Global Phasing Limited

A system for producing simple but comprehensive user summaries of large data collection campaigns executed by the GPhL Workflow

Rasmus Fogh
Global Phasing Ltd, Cambridge

ISPyB meeting, ALBA
29-30 November 2023

Global Phasing Limited

Making changes in ISPyB

- Long-standing problem of adding new data to ISPyB
  - Multi-sweep experiments cannot be shown
  - Anisotropic diffraction limits cannot be fitted in
  - Quality metrics fixed

- Roadblocks
  - The data model (SQL tables) takes immense effort to modify
    - Because of the tight coupling to very large bodies of code
  - Viewers cannot accommodate new data or experiment types
  - Data shown limited to ‘lowest common denominator’

Global Phasing Limited

Task description

- A user (Ashwin Chari, Max Planck Institute) had 1000+ workflow experiments to track (*now 3000+*), including home institution processing results So we *had* to address the limitations of ISPyB
- How to get an overview?
  - *You should have only one line per result*
- Complex multi-sweep experiments
  - *Additional information; organised per experiment, **not** per sweep, or per processing program*
- Details view for experiment length, dose, number and orientation of sweeps, …
- Applicable to already acquired experiments

Solutions

- Prototype: extract data from
  - existing GΦL workflow output files
  - associated autoPROC processing output
  - ISPyB only where there is no other source
    - ISPyB is only available on the parent synchrotron
- Future plans:
  - Save all relevant information in structured file while the workflow experiment runs
  - Combine to overview after the fact
  - Coordinate data model for exported file with ISPyB metadata??

GΦL
Global Phasing Limited

Overview spreadsheet

session	sampleId	strategy	variant	SG_in	n_sweeps	resolution	dose	energies	total_length°	final_spacegroup	diffraction_limits	path
20231106	PPIG-1-CP273A_xtl02_001	native	full	P222	n * 3	0.57			n * 816	No processing start	No results	.../PPIG-1-CP273A
20231106	PPIG-1-CP273A_xtl02_003	native	full	P222	n * 3	0.57			n * 816	No processing start	No results	.../PPIG-1-CP273A
20231106	PPIG-1-CP273A_xtl02_004	native	full	P222	n * 3	0.571			n * 816	No processing start	No results	.../PPIG-1-CP273A
20231106	PPIG-1-CP294_xtl08_001	phasing (SAD)	full	P222	5 * 5	0.65		23.4163	3100	P212121	[0.638, 0.655, 0.661]	.../PPIG-1-CP294_
20231106	PPIG-1-CR218A_xtl01_001	native	full	P222	n * 3	0.65			n * 816	No processing start	No results	.../PPIG-1-CR218A
20231106	PPIG-1-CR218A_xtl01_003	native	full	P222	3	0.65		23.4163	816	P212121	[0.637, 0.651, 0.665]	.../PPIG-1-CR218A
20231106	PPIG-1-CR218A_xtl05_001	native	full	P222	3	0.65		23.4163	816	P212121	[0.654, 0.666, 0.693]	.../PPIG-1-CR218A
20231106	PPIG-1-CR218A_xtl06_001	native	full	P222	3	0.65		23.4163	684	P212121	[0.681, 0.705, 0.705]	.../PPIG-1-CR218A
20231106	PPIG-1-CR218A_xtl07_001	native	full	P222	3	0.65		23.4163	684	P212121	[0.642, 0.676, 0.66]	.../PPIG-1-CR218A
20231106	PPIG-1-CR218A_xtl09_001	native	full	P222	3	0.65		23.4163	816	P212121	[0.664, 0.681, 0.698]	.../PPIG-1-CR218A
20231106	PPIG-1-CR235A_xrtl010_001	phasing (SAD)	full	P222	5 * 2	0.65		23.4163	2700	P222	[0.636, 0.64, 0.661]	.../PPIG-1-CR235A
20231106	PPIG-1-CR235A_xrtl01_001	phasing (MAD)	ultralong	P222	2 * 5	1.204		[12.6693, 12.6593]	1232	P222	[0.989, 1.0, 0.999]	.../PPIG-1-CR235A
20231106	PPIG-1-CR235A_xrtl02_001	phasing (MAD)	ultralong	P222	10 * 2	1.204		[12.6693, 12.6593]	3600	P21212	[1.021, 1.013, 1.005]	.../PPIG-1-CR235A
20231106	PPIG-1-CR235A_xrtl03_001	phasing (MAD)	ultralong	P222	10 * 2	1.204		[12.6593, 12.6693]	3600	No processing start	No results	.../PPIG-1-CR235A
20231106	PPIG-1-CR235A_xrtl04_004	phasing (SAD)	full	P222	5 * 2	0.85		18	2700	P212121	[0.754, 0.741, 0.744]	.../PPIG-1-CR235A
20231106	PPIG-1-CR235A_xrtl04_006	native	full	P222	n * 3	0.8			n * 934	No processing start	No results	.../PPIG-1-CR235A
20231106	PPIG-1-CR235A_xrtl04_007	phasing (SAD)	full	P222	5 * 2	0.8		26.6797	2700	P212121	[0.759, 0.733, 0.733]	.../PPIG-1-CR235A
20231106	PPIG-1-CR235A_xrtl07_001	phasing (SAD)	full	P222	5 * 5	0.65		23.4163	3100	No processing start	No results	.../PPIG-1-CR235A
20231106	PPIG-1-CR236A_xtl05_002	native	full	P222	3	0.75		23.4164	802	P212121	[0.739, 0.738, 0.757]	.../PPIG-1-CR236A
20231106	PPIG-1-CR236A_xtl07_001	native	full	P222	3	0.75		26.6797	648	P212121	[0.744, 0.731, 0.762]	.../PPIG-1-CR236A
20231106	PPIG-1-CR301A_xtl03_001	native	full	P222	3	0.75		23.4163	802	P212121	[0.752, 0.782, 0.751]	.../PPIG-1-CR301A
20231106	PPIG-1-CR301A_xtl06_001	native	full	P222	3	0.75		23.4163	941	P212121	[0.816, 0.81, 0.806]	.../PPIG-1-CR301A
20231106	Vhaas2-CP293A_xtl01_001	native	full	C2	4	2.6		23.4162	936	C2	[2.707, 2.425, 2.339]	.../Vhaas2-CP293A
20231106	Vhaas2-CP293A_xtl02_001	native	full	C2	4	2.5		23.4162	936	C2	[3.09, 2.669, 2.569]	.../Vhaas2-CP293A

