Programming pdCIF and Rietveld:

Brian H. Toby
NIST Center for Neutron Research

Talk Outline

- *Motivation/Goals*
- *Data Grammars vs. Data Languages*
- *CIF*
- *Informatics and the next generation of Rietveld software*
- *Programming writing & reading CIFs*

Motivation for Standardized Data Formats

Back in the dark ages of crystallography every program used its own data format.

– Electronic data communication was unusual and even then by sneakernet.

Even then many crystallographers felt there must be a better way…

Modern goals:

- Direct communication between instruments and data analysis tools
- Interoperability between programs
- Electronic communication of results
- Facile publication

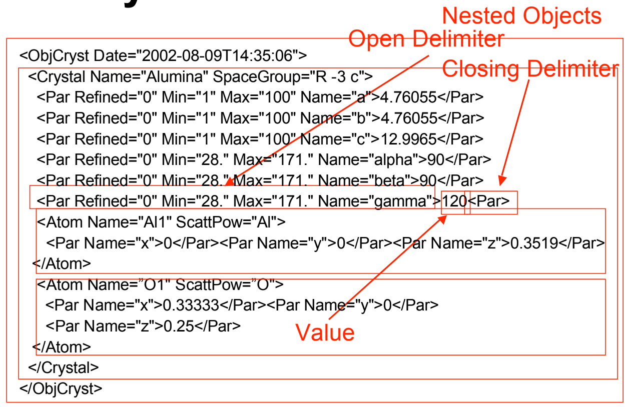
Productivity increases when computers function for scientists rather than the other way round!

Data Grammars vs. Data Languages

A Data Grammar specifies how information will be formatted so that a computer program can interpret it.

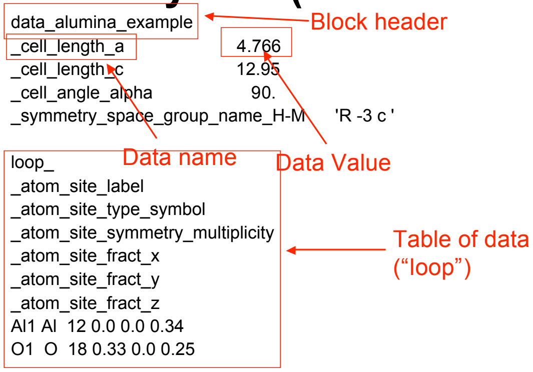
– JCAMP-DX
– Spreadsheet (.csv)
– HDF
– STAR
– XML

XML syntax



```
<ObjCryst Date="2002-08-09T14:35:06">
 <Crystal Name="Alumina" SpaceGroup="R -3 c">
  <Par Refined="0" Min="1" Max="100" Name="a">4.76055</Par>
  <Par Refined="0" Min="1" Max="100" Name="b">4.76055</Par>
  <Par Refined="0" Min="1" Max="100" Name="c">12.9965</Par>
  <Par Refined="0" Min="28." Max="171." Name="alpha">90</Par>
  <Par Refined="0" Min="28." Max="171." Name="beta">90</Par>
  <Par Refined="0" Min="28." Max="171." Name="gamma">120<Par>
  <Atom Name="Al1" ScattPow="Al">
   <Par Name="x">0</Par><Par Name="y">0</Par><Par Name="z">0.3519</Par>
  </Atom>
  <Atom Name="O1" ScattPow="O">
   <Par Name="x">0.33333</Par><Par Name="y">0</Par>
   <Par Name="z">0.25</Par>
  </Atom>
 </Crystal>
</ObjCryst>
```

STAR syntax (used in CIF)



