WORKSHOP ON WHEN SHOULD SMALL MOLECULE CRYSTALLOGRAPHERS PUBLISH RAW DIFFRACTION DATA?

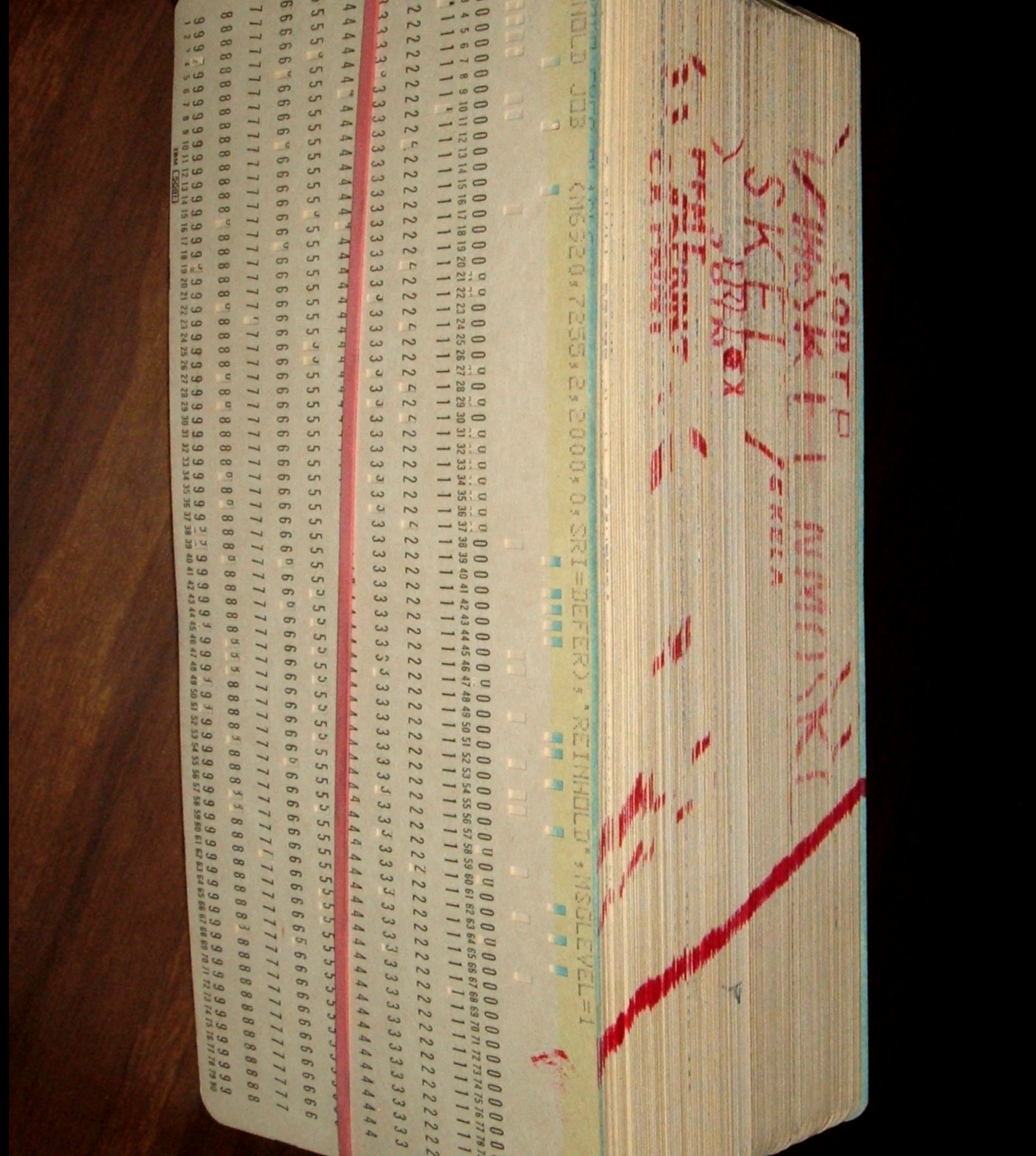
GRAEME WINTER / DIAMOND LIGHT SOURCE

FUTURE OUTLOOK FOR CURATED DATA ARCHIVE

diamond

OVERVIEW

- Historical viewpoint - archiving in crystallography
- Features of a useful raw data archive
- Curation
- Supporting data archiving - cost / benefit
- Conclusions



I am considering biological and chemical crystallography as “the same problem” here

CRYSTALLOGRAPHY AS A DATA SCIENCE

Crystallography highly data driven -

- Determine results fully from experimental data and prior knowledge
- The “shape” of the data are well known in advance (i.e. not “messy”)
- Influence of interpretation much reduced compared to e.g. geology
- Process significantly automated, powerful library of tools used for analysis
- Comparable with radio astronomy as observational / data driven science