Details file - parameters

```
Parameters:
===========
Session:                  20231106              Sample ID:               PPIG-1-CR235A_xrtl01_001
Run number:               2                     File prefix:             PPIG-1-CR235A_xrtl01_G1B1
Strategy:                 phasing (MAD)         Variant:                 ultralong
Input spacegroup:         P222                  First wavelength (Å):    0.979
Detector distance (mm):   139.7                 Resolution (Å):          1.204
Sweep count:               2 *  5               Total length (°):        2 * 616.0
Image width (°):          0.1                   Exposure time (s):       0.012404
Radiation Sensitivity:    Missing               Dose Budget (MGy):       4.019
Transmission (%):         Missing               Acquisition dose (MGy):Missing
Beam position (pixels):   [2068.32, 2186.36]    Flux (photons/s):        Missing
Beam Size (mm):           Missing               Beam Setting:            Missing
energies:                 [12.6693, 12.6593]
path:                     …/20231106/PROCESSED_DATA/GPhL_WF/PPIG-1-CR235A_xrtl01_001
```

Details file – individual sweeps

```
Sweeps (for each wavelength)
============================
  180°, ω= -55.1°, κ=  21.7°, φ=-118.9°, on-axis
   57°, ω=  70.3°, κ= 180.3°, φ= -19.1°, unaligned
   19°, ω=  51.3°, κ= 180.3°, φ= -19.1°, unaligned
  180°, ω=-128.7°, κ= 180.3°, φ= -19.1°, unaligned
  180°, ω= -49.5°, κ=   9.3°, φ=-113.2°, off-axis
```