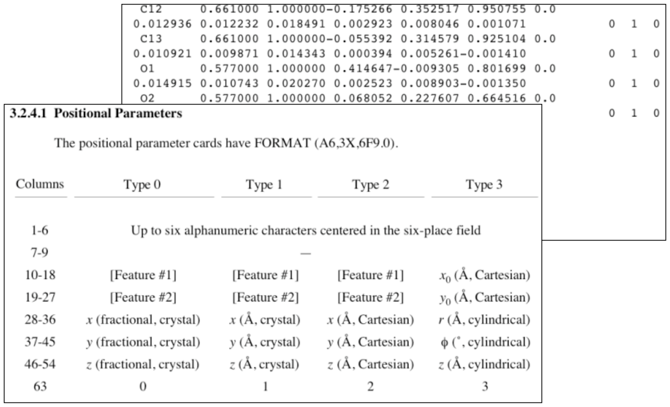
```
data_alumina_example
_cell_length_a             4.766
_cell_length_c             12.95
_cell_angle_alpha           90.
_symmetry_space_group_name_H-M     'R -3 c '

loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_symmetry_multiplicity
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
Al1 Al  12 0.0 0.0 0.34
O1  O  18 0.33 0.0 0.25
```

Data Language:

- Built on a data grammar (usually)
- Provides rigorous definitions for each data value
- Establishes validation information (usually)
- Defines logical relationships between data items (optional)

CIF: first comprehensive & interoperable data language for crystallography

Data languages are not new



```
 C12      0.661000 1.000000-0.175266 0.352517 0.950755 0.0
0.012936 0.012232 0.018491 0.002923 0.008046 0.001071          0  1  0
 C13      0.661000 1.000000-0.055392 0.314579 0.925104 0.0
0.010921 0.009871 0.014343 0.000394 0.005261-0.001410          0  1  0
 O1       0.577000 1.000000 0.414647-0.009305 0.801699 0.0
0.014915 0.010743 0.020270 0.002523 0.008903-0.001350          0  1  0
 O2       0.577000 1.000000 0.068052 0.227607 0.664516 0.0
                                                               0  1  0
```

3.2.4.1 Positional Parameters

The positional parameter cards have FORMAT (A6,3X,6F9.0).

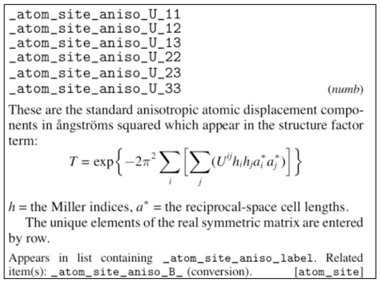
Columns	Type 0	Type 1	Type 2	Type 3
1-6	Up to six alphanumeric characters centered in the six-place field			
7-9	—			
10-18	[Feature #1]	[Feature #1]	[Feature #1]	x_0 (Å, Cartesian)
19-27	[Feature #2]	[Feature #2]	[Feature #2]	y_0 (Å, Cartesian)
28-36	x (fractional, crystal)	x (Å, crystal)	x (Å, Cartesian)	r (Å, cylindrical)
37-45	y (fractional, crystal)	y (Å, crystal)	y (Å, Cartesian)	ϕ (°, cylindrical)
46-54	z (fractional, crystal)	z (Å, crystal)	z (Å, Cartesian)	z (Å, cylindrical)
63	0	1	2	3

What's so special about CIF?

Each data item in CIF is defined in a computer-readable dictionary

>3,000 defined terms (250+ pages in Int. Tabl.)

uses subset of STAR data grammar



`_atom_site_aniso_U_11`
`_atom_site_aniso_U_12`
`_atom_site_aniso_U_13`
`_atom_site_aniso_U_22`
`_atom_site_aniso_U_23`
`_atom_site_aniso_U_33` *(numb)*

These are the standard anisotropic atomic displacement components in ångströms squared which appear in the structure factor term:

$$T = \exp\left\{-2\pi^2 \sum_i \left[\sum_j (U^{ij} h_i h_j a_i^* a_j^*)\right]\right\}$$

h = the Miller indices, a^* = the reciprocal-space cell lengths.

The unique elements of the real symmetric matrix are entered by row.

Appears in list containing `_atom_site_aniso_label`. Related item(s): `_atom_site_aniso_B_` (conversion). [atom_site]

>20 years of development effort (adopted by IUCr in 1990.)