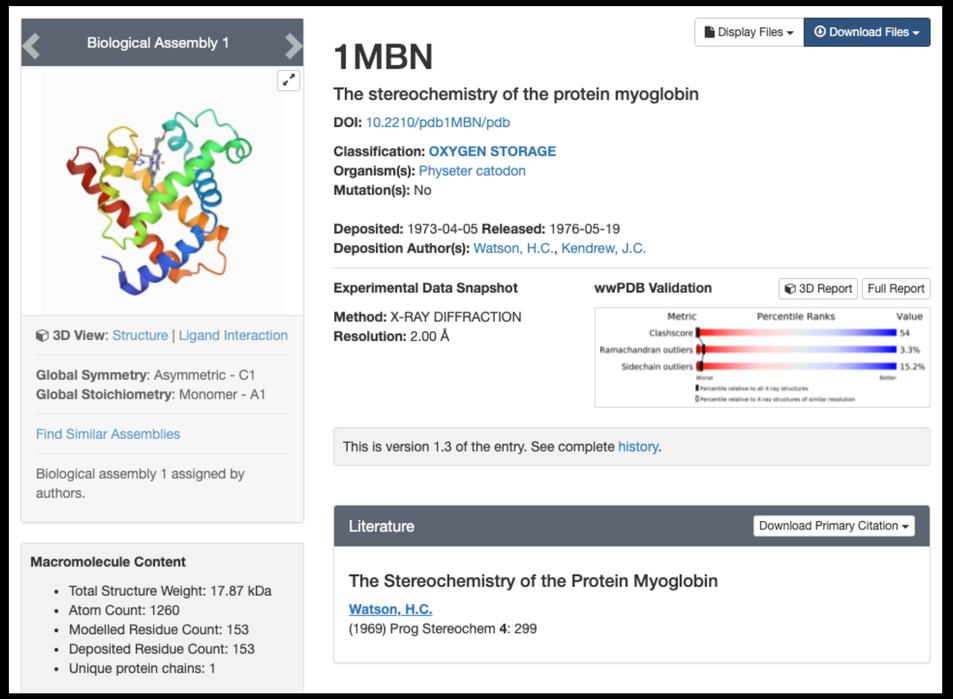
ARCHIVING IN CRYSTALLOGRAPHY

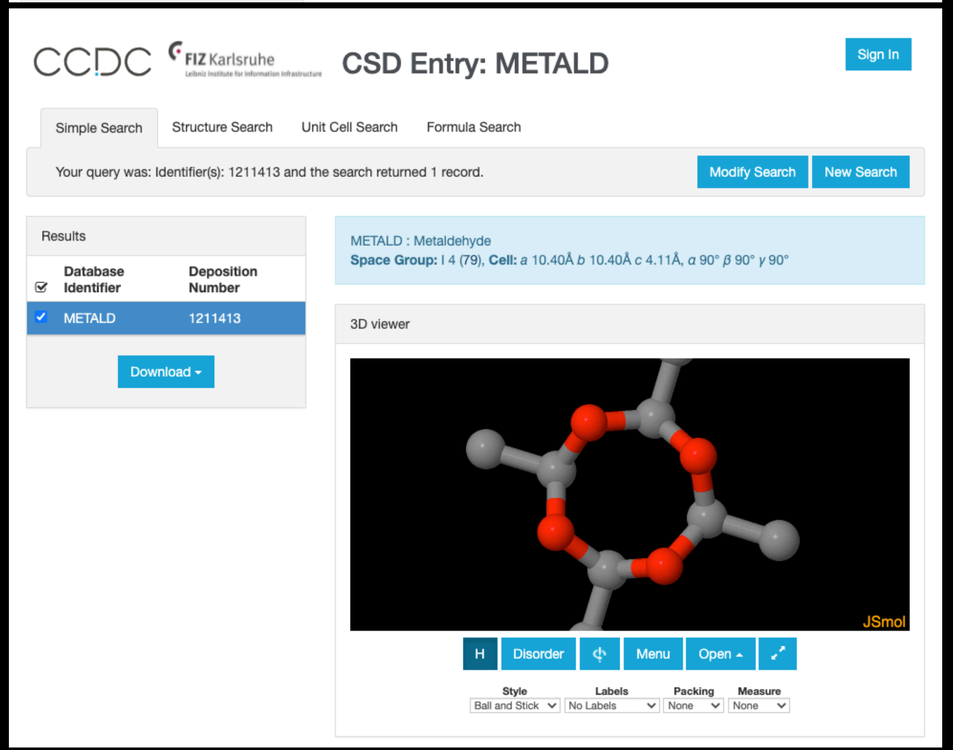
Long history of data banks / archives in X-ray crystallography

- CSD - 1965
- PDB - 1971
- ICSD - 1979

Crystallography pioneered open data archives

Computer-based archiving fundamental





CONTENT OF AN ARCHIVE RECORD

- 3D atomic coordinates
- Report of experiment - maybe
- Reference to publication - often
- Experimental data (processed) - maybe
- Reference to external data - maybe

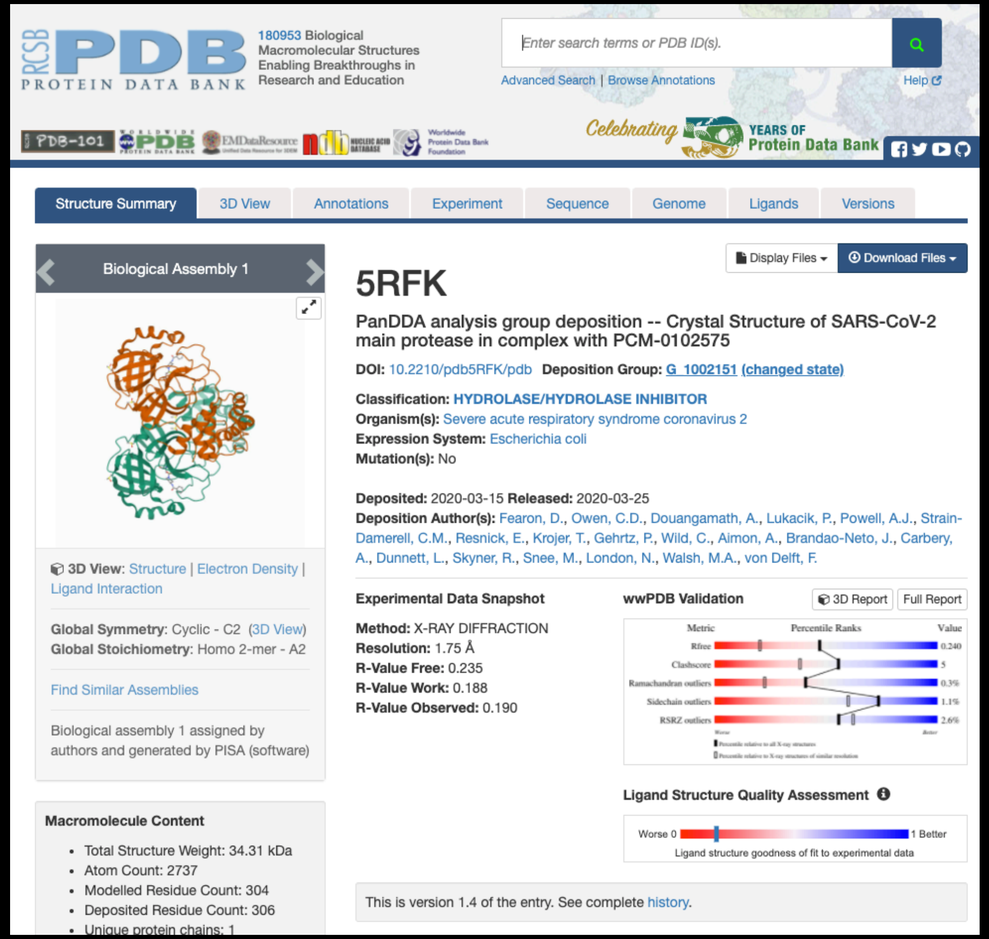
```
##################################################################
#
# This file contains crystal structure data downloaded from the
# Cambridge Structural Database (CSD) hosted by the Cambridge
# Crystallographic Data Centre (CCDC).
#
# Full information about CCDC data access policies and citation
# guidelines are available at http://www.ccdc.cam.ac.uk/access/V1
#
# Audit and citation data items may have been added by the CCDC.
# Please retain this information to preserve the provenance of
# this file and to allow appropriate attribution of the data.
#
##################################################################

data_CAFINE
#This CIF has been generated from an entry in the Cambridge Structural Database
_database_code_depnum_ccdc_archive 'CCDC 1119028'
_database_code_CSD               CAFINE
loop_
_citation_id
_citation_doi
_citation_year
1 10.1107/S0365110X58001286 1958
_audit_creation_method           'Created from the CSD'
_audit_update_record
;
2021-08-09 downloaded from the CCDC.
;
_database_code_NBS               504758
_chemical_name_common            'Caffeine monohydrate'
_chemical_formula_moiety         'C8 H10 N4 O2,H2 O1'
_chemical_name_systematic        '1,3,7-Trimethyl-purine-2,6-dione monohydrate'
_chemical_properties_biological  'stimulant which increases CNS activity'
_chemical_absolute_configuration unk
_diffrn_ambient_temperature      295
_exptl_crystal_density_diffrn    1.447
#These two values have been output from a single CSD field.
_refine_ls_R_factor_gt           0.146
_refine_ls_wR_factor_gt          0.146
_diffrn_radiation_probe          x-ray
_symmetry_cell_setting           monoclinic
_symmetry_space_group_name_H-M   'P 21/a'
_symmetry_Int_Tables_number      14
_space_group_name_Hall           '-P 2yab'
loop_
_symmetry_equiv_pos_site_id
_symmetry_equiv_pos_as_xyz
1 x,y,z
2 1/2-x,1/2+y,-z
3 -x,-y,-z
4 1/2+x,1/2-y,z
_cell_length_a                   14.8(1)
_cell_length_b                   16.7(1)
_cell_length_c                   3.97(3)
_cell_angle_alpha                90
_cell_angle_beta                 97.0(5)
_cell_angle_gamma                90
_cell_volume                     973.911
_cell_formula_units_Z            4
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
C1 C 0.24140 0.22250 -0.09980
C2 C 0.10030 0.25330 0.12950
C3 C 0.08410 0.17590 0.19440
C4 C 0.14630 0.11430 0.11550
C5 C -0.01990 0.25200 0.36380
C6 C 0.28910 0.08320 -0.12100
C7 C 0.19590 0.36380 -0.07910
C8 C -0.04640 0.10530 0.45840
N1 N 0.21960 0.14150 -0.02650
N2 N 0.18010 0.27690 -0.01520
N3 N 0.00200 0.17490 0.33760
N4 N 0.04030 0.30080 0.24400
O1 O 0.30630 0.24000 -0.23860
O2 O 0.13630 0.04040 0.16160
H1 H -0.08700 0.26100 0.47400
H2 H -0.01300 0.06200 0.59900
H3 H -0.06500 0.06300 0.27800
H4 H -0.10500 0.13700 0.51000
H5 H 0.26300 0.36200 -0.14300
H6 H 0.22800 0.39600 0.10500
H7 H 0.14200 0.37700 -0.21700
H8 H 0.34800 0.10000 -0.22800
H9 H 0.30000 0.03300 0.02200
H10 H 0.25700 0.06000 -0.32400
O3 O 0.01840 0.47050 0.27050

#END
```

