeting ‘ECM27’ in Bergen

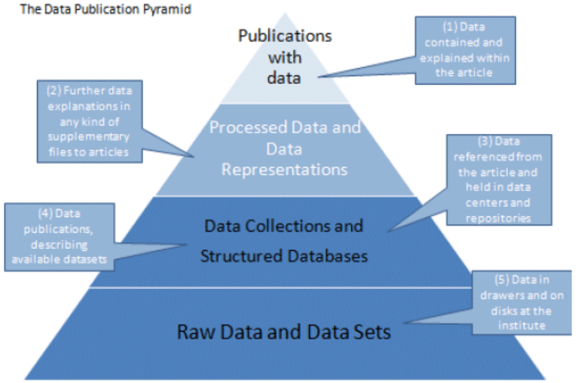
  http://www.iucr.org/resources/data/dddwg/bergen-workshop

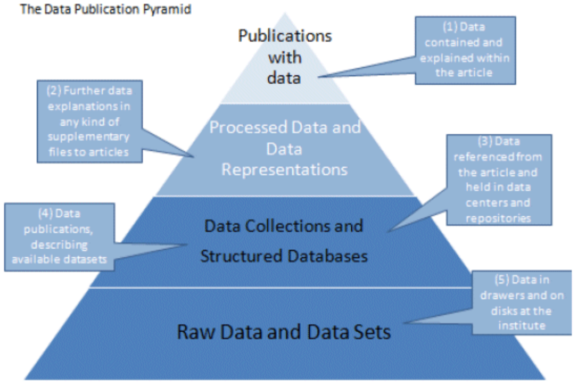
Dr Tom Terwilliger is bringing together a Special Issue for Acta Cryst D derived from the Workshop Lectures.


The Data Publication Pyramid
Publications with data
Processed Data and Data Representations
Data Collections and Structured Databases
Raw Data and Data Sets
(1) Data contained and explained within the article
(2) Further data explanations in any kind of supplementary files to articles
(3) Data referenced from the article and held in data centers and repositories
(4) Data publications, describing available datasets
(5) Data in drawers and on disks at the institute


The Data Publication Pyramid
Publications with data
Processed Data and Data Representations
Data Collections and Structured Databases
Raw Data and Data Sets
(1) Data contained and explained within the article
(2) Further data explanations in any kind of supplementary files to articles
(3) Data referenced from the article and held in data centers and repositories
(4) Data publications, describing available datasets
(5) Data in drawers and on disks at the institute

Thank you


The Data Publication Pyramid
Publications with data
Processed Data and Data Representations
Data Collections and Structured Databases
Raw Data and Data Sets
(1) Data contained and explained within the article
(2) Further data explanations in any kind of supplementary files to articles
(3) Data referenced from the article and held in data centers and repositories
(4) Data publications, describing available datasets
(5) Data in drawers and on disks at the institute


The Data Publication Pyramid
Publications with data
Processed Data and Data Representations
Data Collections and Structured Databases
Raw Data and Data Sets
(1) Data contained and explained within the article
(2) Further data explanations in any kind of supplementary files to articles
(3) Data referenced from the article and held in data centers and repositories
(4) Data publications, describing available datasets
(5) Data in drawers and on disks at the institute