CIF has redefined small molecule publishing

CIF is the uncontested standard for communication of structure factors & crystal structure results

- Reduces errors in print: Journals can use structure validation software
- Bond distance & angle tables are generated directly from the CIF
- Required for IUCr Journals (*Acta Cryst.*, etc.)

Impact on mm (via PDB) is probably even larger

How does CIF work: CIF Syntax

```
_cell_length_a        4.766
_cell_length_c        12.95
_cell_angle_alpha       90.
```

- Data names (tags) & values
- loop_: links sets of data names & sets of values (Tables)
- Dictionary specifies:
  - Definition
  - Rules on allowed values
  - Category
    - All data names in loop must be in same c
  - Loop rules
    - Specifies which data items can/must/cannot be looped
    - Specifies logical connections between loops
    - Specifies a unique item for each loop

```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
Al1 Al  0.0 0.0 0.34
O1  O  0.33 0.0 0.25
```

CIF Dictionaries

CIF definitions are developed by teams with widespread interests

- Core -- fundamental & single-crystal terms
- mmCIF -- macromolecular
- pdCIF -- powder diffraction
- msCIF -- modulated structures
- imgCIF -- 2D images
- symCIF -- symmetry
- rhoCIF -- electron density
- diffCIF -- diffuse scattering (in progress)

CIF Information Sources

- Formal specifications, see:
  http://www.iucr.org/iucr-top/cif/
- Also see templates & examples on Acta Cryst. Author's Guides
- International Tables Vol. G (today in Florence!)
- Developer's discussion list
  http://www.iucr.org/iucr-top/lists/cif-developers/

Editorial comment:

pdCIF is far more important than a mechanism for data interchange & review.

- CIF has the potential to be the cornerstone of the next generation of Rietveld analysis software

The two dialects of CIF Dictionaries

- In the course of defining CIF dictionaries, the mmCIF designers wanted more database structure than required for CIF initially.
  - Created a dictionary for defining dictionaries:
    - DDL (data definition language)
  - After inflicting many database structures into DDL v1.x, the mmCIF was written using DDL2.x
  - Programs that read dictionaries need to be aware of DDL1 vs DDL2 differences,
  - programs that only read/write CIFs do not.
  - Discussions on merging DDLs are underway

CIF for Powder Diffraction (pdCIF)

Universal data format for powder diffraction

Goals beyond those of CIF:

- Accommodate all types of powder data
  - Flexibility in conflict with mmCIF
- Document experimental geometry, conditions
- Record "raw" and processed diffraction data & Rietveld fits

The future of "Rietveld"

- Problems increasingly are more complex than can be accommodated with powder data at any resolution
  - Need incorporate many additional types of observations and constraints
- Characterize non-periodic aspects of crystal structures
  - Local order; stacking faults; defects

The Next Generation of Data Fitting

- Codes should be modular, glued by a scripting language: customization
- Data modules can compute contributions to design matrix & least-squares vector against data & restraints
- Hard constraint modules can reduce parameters using GSAS or Finger-Prince approach
- Minimizer modules can develop & apply shifts from Hessian
- Cost function modules can keep parameters in bounds by adding to design matrix or Hessian

CIF is the control file for next generation data fitting

CIF defines the basis for state-of-art refinement data objects

Where models cannot be described, CIF must expand

- Data item descriptions are rigorous so that structure factors can be computed directly from the CIF
  - Modulated structures
  - Need more defect model descriptions
- Powder data can be simulated to match data items found in CIF
- With CIF additions PDF fitting becomes straight-forward

CIF without STAR?