CONTENT OF A USEFUL ARCHIVE RECORD

- 3D atomic coordinates
- Report of experiment
- Reference to publication
- Experimental data (processed)
- Reference to external data



ACCESS TO DATA ARCHIVE

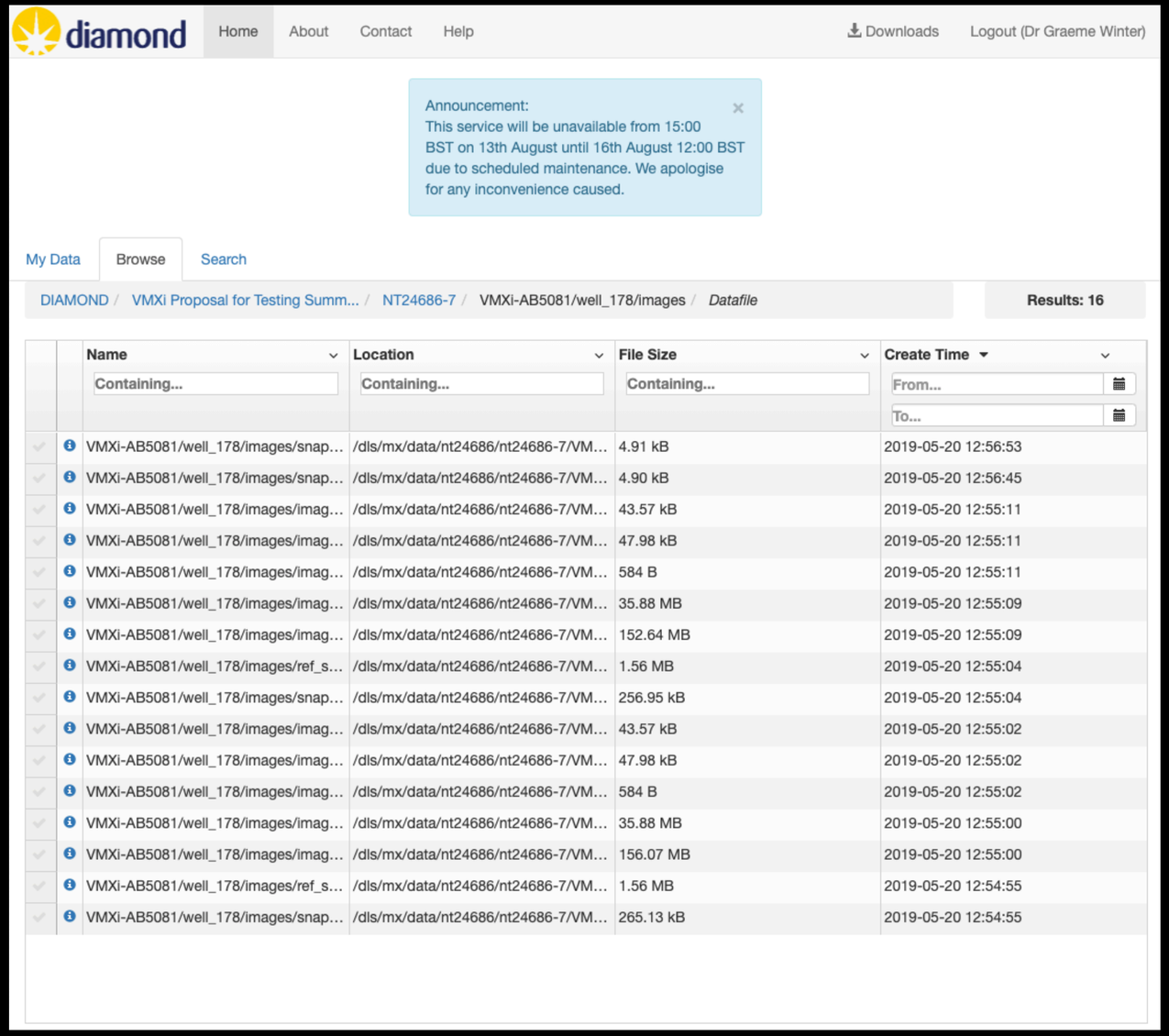
- Data are generally freely available in predefined format (CIF, PDB, mmCIF, …)
- Services may not be freely available e.g. ability to usefully search, links to other data archives
- wwPDB made up of RCSB / EBI (PDBe) and PDBj - “competition” between databases, built on the same underlying databank
- CSD funded by CCDC as not for profit company

FEATURES OF USEFUL RAW DATA ARCHIVE

- Easy to search, well integrated with existing data archives (PDB, CSD etc.)
- Inclusive / open to all depositors / open to all users
- Curated
- Funded / sustainable / long lived