Details file – Table 1

```
Table1:
=======
Spacegroup name              P222
Unit cell parameters         37.5041 65.4832 69.5195 90.0 90.0 90.0
Wavelength                   0.97862 A

Diffraction limits & principal axes of ellipsoid fitted to diffraction cut-off surface:
    0.989         1.0000   0.0000   0.0000       _a_*
    1.000         0.0000   1.0000   0.0000       _b_*
    0.999         0.0000   0.0000   1.0000       _c_*

  Number of active ice-rings within this resolution range = 15
  Number of RUNs (sweeps) contributing to this dataset =   8

Criteria used in determination of diffraction limits:
  -----------------------------------------------------
     local(I/sigI)  >=    1.20

Per-reflection cut-off      Operational Resolution
-----------------------------------------------------------------
     I/sigma(I) >= 2.0     :   1.0582 A   for      78855 reflections
     I/sigma(I) >= 1.0     :   1.0480 A   for      81086 reflections
     I/sigma(I) >= 0.0     :   1.0403 A   for      82859 reflections
     all                   :   1.0396 A   for      83073 reflections
```

Global Phasing Limited

Details file – Table 1 (contd.)

	Overall	InnerShell	OuterShell
Low resolution limit	37.504	37.504	1.048
High resolution limit	1.006	2.821	1.006
Rmerge (all I+ & I-)	0.058	0.041	0.168
Rmerge (within I+/I-)	0.041	0.030	0.141
Rmeas (all I+ & I-)	0.060	0.042	0.191
Rmeas (within I+/I-)	0.043	0.032	0.170
Rpim (all I+ & I-)	0.013	0.009	0.086
Rpim (within I+/I-)	0.013	0.009	0.091
Total number of observations	1465136	86622	16900
Total number unique	83073	4154	4154
Mean(I)/sd(I)	42.4	96.7	5.2
Completeness (spherical)	90.7	93.1	39.8
Completeness (ellipsoidal)	90.7	93.1	39.8
Multiplicity	17.6	20.9	4.1
CC(1/2)	1.000	0.999	0.970
Anomalous completeness (spherical)	89.2	91.3	35.2
Anomalous completeness (ellipsoidal)	89.2	91.3	35.2
Anomalous multiplicity	9.3	12.1	2.3
CC(ano)	0.865	0.814	0.470
|DANO|/sd(DANO)	3.131	4.546	0.666

Acknowledgements

- **Global Phasing colleagues**
  - Peter Keller
  - Rasmus Fogh
  - Wlodek Paciorek
  - Claus Flensburg
  - Clemens Vonrhein
  - Andrew Sharff
  - Ian Tickle
  - Gerard Bricogne
- EMBL-Hamburg / PETRA III (P14)
  - Gleb Bourenkov
- Max Planck Institute, Göttingen
  - Ashwin Chari
- The Global Phasing Consortium
  - Funding, feed-back, and much more

Global Phasing Limited

- Make_summaries program
- **ISPyB discussion**

Global Phasing Limited

Contribution to discussion

- After useful discussion with Alex de Maria and others, my (many) reservations have been answered. So:
-
-
-

Global Phasing Limited

Contribution to discussion

- After useful discussion with Alex de Maria and others, my (many) reservations have been answered. So:
- I think ICAT could be an excellent basis for a new ISPyB
-
-

Global Phasing Limited

Contribution to discussion

- After useful discussion with Alex de Maria and others, my (many) reservations have been answered. So:
- I think ICAT could be an excellent basis for a new ISPyB
- - but it will require significant work from all of us to get there
-

Contribution to discussion

- After useful discussion with Alex de Maria and others, my (many) reservations have been answered. So:
- I think ICAT could be an excellent basis for a new ISPyB
- - but it will require significant work form all of us to get there
- Some points to consider:

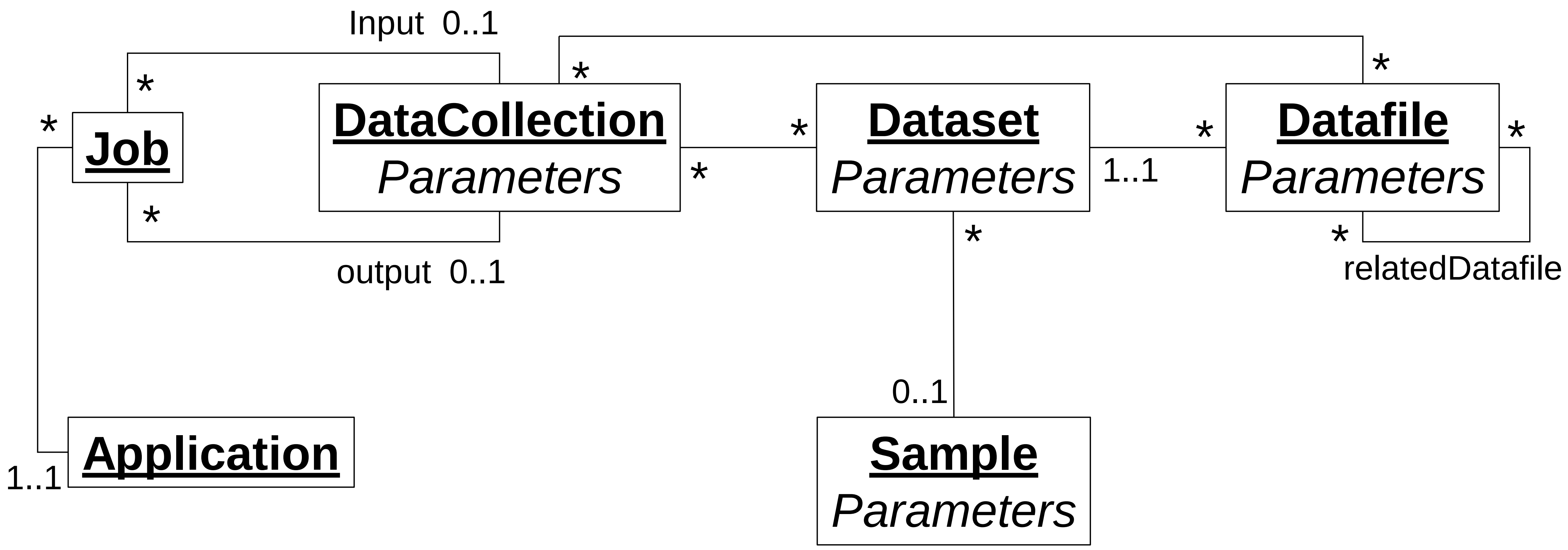
Global Phasing Limited

Scientific data handling

- Detailed data are put as parameters to ICAT objects
- We must have a *Data model* to define the scientific data to store and retrieve
- We must give each Dataset/DataCollection a *type* to emulate a more detailed model, so we can distinguish e.g. EM Datasets from MX datasets
  - Only MX Datasets can have MX type parameters
  - Only MX Datasets can be used for MX calculations
  - Can we have structured (JSON) data in parameters?

Global Phasing Limited

ICAT Data Model



Global Phasing Limited

Organise viewers by DataCollection

- Current prototype ICAT viewer was organized by Sample and Dataset
  - A Dataset is a single sweep on a single sample
  - This can only work well for single-sweep experiments

- The natural ICAT organization unit is the **DataCollection**
  - a Job input or result is a DataCollection
  - easy fit for multi-sweep experiments,
    or results combining different experiments
  - you can have multiple experiments per crystal/sample
  - you can make DataCollections to combine (only) relevant Dataset(s)

Scope for site-specific coding

- In data model separate
  - Core metadata - global agreement and definitions
  - Site/Program-specific data – separate namespace and locally managed

- Allow for customization of views, either at program level or even by user at runtime, to cater for different needs

Global Phasing Limited

Latest from MXCuBE

- The MXCuBE developers have agreed on the need for an abstract LIMS
  - There are now MXCuBE members who do not use ISPyB
  - This requires an interface for how to transfer data in and out to LIMS

- The main part of that work will be agreeing on the nature and structure of the data – a '*metadata model*'

- There is obvious scope for coordinating with others who need to model these same data
  - But MXCuBE has its own needs (and timings) independent of other actors