- CIF contains 20 years of informatics design efforts
- CIF is poor for large data structures
  - HDF is a portable data grammar for large data volumes
  - NeXus (HDF for scattering) not yet a complete data language
- XML is state-of-art ASCII data grammar
  - Again not a data language

CIF definitions can be transferred to other data grammars

- Efforts to pair XML and CIF are underway (*c.f.* IUCr Florence, 2005)
- A marriage between CIF & NeXus would benefit everyone

Programming with CIF: Resources

- International Tables Volume G
- Open-source CIF parsers (see www.iucr.org/iucr-top/cif/software/):
  - CIFtbx 3.0 [Fortran]
  - Rutgers mmCIF lib [C]
  - CBFlib (used in RasMol) [C?]
    http://www.bernstein-plus-sons.com/software/CBF
  - PyCifRW [Python]
    www.ansto.gov.au/natfac/ANBF/CIF/
  - CIFIO in XTAL pkg [RATMAC=FORTRAN]
    xtal.sf.net

Personal Experiences in CIF Programming

- GSAS2CIF
  - Exports GSAS refinements in CIF
  - FillTemplate
    - Enter information into CIF templates (EXPGUI)
  - CIFSelect
    - Set [don't] publish flag in bond distances (EXPGUI)
- pdCIFplot
  - Plots Rietveld fits from CIF
- CIFEDIT
  - CIF validation & editor
- CMPR

GSAS2CIF: Challenges

- Potentially complex data structures:
  - N (≤9) phases
  - M (≤99) data sets
  - NxM+1 CIF Blocks (or 1 if N=M=1)
- Reuse of author-entered information (metadata)
- Avoid use of CIF parser

Toby, B. H., Von Dreele, R. B, and Larson, A. C., "Reporting of Rietveld Results Using pdCIF: GSAS2CIF", *J. Appl. Cryst.* **36**, 1290 (2003)

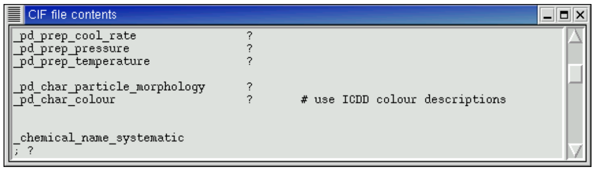
http://www.ncnr.nist.gov/xtal/software/expgui/gsas2cif.html

GSAS2CIF: Solutions

- Divide *Acta Cryst.* template into sections
  1. Publication info
  2. Sample/characterization info *(need N copies)*
  3. Instrument/data collection info *(need M copies)*
  - Remove parameters “known” to GSAS
- CIF is generated by “quilting” together template sections with fit results
- Author-entered info (metadata) goes into template sections not into “final” CIF
  - Quick regeneration of new “final CIF”
  - Sharing of template sections between projects

GSAS2CIF GUI Tools: FillTemplate

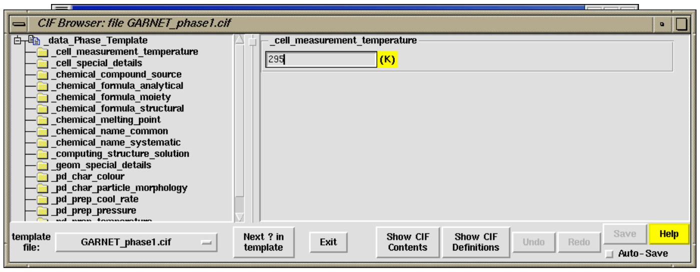
Author must supply
metadata -- entered
into template



```
_pd_prep_cool_rate               ?
_pd_prep_pressure                ?
_pd_prep_temperature             ?

_pd_char_particle_morphology     ?
_pd_char_colour                  ?       # use ICDD colour descriptions


_chemical_name_systematic
; ?
```