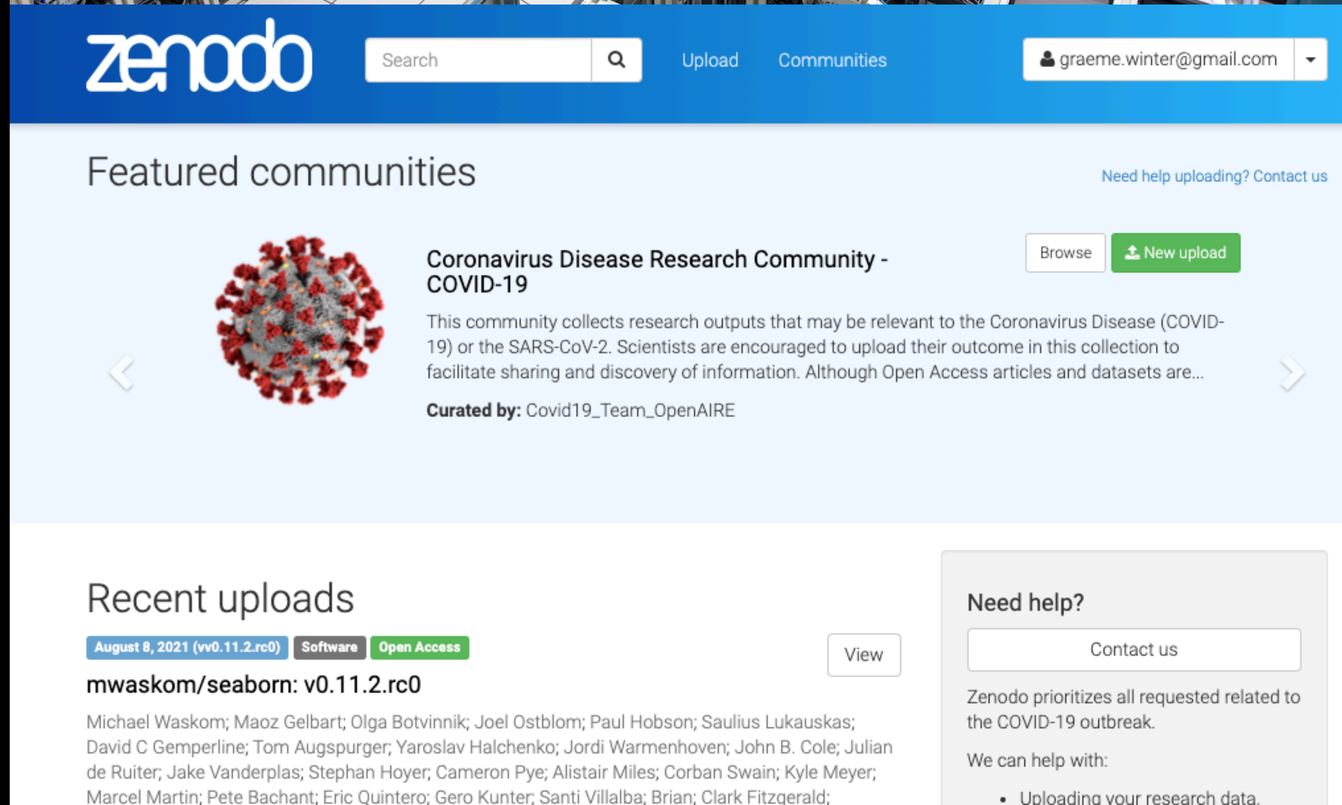
EXAMPLE - ICAT

- Designed for STFC facilities - Diamond, ISIS, …
- Strictly a data archive - no metadata, very limited search - but useful
- Data pulled off tape when needed, to staging or for download
- Archive goes back lifetime of Diamond



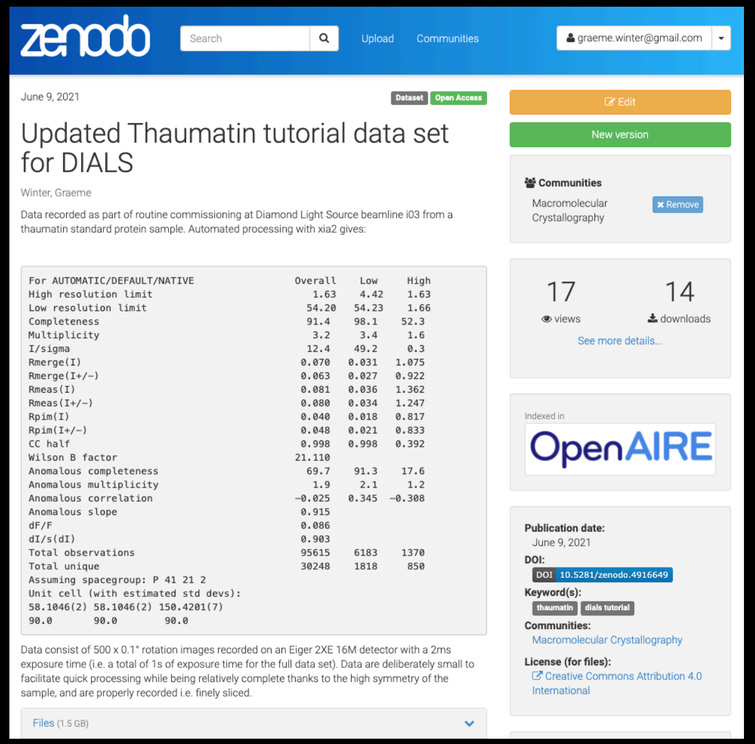
EXAMPLE: ZENODO?

- Funded through EU / openAIRE
- Intended to be free at the point of access for depositors and users
- Allows but does not mandate metadata
- Allows curation via communities
- Provides DOI, search facilities etc.
- Provides open API -> very useful