GSAS2CIF GUI Tools: FillTemplate

Author must supply
metadata -- entered
into template

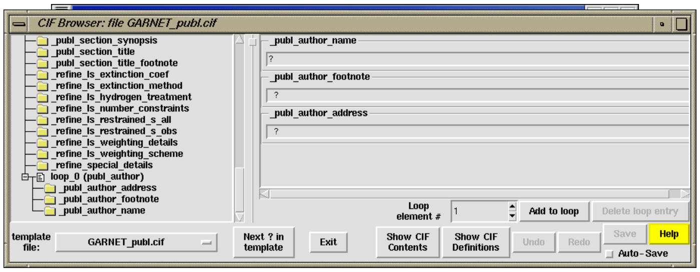


_cell_measurement_temperature: 295 (K)

template file: GARNET_phase1.cif

GSAS2CIF GUI Tools: FillTemplate

Author must supply
metadata -- entered
into template


CIF Browser: file GARNET_publ.cif
_publ_section_synopsis
_publ_section_title
_publ_section_title_footnote
_refine_ls_extinction_coef
_refine_ls_extinction_method
_refine_ls_hydrogen_treatment
_refine_ls_number_constraints
_refine_ls_restrained_s_all
_refine_ls_restrained_s_obs
_refine_ls_weighting_details
_refine_ls_weighting_scheme
_refine_special_details
loop_0 (publ_author)
_publ_author_address
_publ_author_footnote
_publ_author_name
Add to loop
Delete loop entry
Next ? in template
Exit
Show CIF Contents
Show CIF Definitions
Undo
Redo
Save
Help
Auto-Save

_publ_author_name: ?

_publ_author_footnote: ?

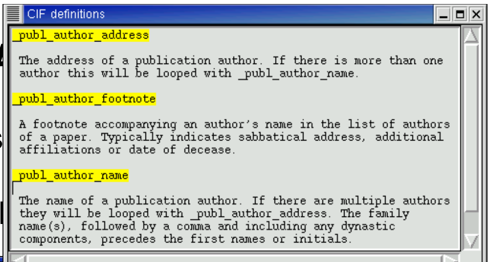
_publ_author_address: ?

Loop element #: 1

template file: GARNET_publ.cif

GSA

Author mus
metadata
into templ



_publ_author_address

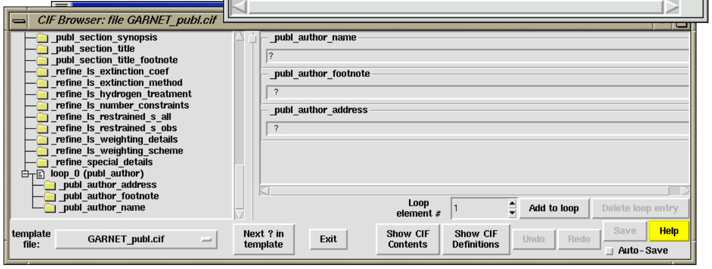
The address of a publication author. If there is more than one author this will be looped with _publ_author_name.

_publ_author_footnote

A footnote accompanying an author’s name in the list of authors of a paper. Typically indicates sabbatical address, additional affiliations or date of decease.

_publ_author_name

The name of a publication author. If there are multiple authors they will be looped with _publ_author_address. The family name(s), followed by a comma and including any dynastic components, precedes the first names or initials.


CIF Browser: file GARNET_publ.cif
_publ_section_synopsis
_publ_section_title
_publ_section_title_footnote
_refine_ls_extinction_coef
_refine_ls_extinction_method
_refine_ls_hydrogen_treatment
_refine_ls_number_constraints
_refine_ls_restrained_s_all
_refine_ls_restrained_s_obs
_refine_ls_weighting_details
_refine_ls_weighting_scheme
_refine_special_details
loop_0 (publ_author)
_publ_author_address
_publ_author_footnote
_publ_author_name
Add to loop
Delete loop entry
Next ? in template
Exit
Show CIF Contents
Show CIF Definitions
Undo
Redo
Save
Help
Auto-Save

_publ_author_name: ?

_publ_author_footnote: ?

_publ_author_address: ?