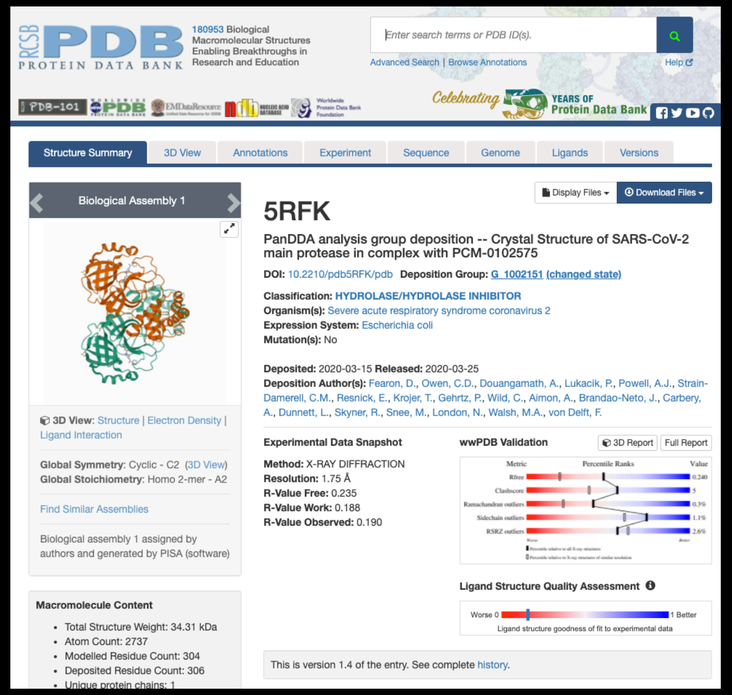


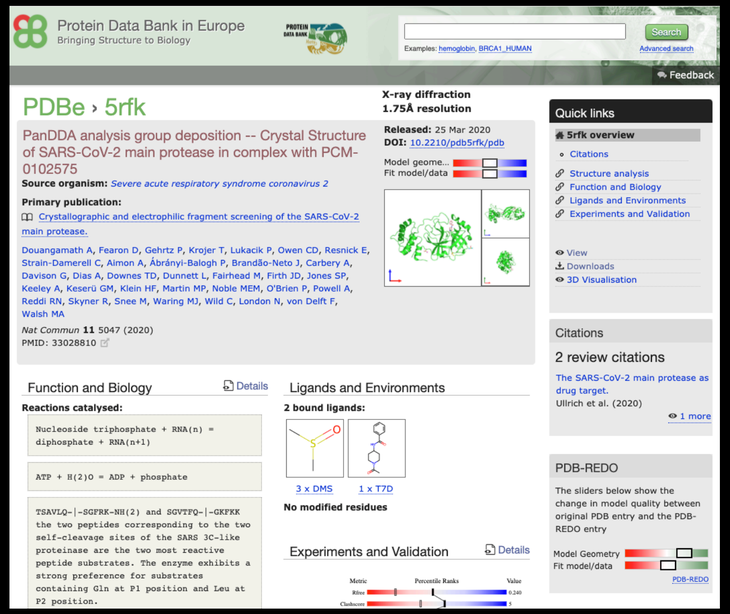
ZENODO FOR CRYSTALLOGRAPHY?

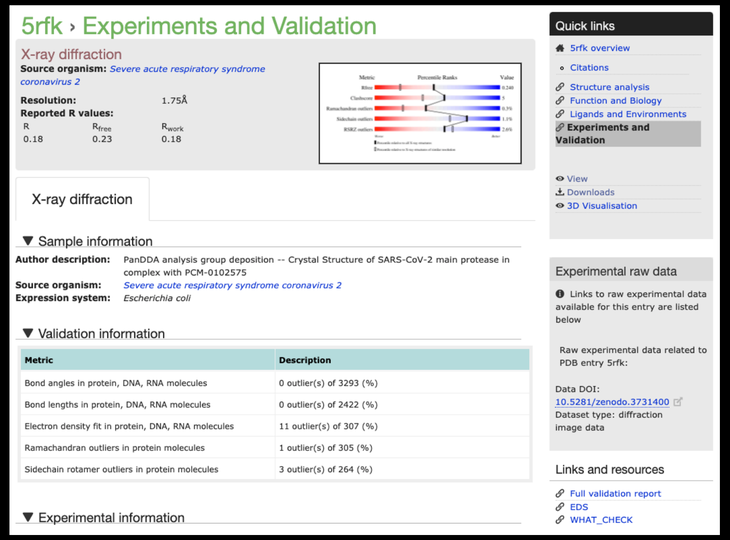
- General archive - so “mandatory data” does not include everything for e.g. CIF - but it could
- Not optimised for our use case - no scope for adding structured metadata
- If we started uploading 100,000 raw data sets / year someone would notice
- Great for “one off” type uploads

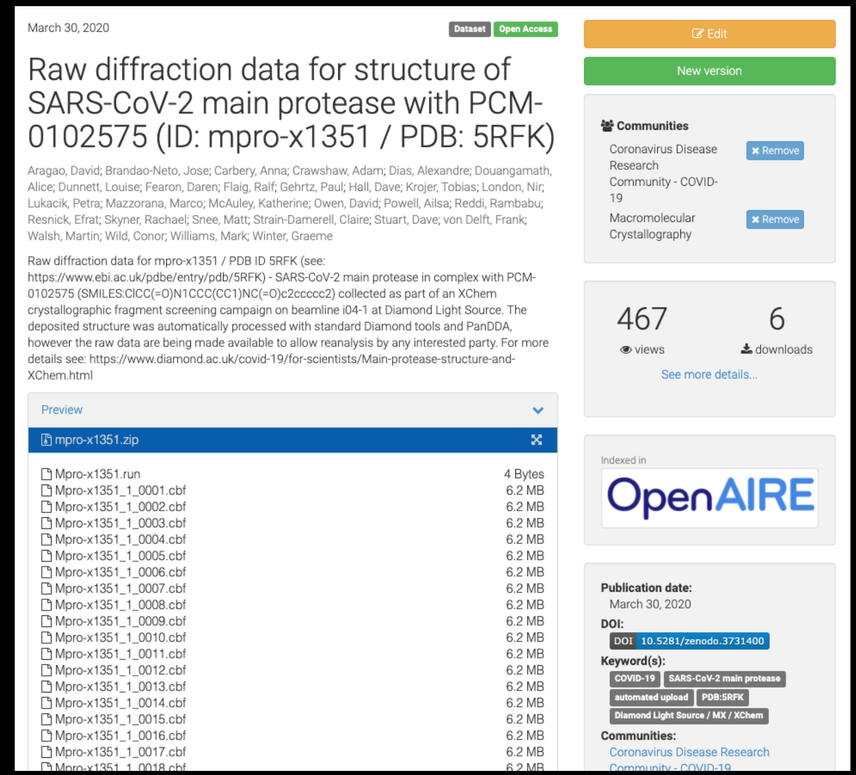


ZENODO FOR CRYSTALLOGRAPHY









ZENODO FOR CRYSTALLOGRAPHY

March 30, 2020

Dataset | Open Access

Raw diffraction data for structure of SARS-CoV-2 main protease with PCM-0102575 (ID: mpro-x1351 / PDB: 5RFK)

Aragao, David; Brandao-Neto, Jose; Carbery, Anna; Crawshaw, Adam; Dias, Alexandre; Douangamath, Alice; Dunnett, Louise; Fearon, Daren; Flaig, Ralf; Gehrtz, Paul; Hall, Dave; Krojer, Tobias; London, Nir; Lukacik, Petra; Mazzorana, Marco; McAuley, Katherine; Owen, David; Powell, Ailsa; Reddi, Rambabu; Resnick, Efrat; Skyner, Rachael; Snee, Matt; Strain-Damerell, Claire; Stuart, Dave; von Delft, Frank; Walsh, Martin; Wild, Conor; Williams, Mark; Winter, Graeme

Raw diffraction data for mpro-x1351 / PDB ID 5RFK (see: https://www.ebi.ac.uk/pdbe/entry/pdb/5RFK) - SARS-CoV-2 main protease in complex with PCM-0102575 (SMILES:ClCC(=O)N1CCC(CC1)NC(=O)c2ccccc2) collected as part of an XChem crystallographic fragment screening campaign on beamline i04-1 at Diamond Light Source. The deposited structure was automatically processed with standard Diamond tools and PanDDA, however the raw data are being made available to allow reanalysis by any interested party. For more details see: https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html