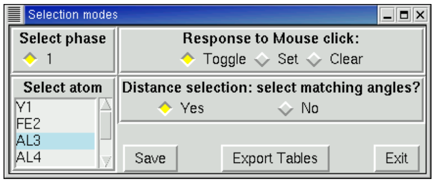
Loop element #: 1

template file: GARNET_publ.cif

GSAS2CIF GUI Tools: CIFSelect

Select distances & angles for publication

– Keep flags in separate file



Select phase: ● 1

Response to Mouse click: ● Toggle ○ Set ○ Clear

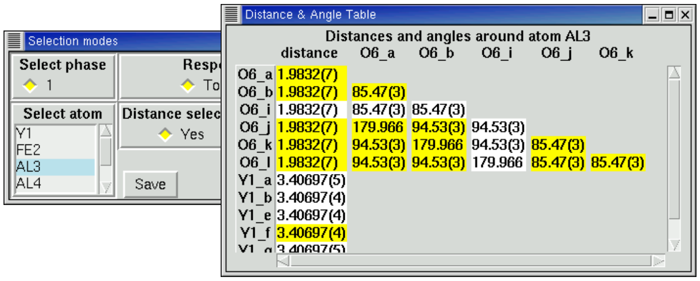
Select atom

Distance selection: select matching angles? ● Yes ○ No

GSAS2CIF GUI Tools: CIFSelect

Select distances & angles for publication

- Keep flags in separate file



Selection modes

Select phase ◆ 1

Resp… ◆ To…

Select atom: Y1, FE2, AL3, AL4

Distance selec… ◆ Yes

Save

Distance & Angle Table

Distances and angles around atom AL3

	distance	O6_a	O6_b	O6_i	O6_j	O6_k
O6_a	1.9832(7)					
O6_b	1.9832(7)	85.47(3)				
O6_i	1.9832(7)	85.47(3)	85.47(3)			
O6_j	1.9832(7)	179.966	94.53(3)	94.53(3)		
O6_k	1.9832(7)	94.53(3)	179.966	94.53(3)	85.47(3)	
O6_l	1.9832(7)	94.53(3)	94.53(3)	179.966	85.47(3)	85.47(3)
Y1_a	3.40697(5)					
Y1_b	3.40697(4)					
Y1_e	3.40697(4)					
Y1_f	3.40697(4)					
Y1_g	3.40697(5)					

pdCIFplot: Challenge

- Change pdCIF from write-only to RW: Plot powder diffraction data/results from CIF
- Requirements
  - Select from many relevant data fields
  - Need Tcl/Tk CIF parser
- Task 1: tabulate possible data names for abscissa & ordinates

Toby, B. H., "Inspecting Rietveld Fits from pdCIF: pdCIFplot", *J. Appl.Cryst.* **36**, 1285 (2003)
http://www.ncnr.nist.gov/xtal/software/cif/pdCIFplot.html

Table 1
pdCIF data items used for recording the powder diffraction dependent variable.

_pd_meas_2theta_range_†	Uncorrected 2θ values with constant steps
_pd_meas_2theta_scan	Uncorrected 2θ values, which may not have constant steps
_pd_proc_2theta_range_†	Calibration corrected 2θ values with constant steps
_pd_proc_2theta_corrected	Calibration corrected 2θ values, which may not have constant steps
_pd_meas_time_of_flight	Time for time-of-flight neutron diffraction measurements
_pd_meas_position	Linear detector position
_pd_proc_energy_incident	X-ray energy for energy-dispersive measurements
_pd_proc_wavelength	X-ray or neutron wavelength, when not constant
_pd_proc_d_spacing	d spacing corresponding to an intensity value
_pd_proc_recip_len_Q	Momentum transfer ($Q = 4\pi \sin\theta/\lambda$) for an intensity value

† The data names indicated as _pd_XXXX_2theta_range_ actually correspond to three CIF data items, _pd_XXXX_2theta_range_min, _pd_XXXX_2theta_range_max and _pd_XXXX_2theta_range_inc, which define a range of equally spaced values.

Table 2
pdCIF data items used for powder diffraction intensity values.

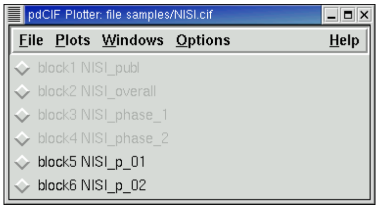
Observed intensities y(obs)	Uncertainty values $\sigma_{y(\mathrm{obs})}$
_pd_meas_counts_total	_pd_meas_counts_total†
_pd_meas_intensity_total	_pd_meas_intensity_total‡
_pd_proc_intensity_total	_pd_proc_intensity_total‡
_pd_proc_intensity_net	_pd_proc_intensity_net‡
	_pd_proc_ls_weight§

† Standard uncertainty is the square-root of the counts for this data item.

Background intensity y(bck)	Calculated intensities y(calc)
_pd_meas_counts_background	_pd_calc_intensity_net
_pd_meas_counts_container	_pd_calc_intensity_total
_pd_meas_intensity_background	
_pd_meas_intensity_container	
_pd_proc_intensity_bkg_calc	
_pd_proc_intensity_bkg_fix	

Sequential GUI programming (ugh)

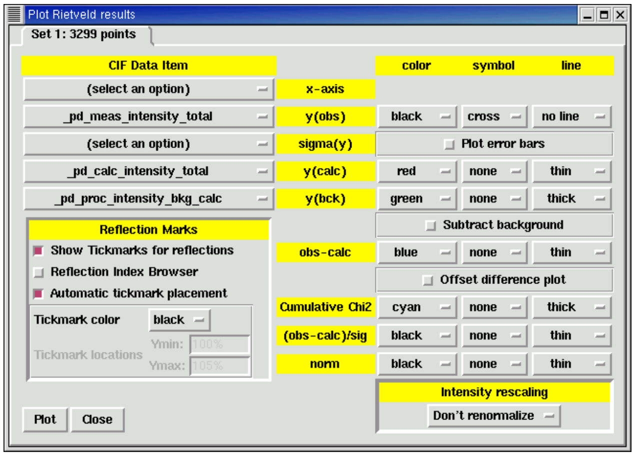
Select block (skipped if no choices)



pdCIF Plotter: file samples/NISI.cif

◇ block1 NISI_publ
◇ block2 NISI_overall
◇ block3 NISI_phase_1
◇ block4 NISI_phase_2
◇ block5 NISI_p_01
◇ block6 NISI_p_02

Specify plot contents



Plot Rietveld results

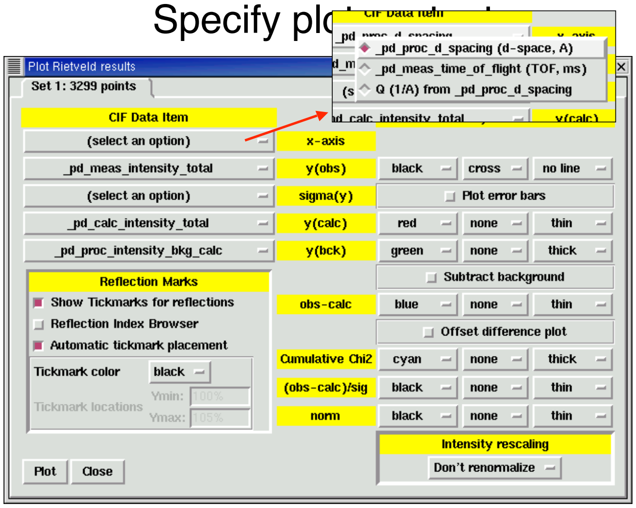
Set 1: 3299 points

CIF Data Item		color	symbol	line
(select an option)	x-axis			
_pd_meas_intensity_total	y(obs)	black	cross	no line
(select an option)	sigma(y)	□ Plot error bars		
_pd_calc_intensity_total	y(calc)	red	none	thin
_pd_proc_intensity_bkg_calc	y(bck)	green	none	thick
		□ Subtract background		
	obs-calc	blue	none	thin
		□ Offset difference plot		
	Cumulative Chi2	cyan	none	thick
	(obs-calc)/sig	black	none	thin
	norm	black	none	thin