Preview

mpro-x1351.zip

Mpro-x1351.run	4 Bytes
Mpro-x1351_1_0001.cbf	6.2 MB
Mpro-x1351_1_0002.cbf	6.2 MB
Mpro-x1351_1_0003.cbf	6.2 MB
Mpro-x1351_1_0004.cbf	6.2 MB
Mpro-x1351_1_0005.cbf	6.2 MB
Mpro-x1351_1_0006.cbf	6.2 MB
Mpro-x1351_1_0007.cbf	6.2 MB
Mpro-x1351_1_0008.cbf	6.2 MB
Mpro-x1351_1_0009.cbf	6.2 MB
Mpro-x1351_1_0010.cbf	6.2 MB
Mpro-x1351_1_0011.cbf	6.2 MB
Mpro-x1351_1_0012.cbf	6.2 MB
Mpro-x1351_1_0013.cbf	6.2 MB
Mpro-x1351_1_0014.cbf	6.2 MB
Mpro-x1351_1_0015.cbf	6.2 MB
Mpro-x1351_1_0016.cbf	6.2 MB
Mpro-x1351_1_0017.cbf	6.2 MB
Mpro-x1351_1_0018.cbf	6.2 MB

Edit

New version

Communities

Coronavirus Disease Research Community - COVID-19 — Remove

Macromolecular Crystallography — Remove

467 views

6 downloads

See more details...

Indexed in OpenAIRE

Publication date: March 30, 2020

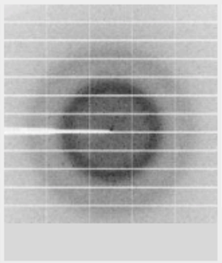
DOI: 10.5281/zenodo.3731400

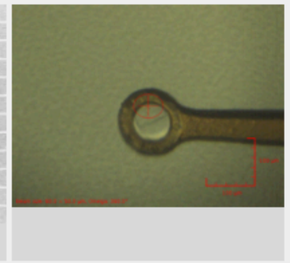
Keyword(s): COVID-19, SARS-CoV-2 main protease, automated upload, PDB:5RFK, Diamond Light Source / MX / XChem

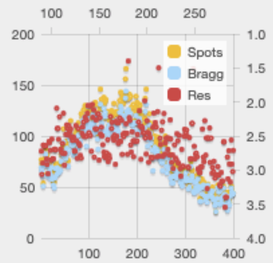
Communities: Coronavirus Disease Research Community - COVID-19

0 1 05-03-2020 20:53:46 - Mpro/Mpro-x1351/Mpro-x1351_1_####.cbf

Sample: Mpro-x1351	Flux: 3.50e+11
Ω Start: 90.0°	Ω Osc: 0.50°
Ω Overlap: 0°	No. Images: 400
Resolution: 1.80Å	Wavelength: 0.9119Å
Exposure: 0.040s	Transmission: 100.00%
Beamsize: 60x50µm	Type: SAD
Comment: (120006,13,184) Xray centring boxes: ['26.2s (8s)', 45, '19.3s (4s)', 21]. Aperture: 70um	







Auto Processing — xia2 dials: 2x ✔ | autoPROC: 2x ✔ | fast_dp: ✔ | xia2 3dii: 2x ✔ | autoPROC+STARANISO: 2x ✔

Type	Resolution	Spacegroup	Mn<I/sig(I)>	Rmeas Inner	Rmeas Outer	Completeness	Cell	Status
xia2 dials	54.96 - 1.75	C 1 2 1	4.9	0.071	2.039	99.7	126.95 52.77 111.24 90.00 159.66 90.00	processing successful
autoPROC	54.79 - 1.85	C 1 2 1	6.0	0.056	2.240	99.9	112.51 52.78 44.81 90.00 103.10 90.00	processing successful
fast_dp	29.98 - 2.06	C 1 2 1	8.1	0.043	1.223	96.4	112.32 52.68 44.74 90.00 103.12 90.00	processing successful
autoPROC	47.55 - 1.85	C 1 2 1	6.0	0.056	2.253	99.9	112.51 52.78 44.81 90.00 103.10 90.00	processing successful
xia2 3dii	52.73 - 1.88	P 1 21 1	3.2	0.074	3.541	99.8	44.77 52.73 111.15 90.00 99.98 90.00	processing successful
xia2 dials	54.95 - 1.75	C 1 2 1	5.0	0.071	2.036	99.7	112.43 52.76 44.79 90.00 103.08 90.00	processing successful
xia2 3dii	54.73 - 1.84	C 1 2 1	6.1	0.053	2.180	99.8	112.38 52.72 44.77 90.00 103.11 90.00	processing successful
autoPROC+STARANISO	47.55 - 1.75	C 1 2 1	7.2	0.055	1.042	91.5	112.51 52.78 44.81 90.00 103.10 90.00	processing successful
autoPROC+STARANISO	54.79 - 1.75	C 1 2 1	7.2	0.055	1.036	91.3	112.51 52.78 44.81 90.00 103.10 90.00	processing successful

xia2 dials | autoPROC | fast_dp | autoPROC | xia2 3dii | xia2 dials | xia2 3dii | autoPROC+STARANISO | autoPROC+STARANISO

Plots | Archive | Logs & Files | Lookup Cell

Beam Centre	X	Y
Start	212.35	251.34
Refined	212.71	251.60
Δ	-0.36	-0.26

Space Group	A	B	C	α	β	γ
C 1 2 1	126.95	52.77	111.24	90.00	159.66	90.00

Shell	Observations	Unique	Resolution	Rmeas	I/sig(I)	CC Half	Completeness	Multiplicity	Anom Completeness	Anom Multiplicity	CC Anom
outerShell	3995	1305	1.75 - 1.78	2.039	0.5	0.3	99.3	3.1	84.9	1.7	-0.0
innerShell	4933	1372	4.75 - 54.99	0.071	19.2	1.0	99.7	3.6	95.2	2.0	-0.5
overall	91960	26116	1.75 - 54.96	0.217	4.9	1.0	99.7	3.5	92.4	1.9	-0.2

Downstream Processing — dimple: 5x ✔

WHAT DO WE WANT?

- Ability to annotate raw data with processing output, full experiment metadata, sample material etc.
- Link to published structure - but not mandatory? Publishing *unsuccessful* data very interesting
- Validation to ensure that the data correspond to the claimed structure
- Facility to automate publication and update

5rfk redone

This information was created with PDB-REDO version 7.34. Please log in to request an update.

Crystallographic data

From PDB header

Spacegroup	C 1 2 1	*a:* 112.854 Å	*b:* 52.919 Å	*c:* 44.942 Å	*α:* 90.00°	*β:* 103.14°	*γ:* 90.00°
Resolution	1.75 Å	*Reflections*	25961	*Test set*	1294 (5.0%)		
R	0.1880	*R-free*	0.2350				

According to PDB-REDO

Resolution	1.75 Å	*Reflections*	25961	*Test set*	1294 (5.0%)	*Twin*	false

PDB-REDO files

Re-refined and rebuilt structure (PDB | mmCIF | MTZ)	Re-refined (only) structure (PDB | MTZ)	YASARA scenes (for visualisation of the results)	All files (compressed)

Links

PDBe	RCSB PDB	3D bionotes	Proteopedia

Validation metrics from PDB-REDO

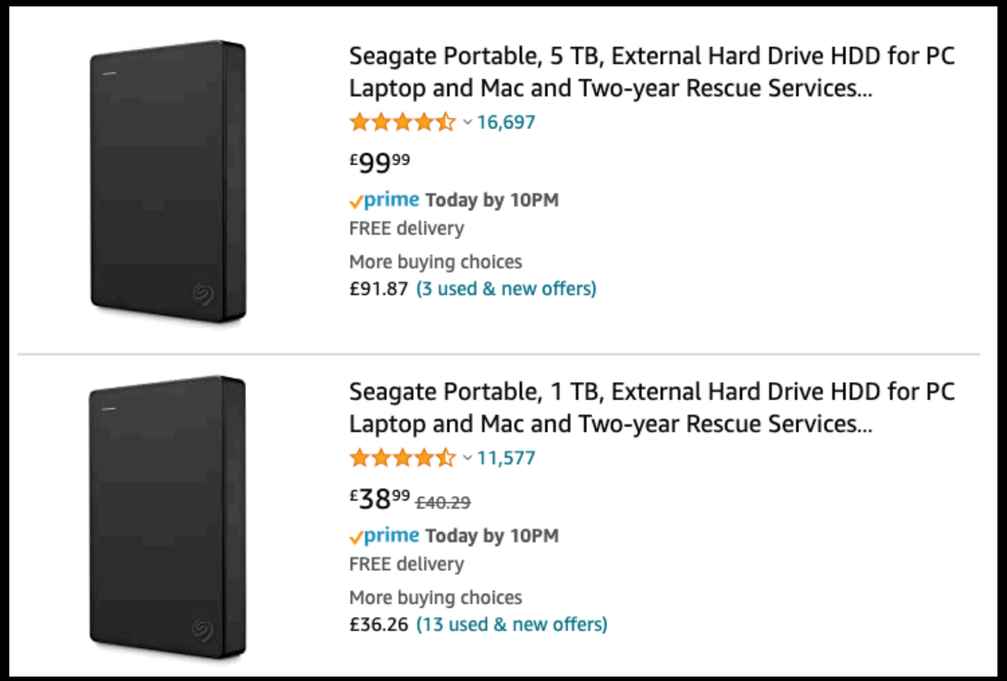
	PDB	PDB-REDO
Crystallographic refinement		
R	0.1970	0.2014
R-free	0.2400	0.2387
Bond length RMS Z-score	0.679	0.376
Bond angle RMS Z-score	1.025	0.653
Model quality (raw scores | percentiles)		
Ramachandran plot appearance	34	44
Rotamer normality	67	86
Coarse packing	34	32
Fine packing	19	23
Bump severity	82	92
Hydrogen bond satisfaction	30	38
WHAT_CHECK	Report	Report

CURATION

- Currently zenodo communities have “light touch” curation - largely done as a hobby by folks - but librarian is a vocation / job
- PDB, CSD etc. have professional curators and annotators - adding value to the raw data and the archive
- Critical to ensure the standards are defined
- Critical to ensure the standards are maintained
- Critical to ensure people are helped as users and depositors

COSTS

- Disks are cheap, small, portable
- Can buy one for every visit to DLS for a small cost compared with other consumables
- Obviously data not public, but can consider making public if someone asks…



COSTS

Data storage is expensive

- disks die
- technology changes - try finding a firewire port in 2021
- failure / accidents happen
- if you have not tried to read the data, assume the worst

COSTS

Data storage is expensive

- disks die
- technology changes - try finding a firewire port in 2021
- failure / accidents happen
- if you have not tried to read the data, assume the worst

REAL COSTS - STORAGE

- Azure as an example - probably priced in a realistic manner
- One “visit” / “shift” ~ 4TB
- £2.48 / month cheapest storage cost - £300 over 10 years
- 10 years time you’ll be paying to store 60 visits worth of data… and the data won’t be getting smaller

Data storage prices pay-as-you-go

All prices are per GB per month.

	Premium	Hot	Cool	Archive
First 50 terabyte (TB) / month	£0.11180 per GB	£0.0135 per GB	£0.00746 per GB	£0.00074 per GB
Next 450 TB/month	£0.11180 per GB	£0.0129 per GB	£0.00746 per GB	£0.00074 per GB
Over 500 TB/month	£0.11180 per GB	£0.0124 per GB	£0.00746 per GB	£0.00074 per GB

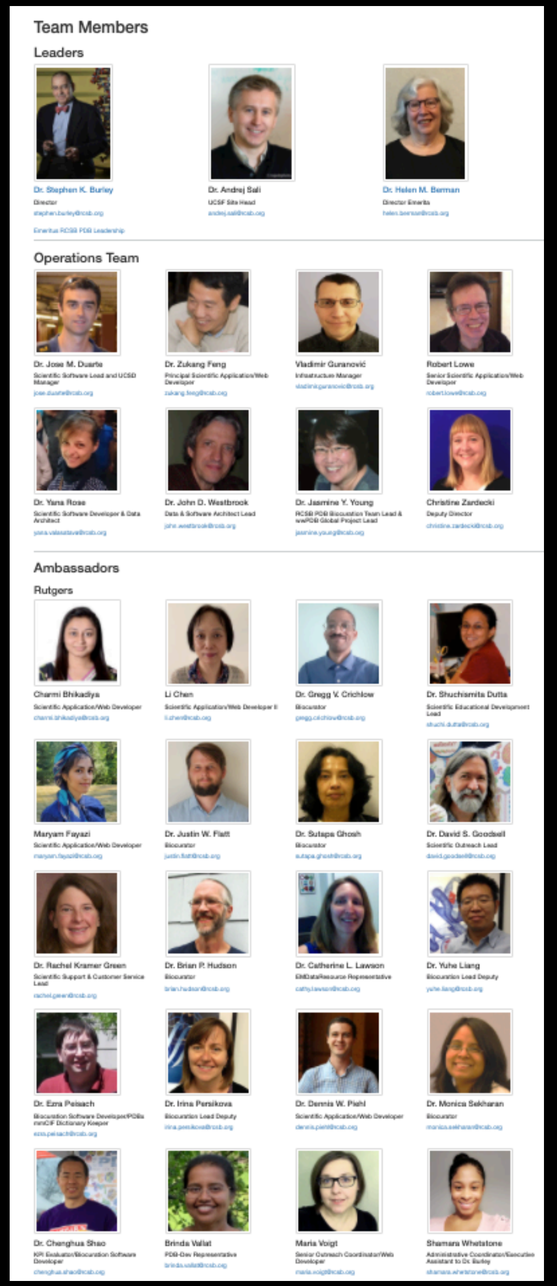
Azure Storage Reserved Capacity

Azure Storage Reserved Capacity helps you lower your data storage cost by committing to one year or three years of Azure Storage. Reserved capacity can be purchased in increments of 100 TB and 1 PB sizes for 1-year and 3-year commitment durations. All prices are per month. For more information, please see documentation.