Reflection Marks
■ Show Tickmarks for reflections
□ Reflection Index Browser
■ Automatic tickmark placement
Tickmark color: black
Tickmark locations Ymin: 100% Ymax: 105%

Intensity rescaling: Don't renormalize

Specify plo



Plot Rietveld results

Set 1: 3299 points

CIF Data Item				
(select an option)	x-axis			
_pd_meas_intensity_total	y(obs)	black	cross	no line
(select an option)	sigma(y)	□ Plot error bars		
_pd_calc_intensity_total	y(calc)	red	none	thin
_pd_proc_intensity_bkg_calc	y(bck)	green	none	thick
		□ Subtract background		
	obs-calc	blue	none	thin
		□ Offset difference plot		
	Cumulative Chi2	cyan	none	thick
	(obs-calc)/sig	black	none	thin
	norm	black	none	thin

Reflection Marks

☒ Show Tickmarks for reflections

□ Reflection Index Browser

☒ Automatic tickmark placement

Tickmark color: black

Tickmark locations Ymin: 100% Ymax: 105%

Intensity rescaling

Don't renormalize

Plot | Close

CIF Data Item

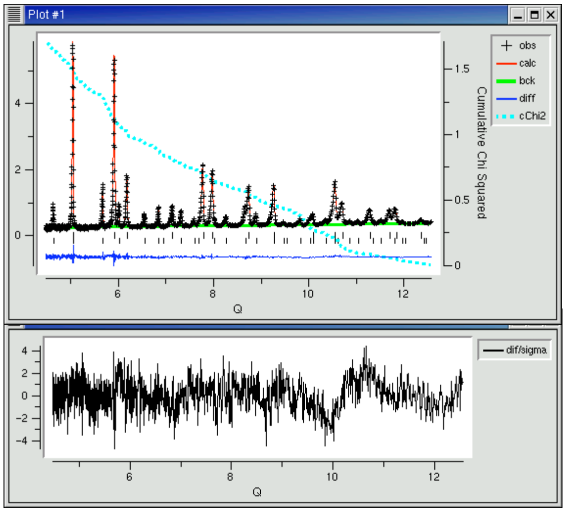
_pd_proc_d_spacing | x-axis

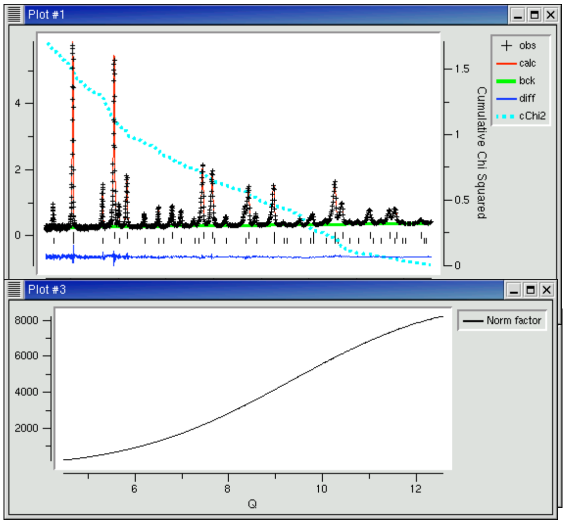
◆ _pd_proc_d_spacing (d-space, A)

◇ _pd_meas_time_of_flight (TOF, ms)

◇ Q (1/A) from _pd_proc_d_spacing

_pd_calc_intensity_total | y(calc)


Plot #1
4
2
0
6
8
10
12
Q
1.5
1
0.5
0
Cumulative Chi Squared
obs
calc
bck
diff
cChi2
4
2
0
−2
−4
6
8
10
12
Q
dif/sigma


Plot #1
4
2
0
1.5
1
0.5
0
Cumulative Chi Squared
obs
calc
bck
diff
cChi2
Plot #3
8000
6000
4000
2000
6
8
10
12
Q
Norm factor