	1-year reserved			3-year reserved		
	Hot	Cool	Archive	Hot	Cool	Archive
100 TB/month	£1,152	£626	£68	£927	£504	£62
1 PB/month	£11,217	£6,096	£658	£8,916	£4,846	£604

REAL COSTS - CURATION

- Different shape to storage - up front / one off rather than annual
- Partly amenable to automation - but still work to verify the data match the publication etc.
- Highly dependent on the goals of the curated data archive - the higher the value, the higher the staff costs

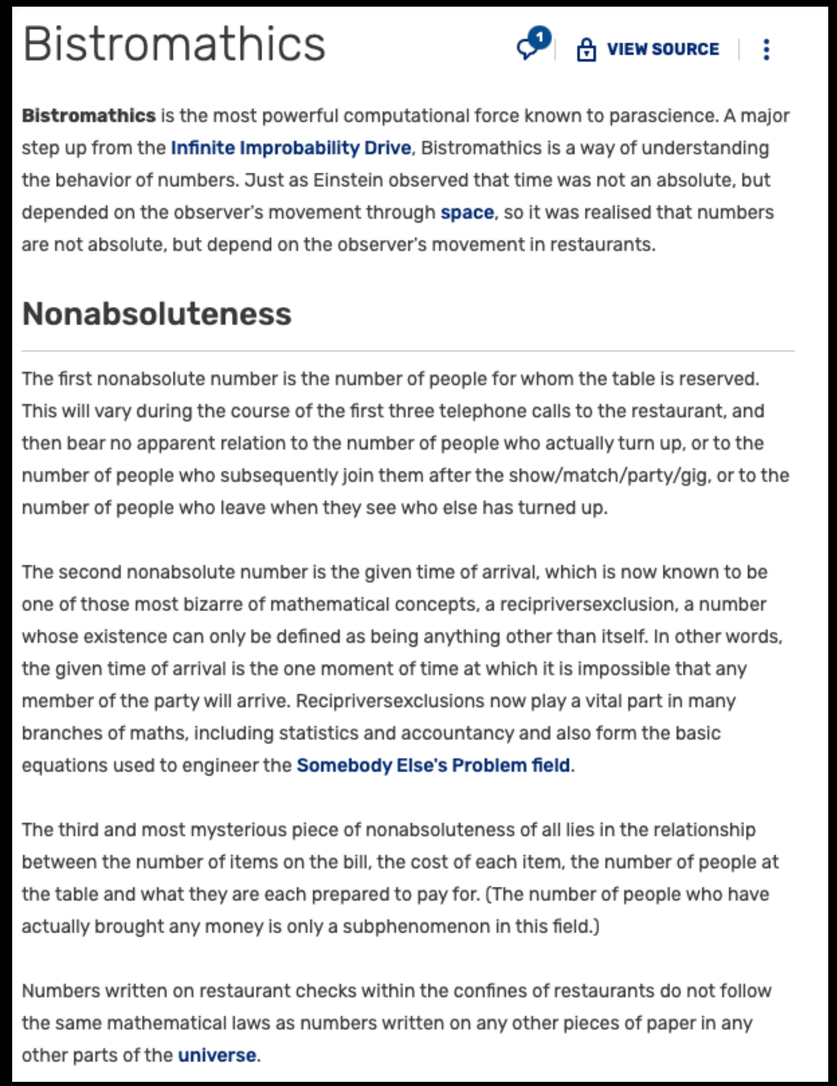


WORKED EXAMPLE

- 25 GB data set -> £2 to store for 10 years at cheapest rate
- Processing time to validate - 15 minutes on 16 core machine - £0.2 (low priority cloud resource)
- People cost to verify data - 5 minutes at £25 / hour -> £2
- Overall about £5 / data set (€6 / $7)
- Taking the data out will cost about £0.75 - £1.25 a go…

WHO PAYS?

- Scientist - reader - traditional manuscript model
- Creator - new “open access” model
- Facility (common in e.g. radio astronomy)
- 3rd party



Bistromathics

VIEW SOURCE

Bistromathics is the most powerful computational force known to parascience. A major step up from the **Infinite Improbability Drive**, Bistromathics is a way of understanding the behavior of numbers. Just as Einstein observed that time was not an absolute, but depended on the observer's movement through **space**, so it was realised that numbers are not absolute, but depend on the observer's movement in restaurants.

Nonabsoluteness

The first nonabsolute number is the number of people for whom the table is reserved. This will vary during the course of the first three telephone calls to the restaurant, and then bear no apparent relation to the number of people who actually turn up, or to the number of people who subsequently join them after the show/match/party/gig, or to the number of people who leave when they see who else has turned up.

The second nonabsolute number is the given time of arrival, which is now known to be one of those most bizarre of mathematical concepts, a recipriversexclusion, a number whose existence can only be defined as being anything other than itself. In other words, the given time of arrival is the one moment of time at which it is impossible that any member of the party will arrive. Recipriversexclusions now play a vital part in many branches of maths, including statistics and accountancy and also form the basic equations used to engineer the **Somebody Else's Problem field**.

The third and most mysterious piece of nonabsoluteness of all lies in the relationship between the number of items on the bill, the cost of each item, the number of people at the table and what they are each prepared to pay for. (The number of people who have actually brought any money is only a subphenomenon in this field.)

Numbers written on restaurant checks within the confines of restaurants do not follow the same mathematical laws as numbers written on any other pieces of paper in any other parts of the **universe**.

WHO PAYS? CHALLENGES

- Scientist - reader - traditional manuscript model - additional expense for hard pressed labs - also implies that publishers have control over your data (same as papers)
- Creator - new “open access” model - additional costs again to labs, though not impossible - advantage that it scales - but lab funding is transient
- Facility (common in e.g. radio astronomy) - very expensive as we don’t know what data will be important, also have to support many disciplines
- 3rd party - how are we going to persuade someone of the need?

HYBRID MODEL

- Data archive - facility / zenodo / azure (assumed to be reliable, may or may not provide DOI) - need not be specialised for crystallography
- Metadata archive - with the publication of the structure - has DOI - is curated and contains a reference back to the raw data (build into CSD / PDB) - see e.g. extensions to imgCIF to allow references to HDF5 raw data

SHOULD WE PAY?

- £6 / data set is / is not good value
- How much does it cost to reproduce the data?
- How much value will the data have? Will anyone ever look at it?

CONCLUSIONS

- Archiving raw experimental data perfectly possible - see Zenodo - easy even
- Defining a standard perfectly possible - see achievements in CIF / mmCIF / PDB etc. - making it part of publication process excellent way of encouraging people
- Deciding who should do the archiving is hard - and who should pay for it, how long the archive should live etc.
- Hybrid model of separating the data archiving from the metadata and curation more likely to meet the community need - just need to ensure link is bidirectional

ACKNOWLEDGEMENTS

- Diamond / STFC IT folks for keeping ICAT running and useful
- Diamond / STFC staff, users
- Commenters on Twitter for raising useful questions
- NeXus / imgCIF (& Herbert Bernstein) for standards definitions